UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Science
School of Electronics and Computer Science

Human Action Recognition Based on Convolutional Neural Networks and Vision Transformers

by

Khaled Alomar

ORCiD: 0000-0002-8303-3240

A thesis for the degree of
Doctor of Philosophy

March 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Science
School of Electronics and Computer Science

Doctor of Philosophy

Human Action Recognition Based on Convolutional Neural Networks and Vision Transformers

by Khaled Alomar

This thesis seeks to deepen our understanding and expand our knowledge of the impacts of deep-learning techniques on human action recognition. It addresses the challenges faced in human action recognition and proposes solutions focused on enhancing feature extraction and optimizing model designs. This is accomplished through the completion of three distinct yet closely interconnected chapters (i.e., papers). These chapters are: (i) *Data Augmentation in Classification and Segmentation: A Survey and New Strategies*; (ii) *TransNet: A Transfer Learning-Based Network for Human Action Recognition*; and (iii) *RNNs, CNNs, and Transformers in Human Action Recognition: A Survey and a Hybrid Model*.

The second chapter provides a survey of the existing data augmentation techniques in computer vision tasks, including segmentation and classification. Data augmentation is a well-established method in computer vision. It can be especially beneficial for human action recognition (HAR) by enhancing feature extraction. This technique addresses challenges such as limited datasets and class imbalance, resulting in more robust feature extraction and reduced overfitting in neural networks. Studies have demonstrated that data augmentation significantly enhances the accuracy and generalizability of models in tasks like image classification and segmentation, which is subsequently utilized in the task of HAR in the third chapter.

The third chapter addresses two significant challenges in HAR: feature extraction and the complexity of HAR models. It introduces a straightforward, yet versatile and effective end-to-end deep learning architecture, termed TransNet, as a solution to these challenges. Extensive experimental results and comparisons with state-of-the-art models demonstrate the superior performance of TransNet in terms of flexibility, model complexity, transfer learning capability, training speed, and classification accuracy. Additionally, this chapter introduces a novel strategy that utilizes autoencoders to form the 2D component of TransNet, referred to as TransNet+. TransNet+ enhances feature extraction by directing the model to extract specific features based on our needs. TransNet+ leverages the encoder part of an autoencoder, trained on computer vision tasks such as human semantic segmentation (HSS), to perform HAR. The extensive experimental results and comparisons with leading models further validate the superior performance of both TransNet and TransNet+ in HAR.

The fourth chapter provides a comprehensive review of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Vision Transformers (ViTs). It examines the progression from traditional methods to the latest advancements in neural network architectures, offering a chronological and extensive analysis of the existing literature on action recognition. The chapter proposes a novel hybrid model that integrates the strengths of CNNs and ViTs. Additionally, it offers a detailed performance comparison of the proposed hybrid model against existing models, highlighting its efficacy in handling complex HAR tasks with improved accuracy and efficiency. The chapter also discusses emerging trends and future directions for HAR technologies.

Contents

List of Figures

List of Tables

Declaration of Authorship

Title of thesis:	Human Action Recognition Based on Convolutional Neural Networks and Vision Transformers
Print name:	Khaled Abdulaziz Alomar

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging*, 9(2):46, 2023
   Khaled Alomar and Xiaohao Cai. Transnet: A transfer learning-based network for human action recognition. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1825–1832. IEEE, 2023

Signed: ....*Khaled Alomar*.................. Date:......*20/03/2025*............

Acknowledgements

I would like to start my acknowledgements in gratitude to *Allah* for having bestowed the blessing and his mercy on me to complete this thesis successfully.

I would like to express my heartfelt gratitude and appreciation to everyone who supported and encouraged me throughout my PhD journey. I am deeply thankful to my supervisors, Dr Xiaohao Cai and Dr Srinandan Dasmahapatra, for their unwavering support, supervision, and guidance throughout this research. My sincere thanks also go to Dr Jo Grundy and Dr Hansung Kim.

I am deeply grateful to my Soton family and my dearest friends in the UK. The wonderful moments and enjoyable times spent with them will be cherished forever. This journey has been the most enjoyable, productive, and stimulating experience of my life. I would especially like to thank my close friend Halil Ibrahim Aysel for being a part of it.

I extend my gratitude to the Saudi Arabian Cultural Bureau in London for financially supporting my PhD scholarship. Special thanks to my wife, whose constant motivation and optimism have been invaluable during this period. Lastly, I am profoundly grateful to my parents, brothers, and sisters for their enduring love and encouragement.

To my children, Faisal and Leena.

Definitions and Abbreviations

HAR	Human action recognition
CNN	Convolutional neural network
RNN	Recurrent neural network
ViT	Vision transformer
NLP	Natural language processing
RGB	Red, green and blue
RGB-D	Red, green and blue with depth
CSI	Channel state information
STIP	Spatio-temporal Interest Points
HOG	Histogram of Oriented Gradients
3D CNNs	3D Convolutional Neural Networks
LSTM	Long Short-Term Memory
GNNs	Graph Neural Networks
SVM	Support Vector Machine
TSN	Temporal Segment Networks
TDD	Trajectory-pooled Deep-convolutional Descriptors
IDT	Improved Dense Trajectory
SPN	Spatiotemporal Pyramid Network
TCLSTA	Two-stream Collaborative Learning with Spatial-Temporal Attention
MBH	Motion Boundary Histogram
HOF	Histogram of Optical Flow
IDT	Improved Dense Trajectories
GRUs	Gated Recurrent Units
BPTT	Back Propagation Through Time
RLR	Random Local Rotation
HSS	Human Semantic Segmentation

Chapter 1

Introduction and Background

1.1 Preamble

This PhD thesis explores the task of Human Action Recognition (HAR) concerning the data used, obstacles encountered, modeling methods, optimization techniques, and applications. It investigates the most significant challenges faced by HAR applications. It seeks to find comprehensive and integrated solutions by connecting the challenges and devising solutions that aim to address multiple issues simultaneously. To this end, this thesis, through three interconnected chapters (i.e., previously papers), aims to advance knowledge and understanding of chapters explore current techniques such as data augmentation and transfer learning, as well as potential model designs that can integrate these solutions for optimal performance. By thoroughly examining these components, the research aims to offer a holistic view of the HAR domain, identifying gaps in existing methodologies and proposing innovative approaches to enhance the accuracy and efficiency of HAR systems. This comprehensive exploration is intended to pave the way for more robust and adaptable HAR models.

The introduction chapter is organized as follows: Section 1.2 provides the research motivation, with the research background discussed in Section 1.3. Section 1.4 details the challenges of the research. The main questions are presented in Section 1.6, while Section 1.7 describes the research objectives. Finally, section 1.8 summarizes each of the three conducted studies.

1.2 Research Motivation

HAR is an increasingly vital area of research within the broader field of computer vision and artificial intelligence. As the world becomes more interconnected and technology-driven, the ability to accurately and efficiently recognize and interpret human actions from video or sensor data has numerous practical applications Morshed et al. (2023). These range from enhancing security systems with advanced surveillance capabilities, improving human-computer interaction and developing assistive technologies for the elderly and disabled Kumar and Kumar (2024). The significance of HAR lies in its potential to revolutionize how we interact with technology, making systems more intuitive, responsive, and beneficial to society.

The growing interest in HAR is driven by its applications in various fields. For instance, in healthcare, HAR can be used for remote patient monitoring, detecting falls, and assessing rehabilitation progress Kumar et al. (2024). In the field of security, HAR can enhance surveillance systems by automatically identifying suspicious activities, reducing the need for constant human monitoring Diraco et al. (2023). Furthermore, in entertainment and sports, HAR can be employed to analyze player movements, enhancing both coaching strategies and viewer experiences Sharshar et al. (2023). These diverse applications underscore the importance of developing robust and reliable HAR systems.

Moreover, a thesis focused on human action recognition can have a profound impact beyond academia. In smart home environments, HAR can enhance user experiences by making systems more adaptive to human needs and behaviors Diraco et al. (2023). Furthermore, in the realm of robotics, HAR can enable more natural and effective human-robot interactions, which is crucial for the development of autonomous systems Kansal et al. (2023). Thus, the potential applications of HAR are vast and varied.

Pursuing a PhD thesis in human action recognition presents a unique opportunity to contribute to a rapidly evolving field with far-reaching implications. The complexity of HAR, which involves understanding complex human movements and behaviors in diverse and often unpredictable environments, provides a rich ground for innovative research. By tackling challenges such as Feature extraction, difficulty of transfer learning, temporal analysis, occlusion, varied action execution styles (i.e. Different people perform the same action in unique ways, and actions can be influenced by factors such as the environment, clothing, and lighting conditions.), and the integration of multi-modal data, the research can push the boundaries of what is currently possible in HAR Singh et al. (2021); Morshed et al. (2023); Pareek and Thakkar (2021); Jegham et al. (2020); Ramanathan et al. (2014). A PhD research project can focus on developing algorithms that can generalize well across these variations. This might involve the use of deep learning techniques, which have shown great promise in handling complex and high-dimensional data. By improving the robustness of HAR systems, the research can make significant contributions to the field.

Finally, the insights gained from HAR research can contribute to the development of safer and more efficient autonomous systems, such as self-driving cars and robotic assistants. By enabling machines to understand and predict human actions, these systems can operate more safely and effectively in dynamic environments. The societal impact of such advancements is immense, as it can lead to reduced accidents and improved quality of life. Therefore, a PhD thesis in human action recognition not only advances scientific knowledge but also holds the promise of creating technologies that improve the quality of life across various sectors.

1.3 Research Background

1.3.1 HAR

HAR is a rapidly evolving field that seeks to automate the recognition of human activities from sensory inputs, such as videos or sensor data. The overarching goal is to build systems capable of accurately identifying various human movements and activities. With advancements in machine learning and artificial intelligence, particularly deep learning, the ability to process and analyze large datasets of human actions has significantly improved, leading to higher accuracy and efficiency. HAR systems have become increasingly sophisticated, capable of handling diverse tasks from detecting simple movements like walking to more complex actions involving multiple actors or environments.

Section 1.3.2 focuses on categorizing HAR based on the type of data utilized. Various data sources contribute to different applications of HAR. For instance, sensor-based methods involve wearable sensors, commonly used in healthcare and fitness, to monitor daily activities and track patient movement. On the other hand, vision-based methods, which are increasingly popular, rely on RGB cameras and depth sensors to capture video data for tasks like video surveillance or gaming. Emerging technologies, such as radar- and WiFi-based recognition, are also gaining attention for their ability to detect movement without relying on visual inputs, making them suitable for privacy-sensitive environments.

In contrast, Section 1.3.3 discusses HAR based on the methods employed. The methodologies employed in HAR can be broadly divided into two categories: traditional hand-crafted feature methods and modern deep learning approaches. Hand-crafted methods involve designing specific features, such as motion trajectories or interest points, and were initially used in early HAR systems. However, as datasets grew in size and complexity, deep learning-based methods emerged, providing the ability to automatically learn complex patterns from large datasets without manual feature engineering. Methods like CNNs, RNNs, and more recently, Transformers, have revolutionized HAR by improving accuracy and adaptability across diverse applications.

These distinctions, based on data types and methodologies, set the stage for exploring both the challenges and potential of HAR. Understanding these classifications is critical as they directly influence the choice of algorithms, model design, and the overall effectiveness of HAR systems.

1.3.2 HAR Based on Data Type

Human action recognition can be categorized into distinct approaches based on the type of input data used, each offering unique capabilities and applications. Video-based HAR uses RGB video sequences to capture spatial and temporal dynamics of human actions, providing a rich representation of motion and environmental context. Input data for this category consists of consecutive frames, enabling detailed analysis of human actions. This approach is widely applied in video surveillance, sports performance analysis, and gesture-based human-computer interaction, where contextual information is essential.

Skeleton-based HAR models human actions by representing the human body as a set of joint coordinates in 2D or 3D space. This abstraction focuses exclusively on movement and posture, making it robust against variations in lighting and background conditions. The input data consists of sequences of joint positions, often captured using motion capture systems or pose estimation algorithms. Applications for this method include gaming, virtual reality, healthcare monitoring, and fitness tracking, where precise motion analysis is crucial.

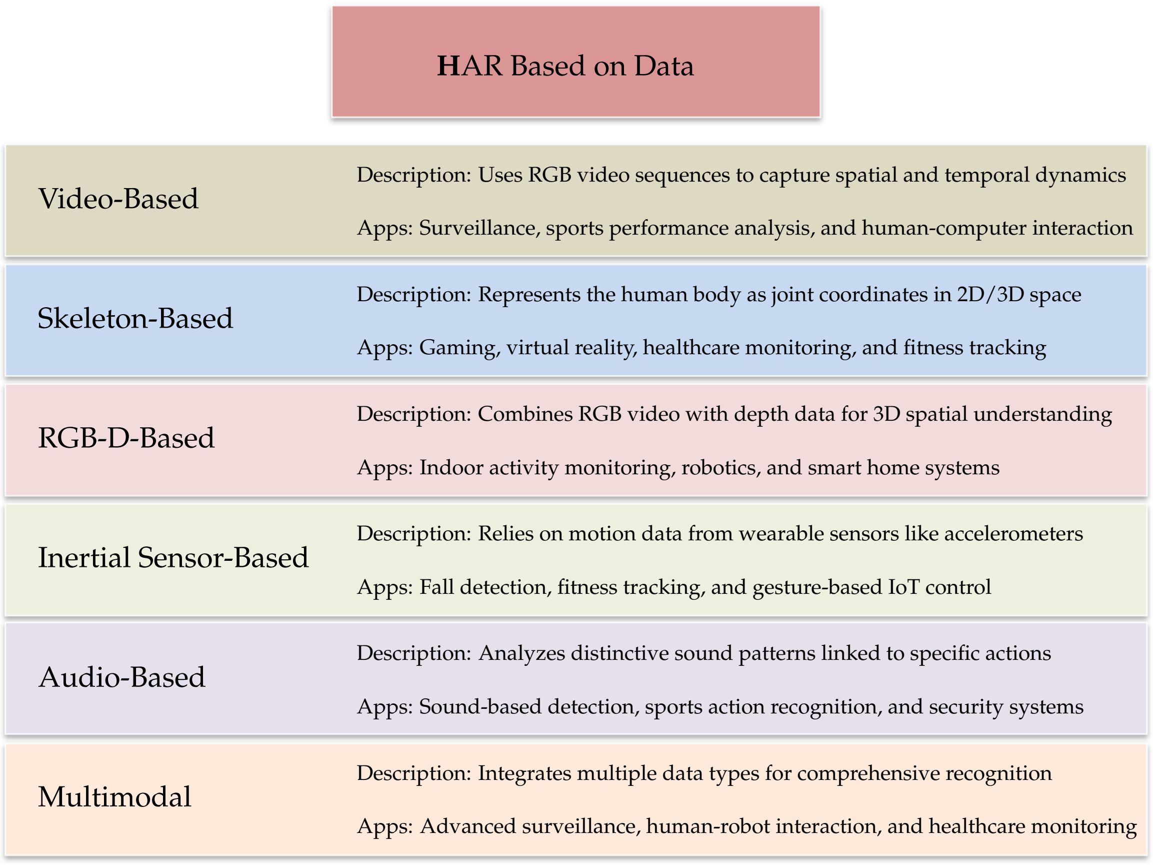
RGB-D-based HAR enhances traditional video-based methods by integrating depth information, enabling a 3D spatial understanding of human actions and their environment. Depth data complements RGB images by providing structural details, making this approach particularly effective in cluttered or low-light settings. Input data consists of paired RGB frames and depth maps. This category is commonly used in indoor activity monitoring, robotics for human-object interaction, and smart home automation, where understanding spatial relationships is critical.

Inertial sensor-based HAR relies on motion data collected from wearable devices or embedded sensors, such as accelerometers, gyroscopes, and magnetometers. These sensors capture the intensity and direction of movement, with data stored as time-series signals. Applications include fall detection in elderly care, fitness tracking, and gesture-based control for smart devices. This lightweight and non-intrusive approach is particularly useful in scenarios where visual data may not be feasible or where privacy concerns are significant.

Audio-based HAR identifies human actions through the analysis of distinctive sound patterns produced by specific activities. Input data is provided as audio signals or spectrograms. This method finds applications in sound-based event detection for smart homes, sports action recognition (e.g., detecting referee whistles or ball hits), and security systems for anomaly detection. Audio-based HAR offers an effective alternative in situations where visual input is unavailable or impractical.

Multimodal HAR integrates multiple data types, such as RGB video, depth maps, skeleton data, audio signals, and inertial sensor readings, to leverage complementary information for enhanced recognition accuracy. Input data is stored in a combination of formats, enabling a holistic representation of human actions by combining spatial, temporal, and contextual data. This approach is suitable for advanced surveillance systems, human-robot interaction, and healthcare monitoring. By integrating diverse data sources, multimodal HAR overcomes the limitations of individual modalities, ensuring robust and reliable performance across various applications.

This categorization highlights the diverse methodologies in HAR, each tailored to specific input data formats and application domains. Together, these approaches form a comprehensive framework for understanding and recognizing human actions in a variety of contexts.



HAR Based on Data

Video-Based	Description: Uses RGB video sequences to capture spatial and temporal dynamics	Apps: Surveillance, sports performance analysis, and human-computer interaction
Skeleton-Based	Description: Represents the human body as joint coordinates in 2D/3D space	Apps: Gaming, virtual reality, healthcare monitoring, and fitness tracking
RGB-D-Based	Description: Combines RGB video with depth data for 3D spatial understanding	Apps: Indoor activity monitoring, robotics, and smart home systems
Inertial Sensor-Based	Description: Relies on motion data from wearable sensors like accelerometers	Apps: Fall detection, fitness tracking, and gesture-based IoT control
Audio-Based	Description: Analyzes distinctive sound patterns linked to specific actions	Apps: Sound-based detection, sports action recognition, and security systems
Multimodal	Description: Integrates multiple data types for comprehensive recognition	Apps: Advanced surveillance, human-robot interaction, and healthcare monitoring

FIGURE 1.1: HAR based on data type.

1.3.3 HAR Based on Methods

HAR can be broadly classified into two main categories based on the methods used: handcrafted feature-based methods and deep learning-based methods. Handcrafted feature-based methods rely on engineered features to represent actions, utilizing traditional image processing and computer vision techniques. Key techniques in this category include spatio-temporal interest points (STIP) Willems et al. (2008); Laptev (2005); Das Dawn and Shaikh (2016); Chakraborty et al. (2011), which detect significant motion points in both spatial and temporal dimensions. Optical flow and optical-flow-like features, such as histogram of optical flow (HOF) and motion boundary histogram (MBH), Horn and Schunck (1981); Ilg et al. (2017); Zhu et al. (2020b); Danafar and Gheissari (2007); Guo et al. (2010); Mahbub et al. (2011), which measures the motion of objects by calculating pixel flow between consecutive frames. Techniques like histogram of oriented gradients (HOG) Dalal and Triggs (2005); Baumann (2013); Ohn-Bar and Trivedi (2013); Lu and Little (2006), capture the shape and appearance of objects by counting occurrences of gradient orientation in localized image portions. Dense and improved dense trajectories (IDT) techniques track the movement of key points or objects across consecutive frames to capture motion dynamics Wang and Kl (2011); Wang and Schmid (2013). These methods were foundational in early video surveillance, sports analysis, and motion detection systems.

On the other hand, deep learning-based methods leverage neural networks to automatically learn features from data, offering high accuracy and scalability with large datasets LeCun et al. (2015); Chai et al. (2021). Prominent techniques in this category include Two-Stream convolutional neural networks (Two-Stream CNNs), which encompass multi-stream CNNs that can process various modalities, such as RGB frames, optical flow, and other sensor data, separately and then combine the results to recognize actions. Additionally, 3D convolutional neural networks (3D CNNs) extract spatio-temporal features from video data. RNNs and long short-term memory (LSTM) networks are used to capture temporal dependencies in sequential data, modeling the sequence of actions over time. Furthermore, Transformers handle long-range dependencies and interactions within the data using self-attention mechanisms. Graph neural networks (GNNs) are also employed to model the relationships and interactions between different body parts by representing the human body as a graph structure Yan et al. (2018); Shi et al. (2019). These methods are widely applied in video action recognition, sports analytics, gesture recognition, and large-scale video analysis, marking significant advancements in HAR technology Adel et al. (2022); Saini and Maan (2020).

In this thesis, we will explore deep learning methods for human action recognition due to their superior performance and adaptability compared to traditional handcrafted methods. Deep learning approaches, particularly those using neural networks, have demonstrated remarkable success in automatically learning complex patterns and features from large-scale datasets without the need for manual feature engineering. This ability to learn hierarchical representations from raw data allows deep learning models to capture complex details of human actions, leading to higher accuracy and robustness in various applications. Additionally, the scalability of deep learning methods makes them suitable for processing extensive video data, which is essential for modern HAR systems. By focusing on deep learning techniques, this research aims to leverage the latest advancements in artificial intelligence to develop more effective and efficient HAR solutions.

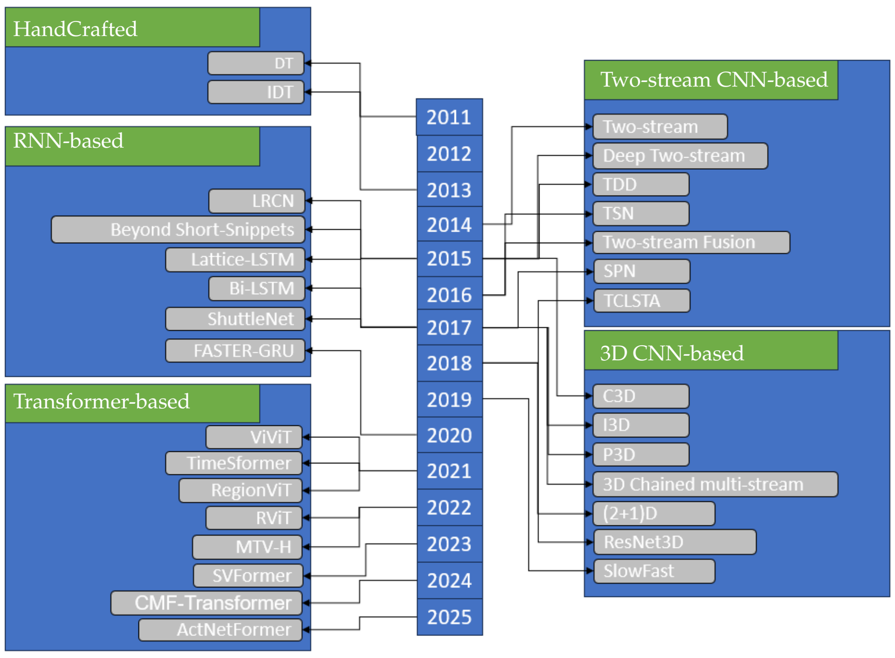


FIGURE 1.2: Categorically-chronologically ordered HAR methods.

1.3.3.1 Two-stream CNN-based Methods

The fundamental concept behind Two-Stream CNNs is to process spatial and temporal data separately before combining them to make a final prediction. Typically, one stream handles spatial information using RGB frames from a video, capturing the appearance of objects and their locations. The other stream processes temporal information, often using optical flow or motion vectors, to capture movement and dynamics over time. Simonyan et al. Simonyan and Zisserman (2014a) introduced a two-stream CNN model that integrates spatial and temporal flows. The spatial flow is responsible for capturing appearance information, while the temporal flow focuses on motion information. The classification scores from both flows are then fused using either an averaging method or a support vector machine (SVM). Following this foundational work, numerous approaches have been developed to enhance the two-stream model Wang et al. (2015a,b, 2016); Feichtenhofer et al. (2016); Wang et al. (2017); Peng et al. (2018).

Wang et al. (2015b) identified that many existing two-stream CNNs are relatively shallow. Consequently, they designed deeper two-stream CNNs to achieve superior recognition performance. Feichtenhofer et al. Feichtenhofer et al. (2016) explored various fusion strategies and demonstrated that fusing the spatial and temporal flows at the final convolutional layer is effective in reducing the number of parameters while maintaining accuracy. To address the challenge of modeling long temporal structures, Wang et al. (2016) introduced temporal segment networks (TSN). TSN employs a sparse sampling scheme to represent temporal features, enabling the modeling of entire videos. Additionally, the authors proposed two supplementary input modalities: RGB difference and warped optical flow, which enhance the learning efficiency of the original two-stream network. Wang et al. (2015a) introduced trajectory-pooled deep-convolutional descriptors (TDD), which integrate classical improved dense trajectory (IDT) handcrafted features with two-stream deep learning features. To further enhance spatiotemporal feature integration in two-stream architecture, a novel spatiotemporal pyramid network (SPN) was proposed by Wang et al. (2017), leveraging a pyramid structure to amplify feature interactions. Peng et al. Peng et al. (2018) advanced this field with their two-stream collaborative learning with spatial-temporal attention (TCLSTA) method, comprising a spatial-temporal attention model and a static-motion collaborative model.

The integration of the spatial and temporal streams is a critical step in Two-Stream CNNs. Various fusion strategies have been proposed, including early fusion (combining data before feature extraction), late fusion (combining after separate feature extraction), and mid-level fusion (combining features at intermediate layers). Late fusion is commonly used, where the outputs of both streams are concatenated or averaged before being fed into a final classifier. This approach allows the model to learn complementary features from both streams effectively Feichtenhofer et al. (2016); Simonyan and Zisserman (2014a); Wang et al. (2016).

Training Two-Stream CNNs involves optimizing both streams simultaneously, which can be challenging due to their different nature. Transfer learning is often employed, using pre-trained networks on large-scale image and video datasets. Fine-tuning these networks on HAR datasets helps in adapting the learned features to specific activities. Additionally, techniques such as data augmentation, dropout, and batch normalization are utilized to enhance generalization and prevent overfitting Wang et al. (2015b).

However, Two-Stream CNNs face several challenges. One significant issue is the computational complexity and memory requirements due to the dual processing streams, making real-time implementation difficult, especially on resource-constrained devices. Additionally, the need for accurate computation of handcrafted techniques used in the temporal stream, such as dense trajectories or optical flow, can be a bottleneck, as errors in motion estimation directly affect the temporal stream's performance.

1.3.3.2 3D CNN-based Methods

3D CNNs have emerged as a powerful tool in HAR Tran et al. (2015); Carreira and Zisserman (2017); Qiu et al. (2017); Tran et al. (2018); Feichtenhofer et al. (2019); Zolfaghari et al. (2017) due to their ability to capture spatial and temporal features simultaneously. 3D CNNs extend the traditional 2D CNNs by adding an extra dimension to the convolutional filters. This enables the network to process video data by considering the temporal sequence of frames, thereby capturing motion and spatial information concurrently. The 3D convolutional layers apply filters across the height, width, and time dimensions, allowing for comprehensive feature extraction from sequential data.

The general architecture of 3D CNNs typically includes several layers of 3D convolutions, pooling, and fully connected layers. The 3D convolutional layers extract spatiotemporal features, while the pooling layers reduce the spatial and temporal dimensions, helping to manage computational complexity and reduce overfitting. The final fully connected layers, followed by softmax activation, are used for classification tasks, such as recognizing specific human activities.

Compared to other HAR methods, such as 2D CNNs and RNNs, 3D CNNs provide a more integrated approach to spatiotemporal feature extraction. While 2D CNNs focus primarily on spatial features and RNNs on temporal sequences, 3D CNNs combine these aspects, leading to more accurate and robust activity recognition. This makes them particularly suitable for applications involving continuous video streams.

The C3D (Convolutional 3D) model, introduced by Tran et al. (2015), is a pioneering approach in video analysis that uses 3D CNNs to capture spatiotemporal features. 2D CNNs benefit from pre-training on extensive image datasets like ImageNet Deng et al. (2009) and Places205 Zhou et al. (2017a), which are significantly larger than any available video datasets. A considerable amount of research has concentrated on employing 2D CNN architectures that achieve higher accuracy and better generalization, subsequently adapting them for video classification tasks.

Proposed by Carreira and Zisserman (2017), the I3D model extends the capabilities of 2D CNNs to handle video data by inflating 2D convolutions into 3D. The core idea behind I3D is to leverage pre-trained 2D CNNs, such as Inception-v1 Szegedy et al. (2015), and inflate their 2D filters and pooling kernels into 3D. ResNet3D was introduced by Hara et al. (2018) extends the 2D ResNet He et al. (2016) by replacing 2D convolutional filters with 3D kernels. This allows the network to process temporal information in videos, leveraging the success of 2D CNNs on large-scale image datasets like ImageNet for improved video analysis.

The kernel-level decomposition of 3D convolutions in P3D Qiu et al. (2017) and R2+1D networks Tran et al. (2018) is proposed to effectively capture spatiotemporal features while significantly reducing computational complexity. By splitting 3D convolutional filters into separate spatial and temporal components, specifically a 2D convolutional kernel for spatial dimensions (height and width) and a 1D convolutional kernel for the temporal dimension (time), these networks can leverage the strengths of 3D convolutions without the extensive computational and memory costs.

Another approach to employing 3D CNNs is through the use of two or multiple streams to enhance performance. The SlowFast network by Feichtenhofer et al. (2019) which utilizes a two-stream 3D CNN architecture: the Fast pathway processes video at a higher frame rate, capturing rapid movements and short-term dynamics, while the Slow pathway operates at a lower frame rate, focusing on long-term temporal dynamics. In a related development, Zolfaghari et al. (2017). employed a three-stream 3D CNN to integrate three crucial visual cues: pose, optical flow, and RGB frames.

Despite their advantages, 3D CNNs face several challenges. One major issue is the high computational cost and memory requirements due to the added temporal dimension in the convolutional layers. This makes training and deploying 3D CNNs on resource-constrained devices challenging. Additionally, training 3D CNNs involves optimizing the network parameters to accurately predict activities from input data. This requires a large dataset of labeled videos or sensor data representing various activities Zhu et al.. The need for large annotated datasets for training can be a significant barrier, as collecting and labeling video data is time-consuming and labor-intensive.

1.3.3.3 RNN-based Methods

RNNs are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. This ability to retain information over time makes them ideal for tasks like HAR, where activities unfold over several frames or time steps. Standard RNNs have limitations like vanishing gradients, but advanced variants such as LSTM networks Hochreiter and Schmidhuber (1997) and gated recurrent units (GRUs) address these issues effectively.

LSTM networks are a popular choice in HAR due to their ability to remember long-term dependencies. An LSTM cell consists of gates that control the flow of information, allowing the network to retain or forget information as needed Hochreiter and Schmidhuber (1997). This is particularly useful in HAR, where activities can vary in duration and complexity. LSTMs have been successfully applied to recognize activities like walking, running, and complex gestures by processing sequential sensor data Ordóñez and Roggen (2016). GRUs, a simpler alternative to LSTMs, also perform well in HAR. They merge the input and forget gates of LSTMs into a single update gate, simplifying the architecture and reducing computational complexity. Despite their simpler structure, GRUs have demonstrated comparable performance to LSTMs in various HAR tasks, making them a viable option for applications where computational efficiency is crucial Dua et al. (2021).

The temporal sequence information in video data, essential for HAR, makes RNNs Sun et al. (2017); Ullah et al. (2017); He et al. (2021); Donahue et al. (2015); Ballas et al.

(2015); Yue-Hei Ng et al. (2015) a highly suitable option. Research initiatives like LRCN Donahue et al. (2015) and Beyond-Short-Snippets Yue-Hei Ng et al. (2015) have pioneered the use of LSTM networks for action recognition in videos within a two-stream network setting. In these frameworks, CNNs extract features from individual video frames, which are then fed into LSTM networks. LRCN extracts CNN features from a single frame and inputs them into an LSTM for the HAR task. Beyond-Short-Snippets extends this approach by utilizing pre-trained 2D CNNs to extract features, which are subsequently processed by a stacked LSTM framework.

Ullah et al. (2017); He et al. (2021) employed bidirectional LSTM, which consists of two separate LSTMs designed to learn both forward and backward temporal information for HAR. Lattice-LSTM Sun et al. (2017) enhances the traditional LSTM by learning independent hidden state transitions for memory cells at individual spatial locations, allowing it to effectively model long-term and complex motions. Besides using LSTM, some studies have explored HAR through GRU Ballas et al. (2015); Dwibedi et al. (2018); Kim et al. (2018b); Shi et al. (2017); Zhu et al. (2020a). GRU, which features fewer gates Cho et al. (2014) than LSTM, leads to a reduction in model parameters while typically delivering similar performance for HAR. FASTER-GRU Zhu et al. (2020a) developed a FAST-GRU to aggregate clip-level features from videos, reducing the processing cost of redundant clips and thereby accelerating inference speed.

ShuttleNet Shi et al. (2017) is a biologically inspired deep network embedded within a CNN-RNN framework, featuring a multilayer loop-connected GRU processor. Furthermore, several studies have integrated attention mechanisms Girdhar and Ramanan (2017); Meng et al. (2019); Li et al. (2018b) into LSTM-based frameworks to enhance HAR.

Despite their strengths, RNN-based methods face challenges in HAR. One major issue is the high computational cost associated with training and inference, which can limit real-time applications. Additionally, RNNs can struggle with very long sequences due to issues like gradient vanishing and exploding. Addressing these challenges often requires architectural modifications and efficient training strategies. Recent research has focused on improving RNN architectures and training techniques to enhance HAR performance. Innovations such as attention mechanisms allow the network to focus on relevant parts of the sequence, improving accuracy. Hybrid models that combine RNNs with CNNs have also been explored, leveraging the strengths of both architectures to capture spatial and temporal features.

1.3.3.4 Transformer-based Methods

Transformers, introduced by Vaswani et al. (2017), have revolutionized natural language processing (NLP) by enabling models to understand context over long sequences of data. Recently, their application has extended to HAR, leveraging their powerful ability to capture long-range dependencies and contextual information. This development marks a significant advancement in HAR, where understanding complex temporal patterns is crucial.

Transformers operate using a self-attention mechanism that allows them to weigh the importance of different elements in a sequence. Unlike RNNs, which process sequences step-by-step, transformers process entire sequences simultaneously. This parallel processing capability, coupled with the ability to focus on relevant parts of the input data, makes transformers highly effective for sequence modeling tasks. Additionally, transformers do not suffer from the vanishing gradient problem that affects RNNs, making them more stable and efficient for training on long sequences Han et al. (2020).

The core architecture of a transformer consists of an encoder-decoder structure. The encoder processes the input sequence, generating a set of continuous representations. The decoder uses these representations to produce the output sequence. Each component of the transformer (both encoder and decoder) comprises multiple layers of self-attention and feed-forward neural networks, enabling the model to capture complex dependencies within the data.

Vision Transformers (ViT) are a recent extension of transformers introduced by Dosovitskiy et al. (2020), specifically designed for image and video analysis. ViT primarily uses an encoder-only architecture, rather than the encoder-decoder structure typical of transformers. Unlike traditional CNNs, ViT models split images into a sequence of patches, treating them similarly to tokens in text processing. This approach allows ViT to leverage the self-attention mechanism to capture relationships within the visual data. In the context of HAR, ViT can be applied to video frames, enabling the model to understand spatial and temporal patterns effectively. ViT models have demonstrated strong performance in various visual tasks, making them a promising direction for further advancements in HAR.

Applying transformers to RGB-based action recognition has yielded remarkably superior performance Bertasius et al. (2021); Arnab et al. (2021). Bertasius et al. (2021) expanded the Vision Transformer (ViT), originally designed for image classification Dosovitskiy et al. (2020), to accommodate video data. They achieved this by decomposing each video into a sequence of frame-level patches. Subsequently, they introduced a divided attention mechanism that applies spatial and temporal attention separately within each block of the model. Arnab et al. (2021) introduced ViViT which treats video frames as sequences of image patches and processes them simultaneously, leveraging the self-attention mechanism to capture both spatial and temporal dependencies effectively. Yan et al. (2022) introduced MTV-H which employs separate streams of encoders to process different views of video inputs, utilizing lateral connections to fuse information across these views.

Chen et al. (2021) introduced RegionViT which employs a two-stage attention mechanism: regional attention and local attention. In the regional stage, it processes image patches hierarchically to capture global context. The local stage refines detailed features within each region. This approach allows RegionViT to leverage global self-attention benefits while enhancing focus on important local features. Yang et al. (2022b) presented the Recurring Vision Transformer (RViT) model which integrates a two-layer transformer architecture optimized for video input, utilizing a series of convolutional layers and self-attention mechanisms to capture both spatial and temporal features. Unlike traditional models that handle each frame or segment independently, RViT processes the entire video sequence holistically, allowing it to capture long-range dependencies and complex temporal patterns essential for accurate action recognition. Moreover, Xing et al. (2023) introduced SVFormer, a semi-supervised video Transformer designed to utilize both labeled and unlabeled data, effectively bridging the gap between supervised and unsupervised learning. This model demonstrated notable advancements in action recognition tasks across several standard HAR datasets, including Kinetics-400 and UCF101. Collectively, these studies highlight the pivotal role of Vision Transformers in driving progress in the field of HAR.

Ahmadabadi et al. (2023) introduces a knowledge distillation approach combining CNN and Vision Transformer (ViT) models for improved human action recognition. The study utilizes ConvNeXt as the teacher model to extract local features and various ViT variants (e.g., PVT, Convit, MViT, Swin Transformer, and Twins) as student models to capture global dependencies. Through distillation, the student models achieved significant improvements in accuracy and mean average precision (mAP) on the Stanford 40 dataset compared to regular training methods. This work highlights the effectiveness of integrating local and global feature learning for advancing action recognition tasks.

Moreover, Wang et al. (2024) proposes a novel model that integrates spatio-temporal video features and skeleton joint data to enhance human action recognition. The CMF-Transformer utilizes directional attention for spatio-temporal modality and cross-attention for skeleton joint modality to capture detailed motion information and maintain the temporal order of actions. A multimodal collaborative recognition strategy is employed to synergistically fuse features from both modalities, optimizing overall performance. Experimental evaluations on benchmark datasets such as NTU-RGB+D and Anubis demonstrate the model's superior accuracy compared to state-of-the-art methods, showcasing its effectiveness in cross-modal action recognition tasks. Later, Dass et al. (2025) introduces ActNetFormer a hybrid architecture combining Transformer and ResNet models for effective semi-supervised action recognition. This approach leverages the strengths of ResNet for extracting local spatial features and Transformers for capturing global temporal dependencies, enabling the model to perform well in scenarios with limited labeled data. The model demonstrates superior performance on standard benchmarks by balancing the trade-off between computational efficiency and accuracy, showcasing its potential for scalable and effective video-based action recognition. This study highlights the efficacy of integrating CNN and Transformer components in hybrid architectures to address challenges in human action recognition.

1.4 Research Challenges

HAR presents a multifaceted array of challenges that can be broadly classified into data-based and model-based categories, see Figure 1.3. Data-based challenges primarily revolve around the nature and quality of the data available for training and testing HAR systems. These include the inherent complexity and variability of human actions, which can be performed in numerous ways depending on individual styles and situational contexts. The issues of occlusions and viewpoint variability further complicate data capture, as actions can be partially or fully obscured or viewed from different angles. Additionally, background clutter introduces noise that can distract from the primary actions being analyzed. Inter-class and intra-class variability, where different actions may appear similar and the same action may appear differently when performed by different individuals, adds to the challenge. Finally, the labor-intensive process of data annotation and labeling requires significant expertise and attention to detail, making it difficult to obtain large, accurately labeled datasets.

Model-based challenges, on the other hand, focus on the limitations and complexities inherent in designing and implementing HAR models. Temporal dynamics are crucial, as recognizing actions involves understanding the progression of movements over time, which requires capturing and processing temporal dependencies accurately. Real-time processing requirements impose additional constraints, as many HAR applications necessitate quick and accurate responses. The integration of multimodal data from various sensors can enhance recognition but also adds complexity to the system. Scalability of models is another concern, as the system must remain efficient and effective even as the diversity of actions and volume of data increase. Transfer learning and domain adaptation are needed to ensure models trained in one environment can generalize to others. Overfitting poses a risk, especially when models are trained on limited or unbalanced datasets, leading to poor generalization to unseen data. Feature extraction is crucial for capturing relevant aspects of actions, requiring techniques that can effectively isolate important details from noise. Meanwhile, model complexity must be carefully managed to avoid excessive computational demands and ensure practical application without sacrificing accuracy.

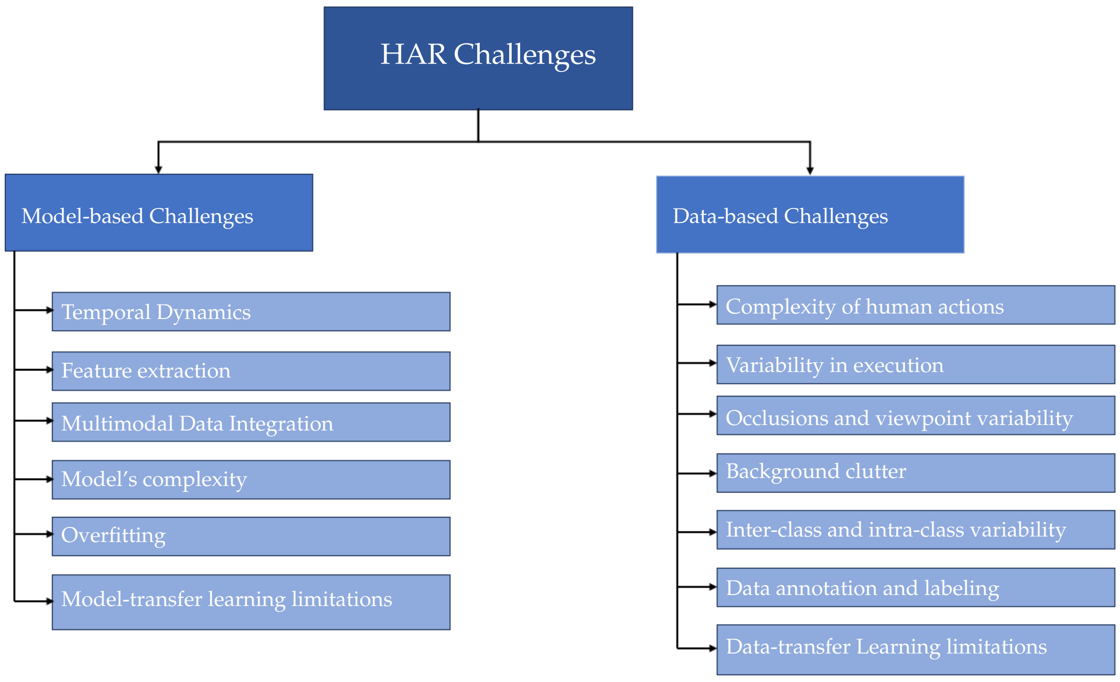

HAR Challenges
Model-based Challenges
Temporal Dynamics
Feature extraction
Multimodal Data Integration
Model's complexity
Overfitting
Model-transfer learning limitations
Data-based Challenges
Complexity of human actions
Variability in execution
Occlusions and viewpoint variability
Background clutter
Inter-class and intra-class variability
Data annotation and labeling
Data-transfer Learning limitations

FIGURE 1.3: HAR challenges taxonomy.

1.4.1 Data-based Challenges

1.4.1.1 Complexity of Human Actions

Human actions involve a wide range of movements and interactions, which can be subtle and complex. This complexity makes it difficult for HAR systems to accurately capture and interpret the small differences of each action. For instance, distinguishing between actions that involve similar motions but different contexts, such as waving hello versus waving to catch someone's attention, requires sophisticated data representation and annotation Weinland et al. (2011); Zhang et al. (2019).

1.4.1.2 Variability in Execution

The way in which an action is performed can vary significantly between individuals and contexts Chaquet et al. (2013). This variability can stem from differences in physical attributes, personal styles, or cultural practices Turaga et al. (2008). HAR systems must be able to generalize across these variations to accurately recognize actions. For example, the same gesture can look different when performed by individuals of different heights or with varying degrees of expressiveness, making it challenging to develop a one-size-fits-all model Singh et al. (2021).

1.4.1.3 Occlusions and Viewpoint Variability

Human actions are often partially or completely obscured by objects, other people, or the performer's own body parts Weinland et al. (2011). Additionally, actions can be observed from various angles and distances, resulting in significant variability in the captured data. Occlusions and changes in viewpoint can lead to incomplete or misleading visual information, complicating the task of accurately recognizing the intended action Jegham et al. (2020). For instance, a handshake might be partially hidden by a tree or another person, making it difficult for the HAR system to detect the action reliably Giannakos et al. (2021).

1.4.1.4 Background Clutter

Actions typically occur in complex environments with dynamic and often noisy backgrounds. Background clutter introduces additional noise that can distract from the primary action being analyzed Jegham et al. (2020). Isolating the action from these backgrounds to ensure accurate recognition is a non-trivial task. For instance, in a crowded street scene, identifying a person waving amidst numerous other activities and moving objects presents a significant challenge for HAR systems.

1.4.1.5 Inter-class and Intra-class Variability

Different actions (inter-class) can sometimes appear similar, such as running versus jogging, while the same action (intra-class) can appear different when performed by different individuals or under different conditions Akila (2022); Jegham et al. (2020). This variability makes distinguishing between actions and ensuring consistency in recognition difficult. For example, walking can vary greatly depending on the person's speed, gait, and external conditions, making it challenging to create a model that accurately captures all variations.

1.4.1.6 Data Annotation and Labelling

Annotating large datasets for action recognition is labor-intensive and time-consuming Jegham et al. (2020). Accurate labeling requires expertise and careful attention to detail, particularly for actions that are ambiguous or involve multiple overlapping movements Kwon et al. (2019). The quality of annotations directly impacts the performance of HAR models, and any errors or inconsistencies can lead to significant degradation in model accuracy. For instance, mislabeling a complex action sequence can mislead the training process, resulting in poor recognition performance.

1.4.1.7 Transfer Learning Limitations Regarding Data

Transfer learning in HAR poses significant limitations when it comes to adapting across diverse data environments. One of the key challenges is that data collected in controlled environments—where lighting, camera angles, and participant actions are carefully managed—differs substantially from the data encountered in real-world settings Jegham et al. (2020). For instance, video footage in controlled lab environments often consists of well-lit, high-resolution recordings, free from noise, obstructions, or occlusions. When a model trained on this data is applied to less controlled environments, such as surveillance footage with poor lighting, low resolution, and unpredictable human behavior, its performance can degrade significantly Kong and Fu (2022). This is because the model's learned features and patterns from the training data may not generalize well to new, more variable data, particularly if the visual characteristics of actions are obscured by factors like occlusion or poor image quality. Such discrepancies highlight the limitations of transfer learning in HAR, as models may struggle to adapt to unforeseen variations in the input data without significant re-training or domain adaptation techniques.

1.4.2 Model-based Challenges

These challenges represent the primary focus of investigation in this thesis, as they encompass critical barriers to the advancement of human action recognition. By addressing these issues, the research aims to propose innovative solutions that improve the robustness, accuracy, and efficiency of action recognition models. This study not only highlights the significance of overcoming these obstacles but also underscores their interconnection with broader challenges in the field of artificial intelligence and computer vision.

1.4.2.1 Temporal Dynamics

Human actions unfold over time, requiring the progression of movements to be understood as a temporal sequence Vasileiou et al. (2021). Accurately capturing these dynamics is critical for effective HAR, but this poses several challenges Jegham et al. (2020). One key difficulty lies in handling variations in action duration and accounting for the temporal dependencies between different movements. Without a clear understanding of how actions evolve over time, HAR models may struggle to distinguish between subtle differences in actions that may appear similar when viewed statically. As a result, modeling the temporal structure of actions becomes vital to ensuring the correct interpretation and recognition of dynamic human behaviors.

For HAR models to distinguish between actions, such as differentiating between sitting down and squatting, they must accurately capture the sequential movements that define these actions Wang et al. (2022). This requires a deep understanding of the temporal relationships between movement phases, allowing the model to recognize the progression and subtle differences between similar actions Bobick and Davis (1997). Successfully identifying these temporal patterns is crucial, as it enables HAR systems to adapt to variations in timing and movement execution, ensuring that they can recognize actions consistently and accurately over a wide range of scenarios.

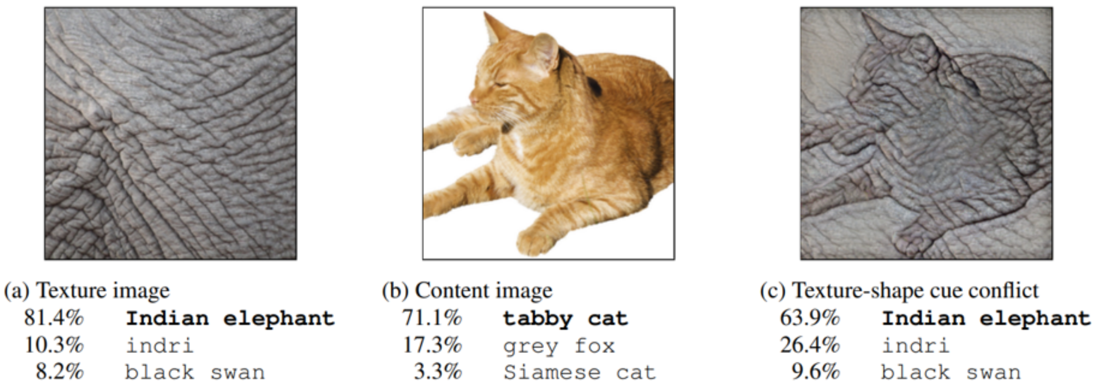
1.4.2.2 Feature Extraction

Effective feature extraction is crucial for capturing the relevant aspects of human actions while minimizing irrelevant information. Extracting meaningful features from raw data requires sophisticated techniques that can isolate important details from noise. For instance, identifying shape of the human body and tracking its movements can provide valuable information for action recognition, but the process needs to be accurate and efficient Hirota and Komuro (2021); Zhang et al. (2021b); Dhiman and Vishwakarma (2020); El-Ghaish et al. (2018).

Shape vs Texture Features: CNNs have the tendency to classify images based on texture rather than shape. Research indicates that CNNs are particularly sensitive to texture features because the convolutional filters in the early layers capture local patterns strongly indicative of texture. This tendency can result in misclassifications when objects share similar textures but differ in shape Hermann et al. (2020); Geirhos et al. (2018).

A CNN is designed to automatically and adaptively detect spatial features Yamashita et al. (2018). However, it remains unclear which specific spatial feature the CNN uses to detect objects (e.g., color, texture, or shape). The study in Geirhos et al. (2018) conducted a quantitative experiment comparing the responses of CNNs and human observers regarding the distinction between shape and texture cues in image classification. The results showed that CNNs trained on ImageNet tend to prioritize texture over shape, as shown in Figure 1.4. This behavior contrasts sharply with that of humans. Conversely, several studies suggest that object shape representations are more critical for action recognition tasks Hirota and Komuro (2021); Zhang et al. (2021b); Dhiman and Vishwakarma (2020); El-Ghaish et al. (2018).

However, for HAR, shape features are crucial. Accurate HAR relies on understanding the overall shape and form of the body and its movements. When models prioritize texture over shape, it can hinder the accurate identification of actions, which depend heavily on the body's shape and movement dynamics. Therefore, enhancing CNNs to



(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

FIGURE 1.4: Conflict between texture and shape features in CNNs trained on ImageNet. Classification of a standard ResNet-50 of (a) a texture image (elephant skin: only texture cues); (b) a normal image of a cat (with both shape and texture cues), and (c) an image with a texture-shape cue conflict, generated by style transfer between the first two images Geirhos et al. (2018)

better capture shape features is essential for improving HAR accuracy Zhang et al. (2019); Kong and Fu (2022); Singh et al. (2021).

Synthetic Frames: In the field of HAR, the use of shape-specific synthetic images, such as body part segmentation, and motion-specific synthetic images, like optical flow, is crucial for improving the accuracy and robustness of action recognition models. Shape-specific synthetic images enable the model to focus on the form and structure of human bodies, capturing critical details about posture and movement dynamics. For instance, segmenting body parts can help distinguish between similar actions that differ slightly in the positioning of limbs, such as running versus walking. This granular level of detail is essential for precise HAR, as it allows the model to learn and recognize the intricate patterns of human movement Zhang et al. (2019); Singh et al. (2021).

Shape-specific synthetic images are typically created using semantic segmentation techniques, which involve partitioning an image into different regions corresponding to various body parts. This process begins with labeling pixels in an image according to the body part they represent, a task usually performed on a large set of training images. These labeled images are then used to train deep learning models, such as convolutional neural networks (CNNs), to recognize and segment different body parts in new, unseen images. Advanced methods employ architectures like U-Net or Mask R-CNN, which are particularly effective in capturing the fine details of body part boundaries and ensuring accurate segmentation even in complex poses and environments Zhang et al. (2019); Singh et al. (2021).

Motion-specific synthetic images, such as optical flow, provide additional context by capturing the movement across frames in a video sequence. Optical flow represents the motion of objects within the visual field, which is vital for understanding dynamic

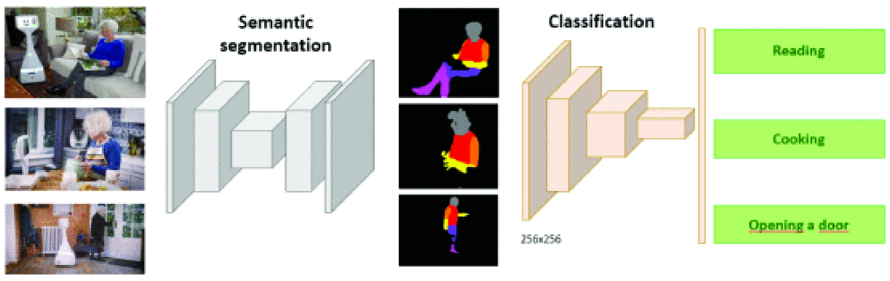


FIGURE 1.5: Artificially created human body part segmentation (synthetic images) Huyghe et al. (2021).

actions. By analyzing these motion patterns, HAR models can better interpret actions that involve complex sequences of movements, such as dancing or sports activities. This temporal information complements the spatial data from body part segmentation, creating a more comprehensive understanding of the action being performed Kong and Fu (2022); Sun et al. (2022)

To address the issue of CNNs favoring texture over shape, many studies have employed techniques such as human semantic segmentation and human body parts segmentation. Additionally, end-to-end autoencoders have been used in action recognition to tackle this specific challenge Huyghe et al. (2021); Tanigawa et al. (2022); Zolfaghari et al. (2017), as illustrated in Figures 1.5 and 1.6. However, the use of autoencoders comes with several disadvantages, including the complexity of multiple modeling steps (see Figure 1.7), the substantial memory required to store synthetic segmentation frames, and the extended training time due to the high computational cost of first training the autoencoder network, then generating synthetic segmentation images, followed by using an action recognition network for classification. These constraints complicate the development of an efficient end-to-end HAR model. Despite their advantages, creating these synthetic images during the preprocessing phase of HAR poses significant challenges. One of the primary difficulties is the computational cost associated with generating high-quality synthetic images. Segmentation and optical flow algorithms require substantial processing power and time, especially when dealing with large datasets. This computational burden can become a bottleneck, delaying the overall processing pipeline and potentially limiting the scalability of the HAR system Kong and Fu (2022).

1.4.2.3 Multimodal Data Integration

Incorporating data from multiple sensors, such as RGB cameras, depth sensors, and motion capture systems, can enhance action recognition by providing complementary information. However, effectively integrating and synchronizing these multimodal inputs adds complexity to the system design and implementation. Yadav et al. (2021) Each modality may have different characteristics and noise levels, requiring

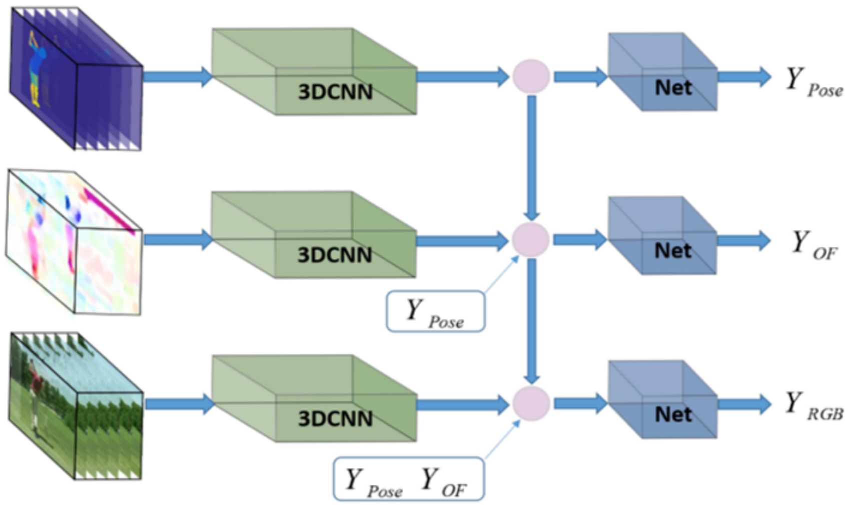


FIGURE 1.6: Artificially created human body part segmentation images (synthetic images) Zolfaghari et al. (2017).

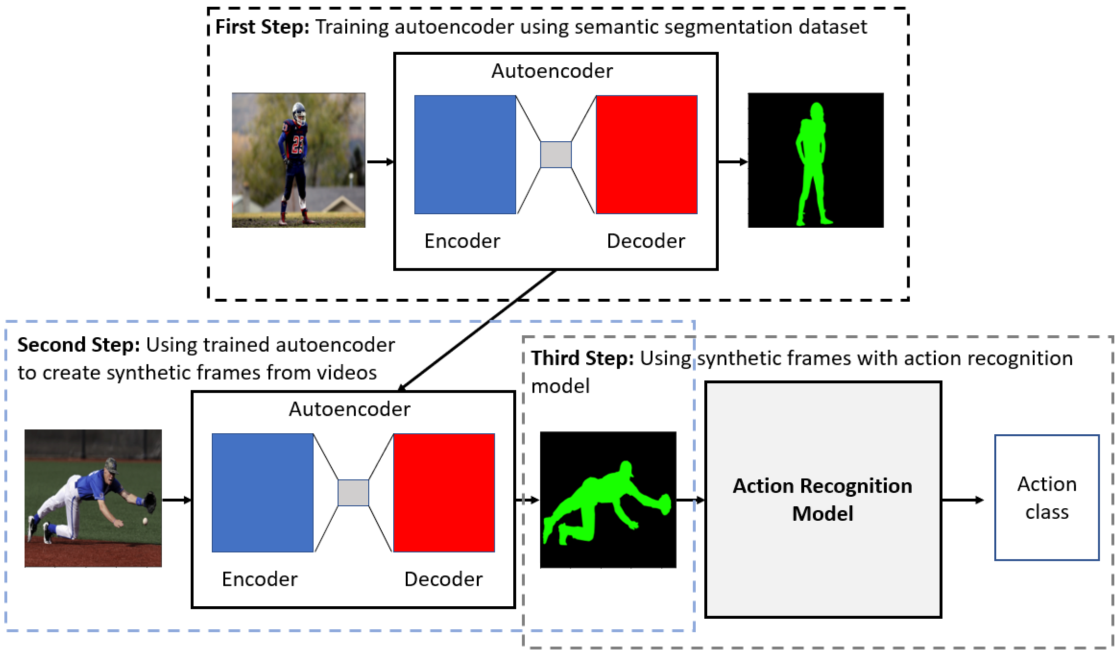

First Step: Training autoencoder using semantic segmentation dataset
Autoencoder
Encoder
Decoder
Second Step: Using trained autoencoder to create synthetic frames from videos
Autoencoder
Encoder
Decoder
Third Step: Using synthetic frames with action recognition model
Action Recognition Model
Action class

FIGURE 1.7: The standard application of autoencoders in HAR tasks. The technique necessitates the use of the autoencoder twice (training and inference) and necessitates a large amount of storage capacity for the synthetic images. Then, the model for action recognition will be trained using the autoencoder-created segmentation images.

sophisticated fusion techniques to leverage the strengths of each sensor while mitigating their weaknesses Sun et al. (2022). For example, combining visual and depth information can improve recognition accuracy in low-light conditions.

In HAR, integrating visual data with other modalities such as audio, depth, or sensor data can enhance recognition accuracy. However, this integration introduces additional challenges. Each modality has different data characteristics, sampling rates, and noise levels, complicating the fusion process Ramanathan et al. (2014). Developing methods that effectively combine these diverse data types while preserving the unique advantages of each modality requires sophisticated modeling and a deep understanding of multimodal data processing.

1.4.2.4 Models' Complexity

The complexity of the models used to process features must be carefully balanced to ensure they are neither too simple nor overly complex. Complex models with numerous parameters can capture intricate patterns but may be computationally expensive and prone to overfitting Ying (2019). Conversely, simpler models may not capture the necessary detail for accurate recognition Bejani and Ghatee (2021). Finding the optimal level of complexity that maximizes accuracy while maintaining efficiency is an ongoing challenge in HAR. For example, deep learning models, while powerful, require significant computational resources and careful tuning to avoid unnecessary complexity.

3D CNNs: Unlike 2D CNNs, which process individual frames independently, 3D CNNs operate on sequences of frames, requiring more complex operations across the spatial and temporal dimensions Tran et al. (2015). This increased complexity results in higher computational costs and greater memory consumption Ji et al. (2012), making it difficult to train and deploy 3D CNNs on standard hardware, let alone on resource-constrained devices like mobile phones or embedded systems.

Training 3D CNNs is a time-consuming process due to the extensive computational requirements Carreira and Zisserman (2017). The need to process and learn from large volumes of video data extends the training time significantly compared to 2D CNNs. This prolonged training period can be a bottleneck in iterative development cycles, where quick experimentation and model refinement are crucial Tran et al. (2018). Additionally, longer training times increase the overall cost and complexity of developing effective HAR systems.

RNNs: and their advanced variants like LSTM or GRU networks are inherently complex due to their sequential processing nature Lipton (2015). This complexity translates into substantial computational requirements and extended training times Greff et al. (2016). Unlike CNNs, which can process data in parallel, RNNs must process each element of a sequence in order, leading to inefficiencies Bai et al. (2018). Training deep RNNs or LSTMs over long sequences, such as those found in HAR tasks, exacerbates this issue, requiring significant computational power and advanced hardware to manage the extensive calculations.

One of the fundamental issues with RNNs is the problem of vanishing and exploding gradients. During back propagation through time (BPTT), gradients used to update the network's weights can become extremely small (vanish) or very large (explode) Bengio et al. (1994). Vanishing gradients cause the network to learn very slowly, while exploding gradients can cause instability and result in large weight updates Pascanu (2013). Although LSTMs were designed to mitigate these problems through their gated structure, they are not completely immune Hochreiter and Schmidhuber (1997). The presence of long sequences in HAR exacerbates these issues, making it challenging to train effective models that capture long-term dependencies.

Many applications of HAR, such as real-time monitoring and surveillance, require low-latency processing Mohammadi (2018). The sequential nature of RNNs and LSTMs poses a challenge for real-time applications because each time step must be processed in order Ordóñez and Roggen (2016). This sequential processing can introduce latency, making it difficult to meet the real-time requirements of certain HAR applications Agarwal and Alam (2020). Optimizing RNNs and LSTMs for real-time performance involves reducing computational complexity and ensuring efficient execution, which is non-trivial.

The complexity of RNNs and LSTMs also impacts their interpretability and explainability. Understanding how these models make decisions is crucial, especially in applications like healthcare and security, where transparency is important Adadi and Berrada (2020). The ″black-box″ nature of these networks makes it challenging to explain their predictions and gain insights into their decision-making process Chen et al. (2020). Developing methods to visualize and interpret the internal workings of RNNs and LSTMs, such as attention mechanisms and saliency maps, is essential to enhance trust and reliability in HAR applications.

1.4.2.5 Overfitting

The development of effective HAR models is significantly hampered by the scarcity of large, labeled datasets. Collecting and annotating video data for HAR is a labor-intensive and costly process, requiring numerous video samples from diverse environments and contexts to ensure robustness and generalization Kumar et al. (2024). This lack of extensive datasets contrasts sharply with fields like image classification, where large-scale datasets such as ImageNet provide a rich source of labeled images for training. Furthermore, the complexity of 3D CNNs compounds these challenges. 3D CNNs, designed to capture spatiotemporal features, require significantly large quantity of data, more computational resources and memory compared to their 2D counterparts Shabanian et al. (2022). This increased complexity makes training and deploying 3D CNNs more difficult, especially on resource-constrained devices.

Moreover, 3D CNNs are prone to overfitting, especially when trained on small datasets Klaiber et al. (2021). Overfitting occurs when a model learns the noise and details in the training data to the extent that it negatively impacts the model's performance on new, unseen data Ying (2019). This issue is particularly prevalent in HAR due to the limited availability of extensive and diverse datasets Yu et al. (2020). Overfitting can lead to poor generalization, where the model performs well on training data but fails to accurately recognize activities in different settings or with different individuals. To mitigate overfitting, techniques such as model complexity reduction, transfer learning, regularization, dropout, and data augmentation are employed Santos and Papa (2022), but these approaches can only partially address the inherent challenges posed by the complexity and data requirements of 3D CNNs.

1.4.2.6 Transfer Learning Limitations Regarding Models' Designs

3D CNNs offer significant benefits for HAR by effectively capturing spatiotemporal features from video data, enabling the modeling of dynamic activities over time Ji et al. (2012). Unlike 2D CNNs, which process individual frames independently, 3D CNNs apply convolutions across both spatial and temporal dimensions, allowing them to understand motion and temporal patterns crucial for accurate activity recognition Tran et al. (2015).

HAR faces significant challenges due to the lack of large, labeled datasets compared to the abundance available for image classification tasks. While datasets like ImageNet provide millions of annotated images for training 2D CNNs, the availability of similarly extensive video datasets for training 3D CNNs in HAR is limited Sargano et al. (2017). Collecting and annotating video data is a labor-intensive and costly process, making it difficult to compile the large datasets needed for robust 3D CNN training. This scarcity hinders the development of highly accurate HAR models and underscores the importance of transfer learning Hoelzemann and Van Laerhoven (2020).

To overcome this limitation, researchers often leverage transfer learning from 2D CNNs pre-trained on vast image classification datasets to 3D CNNs tailored for HAR Carreira and Zisserman (2017); Qiu et al. (2017); Tran et al. (2018); Abdullah et al. (2020). By transferring the rich spatial features learned from extensive 2D image datasets, these models can be fine-tuned to capture temporal dynamics in video data, thus compensating for the lack of large-scale video datasets. This approach not only accelerates the training process but also enhances the model's performance in recognizing complex human activities, thereby bridging the gap between the abundance of image data and the scarcity of annotated video data Ray et al. (2023). While transfer learning from 2D CNNs to 3D CNNs holds promise, it presents several challenges, particularly when transitioning from image classification to HAR. The fundamental differences between handling static images and dynamic video sequences complicate the transfer process.

Learning Temporal Dynamics: A primary challenge in transferring from 2D to 3D CNNs is the necessity to capture temporal dynamics. 2D CNNs are designed to process spatial features in static images, whereas 3D CNNs must also understand motion and changes over time. Adapting spatial features to include temporal information requires significant architectural modifications and additional training Diba et al. (2017). Features learned by 2D CNNs on image classification tasks may not transfer effectively to 3D CNNs for HAR. The pre-trained features are optimized for spatial contexts and may lack the necessary temporal context, resulting in suboptimal performance when applied to video data. Extensive fine-tuning or retraining is often required to adapt these features for effective HAR Leong et al. (2020).

Architectural Modifications and Increased Computational Complexity: Adapting the architecture of 2D CNNs to 3D CNNs involves significant changes, including the modification of convolutional and pooling layers to accommodate the temporal dimension Carreira and Zisserman (2017); Qiu et al. (2017); Tran et al. (2018). These architectural changes must be carefully implemented to avoid degrading the model's performance. Additionally, the transition from 2D to 3D CNNs requires redesigning other network components, such as normalization and activation layers, to handle the increased complexity and data volume. This architectural overhaul often leads to a significant increase in the number of parameters and computational demands.

Converting a 2D CNN into a 3D CNN increases the model's computational load and memory requirements substantially. The added temporal dimension necessitates more parameters and complex operations, which can be computationally prohibitive Liu et al. (2019); Kopuklu et al. (2019). Moreover, the increased computational complexity affects not only the training phase but also the inference phase. Real-time applications, such as surveillance systems or healthcare monitoring, require quick and efficient processing of incoming data streams Li et al. (2024). The heavy computational load of 3D CNNs can result in latency issues, where the time taken to process and recognize activities exceeds acceptable limits for real-time operation. This necessitates the development of optimized algorithms and the use of hardware accelerators, such as GPUs and TPUs, to meet the performance requirements. Techniques like model compression, pruning, and efficient architectural designs are also being explored to mitigate these computational challenges and make 3D CNNs more viable for widespread use in video recognition Sun et al. (2020).

1.5 Technical Background

This section provides a chronological and technical overview of three fundamental types of neural networks: CNNs, RNNs, and Transformers. CNNs, introduced in the late 1980s Fukushima (1980), revolutionized image processing by leveraging local connectivity and shared weights to efficiently detect spatial hierarchies in data. As the field progressed, RNNs emerged in the 1990s, addressing the need for modeling sequential data through their ability to maintain temporal dependencies across sequences. The advent of Transformers in 2017 marked a paradigm shift by utilizing self-attention mechanisms to capture global relationships in data more effectively, thereby enhancing performance in a wide array of tasks beyond sequential data. This background section will delve into the technical intricacies and evolutionary trajectory of these architectures, highlighting their contributions and transitions in the realm of deep learning.

1.5.1 CNNs

The evolution of CNNs has been remarkable since their introduction in the 1980s. Originally, CNNs were designed to process static images, primarily focusing on spatial recognition tasks such as object and pattern recognition. The initial idea was to build layers of convolutional filters that would apply various operations to the image to extract features like edges, textures, and shapes. This structure proved highly effective for tasks like image classification, object detection, image segmentation and more in computer vision.

The Neocognitron Fukushima (1980), developed by Kunihiko Fukushima, presented an early example of neural networks incorporating convolutional operations for image processing, setting the foundations for subsequent progress. Later, Yann LeCun and collaborators introduced LeNet-5 LeCun et al. (1998), a key architecture designed for handwritten digit recognition, showcasing the effectiveness of convolutional layers in pattern recognition tasks. The progress of CNNs reached a turning point in the mid-2010s with the introduction of models like AlexNet Krizhevsky et al. (2012), showcasing their potential in image classification tasks. Alongside architectural innovations, this milestone was achieved thanks to access to large datasets, notably, ImageNet Deng et al. (2009), and computational improvements, including the rise of graphics processing units (GPUs) for parallel computing. Large-scale datasets provided the diversity and complexity necessary for training deep networks, while enhanced computational power accelerated the training of sophisticated CNN architectures.

The architectural enhancements, large datasets, and increased computational capabilities helped CNNs to be a cornerstone in deep learning methodologies, extending their applications beyond image processing to various domains. Notable architectures like VGGNet Simonyan and Zisserman (2014a), distinguished by its uniform design and small convolutional filters, GoogLeNet Szegedy et al. (2015), with its inception modules for capturing features at different scales efficiently, and ResNet He et al. (2016), which introduced residual learning for training very deep networks, have further enriched the landscape of CNNs.

1.5.1.1 Spatio-Temporal CNNs

As CNNs excelled in spatial tasks, researchers began exploring their potential in handling temporal data, such as video and time-series analysis. The challenge was to incorporate the dimension of time into the inherently spatial architecture of CNNs. To address this task, spatio-temporal CNNs were developed. These networks extend traditional CNN architectures by adding a temporal component to analyze dynamic behaviors across time frames. Several approaches have been utilized and main types are as follows.

3D convolution involves extending the 2D kernels to 3D, allowing the network to perform convolution across both spatial and temporal dimensions. This approach is directly applied to video data where the third dimension represents time Hara et al. (2018); Tran et al. (2015). The two-stream CNNs involve running two parallel CNN streams: one for spatial processing of individual frames and another for temporal processing, usually of optical flow, which captures motion between frames Simonyan and Zisserman (2014a); Feichtenhofer et al. (2016). RNNs with CNNs aim to combine CNNs for spatial processing with RNNs like long short-term memory (LSTM) or gated recurrent unit (GRU) to handle temporal dependencies. This hybrid model leverages CNNs' ability to extract spatial features and RNNs' capacity to manage temporal sequences effectively Yue-Hei Ng et al. (2015); Donahue et al. (2015).

1.5.2 From Vanilla RNN to Attention-Based Transformers

This section explores the evolution from RNNs to the Transformers, highlighting the progression in handling time series and sequence data. Initially, RNNs were the go-to deep learning technique for managing temporal tasks, effectively capturing sequential dependencies. However, the development of Transformers marked a significant leap forward, driven by a series of iterative improvements and optimizations that built upon the limitations of RNNs. Transformers, with their focus on NLP, introduced a novel attention mechanism that allows for more efficient and scalable processing of sequential data. By examining the foundational RNN techniques and the subsequent enhancements leading to the Transformer architecture, this section elucidates the transformative journey from traditional RNN models to the sophisticated attention-based frameworks that now dominate the field.

We firstly establish common notations for RNN architectures including vanilla RNNs, LSTM and GRU to streamline discussions in subsequent sections. In these architectures, each iteration involves a cell that sequentially processes an input embedding $\boldsymbol{x}_t \in \mathbb{R}^{n_x}$ and retains information from the previous sequence through the hidden state $\boldsymbol{h}_{t-1} \in \mathbb{R}^{n_h}$ using weight matrices $\boldsymbol{W} \in \mathbb{R}^{n_h \times n_h}$ and $\boldsymbol{U} \in \mathbb{R}^{n_h \times n_x}$. The $\boldsymbol{W}$-like matrices encompass all weights related to hidden-to-hidden connections, while $\boldsymbol{U}$-like matrices encompass all weight matrices related to input-to-hidden connections. Additionally, bias terms are represented by $\boldsymbol{b}$-like vectors. Each cell produces a new hidden state $\boldsymbol{h}_t \in \mathbb{R}^{n_h}$ as its output. More details about symbols and variables used in this section are given in Table 1.1.

1.5.2.1 Vanilla RNNs

Vanilla RNNs Rumelhart et al. (1985); Jordan (1986) lack the presence of a cell state, relying solely on the hidden states as the primary means of memory retention within the RNN framework. The hidden state $\boldsymbol{h}_t$ is subsequently updated and propagated to the subsequent cell, or alternatively, depending on the specific task at hand, it can be employed to generate a prediction. Figure 1.8a illustrates the internal mechanisms of an RNN and a mathematical description of it given as

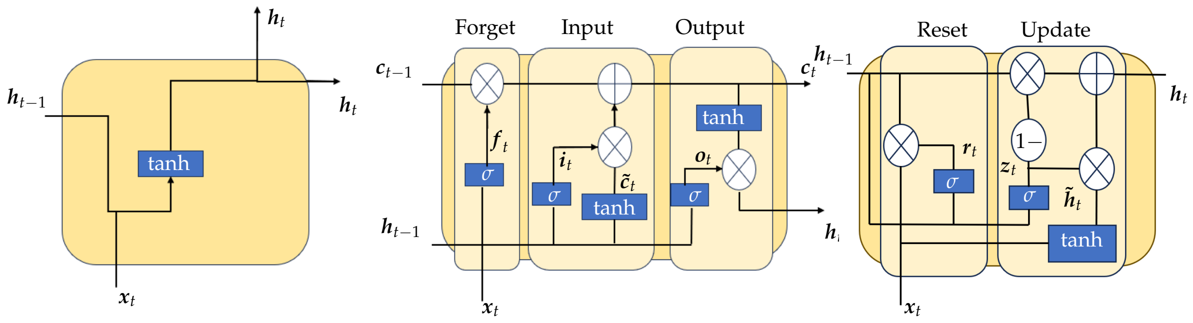
$$\boldsymbol{h}_t = \tanh(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}), \tag{1.1}$$

where tanh is the activation function.

Vanilla RNNs effectively incorporate short-term dependencies of temporal order and past inputs in a meaningful manner. However, they are characterized by certain limitations. Firstly, due to their intrinsic sequential nature, RNNs pose challenges in parallelized computations Graves et al. (2013). Consequently, this limitation can impose restrictions on the overall speed and scalability of the network. Secondly, when processing lengthy sequences, the issue of exploding or vanishing gradients may arise, thereby impeding the stable training of the network Bengio et al. (1994).

Symbol	Definition
$\boldsymbol{x}_t \in \mathbb{R}^{n_x}$	Input embedding at time t
$\boldsymbol{h}_t \in \mathbb{R}^{n_h}$	Hidden state at time t
$\boldsymbol{W} \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for hidden-to-hidden connections
$\boldsymbol{U} \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for input-to-hidden connections
$\boldsymbol{b} \in \mathbb{R}^{n_h}$	Bias vector
$\boldsymbol{i}_t \in \mathbb{R}^{n_h}$	Output of the sigmoid activation function in the input gate in LSTM cell
$\boldsymbol{o}_t \in \mathbb{R}^{n_h}$	Output of the output gate in LSTM cell
$\boldsymbol{c}_t \in \mathbb{R}^{n_h}$	Cell state at time t in LSTM cell
$\tilde{\boldsymbol{c}}_t \in \mathbb{R}^{n_h}$	Candidate cell state at time t in LSTM cell
$\boldsymbol{z}_t \in \mathbb{R}^{n_h}$	Output of the update gate in GRU at time t
$\boldsymbol{r}_t \in \mathbb{R}^{n_h}$	Output of the reset gate in GRU at time t
$\tilde{\boldsymbol{h}}_t \in \mathbb{R}^{n_h}$	Candidate hidden state in GRU at time t
$\boldsymbol{W}_f \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for forget gate in LSTM cell
$\boldsymbol{U}_f \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for forget gate input in LSTM cell
$\boldsymbol{b}_f \in \mathbb{R}^{n_h}$	Bias for forget gate in LSTM cell
$\boldsymbol{W}_i \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for input gate in LSTM cell
$\boldsymbol{U}_i \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for input gate input in LSTM cell
$\boldsymbol{b}_i \in \mathbb{R}^{n_h}$	Bias for input gate in LSTM cell
$\boldsymbol{W}_o \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for output gate in LSTM cell
$\boldsymbol{U}_o \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for output gate input in LSTM cell
$\boldsymbol{b}_o \in \mathbb{R}^{n_h}$	Bias for output gate in LSTM cell
$\boldsymbol{W}_{\tilde{c}} \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for candidate cell state in GRU cell
$\boldsymbol{U}_{\tilde{c}} \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for candidate cell state input in GRU cell
$\boldsymbol{b}_{\tilde{c}} \in \mathbb{R}^{n_h}$	Bias for candidate cell state in GRU cell
$\boldsymbol{W}_z \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for update gate in GRU cell
$\boldsymbol{U}_z \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for update gate input in GRU cell
$\boldsymbol{b}_z \in \mathbb{R}^{n_h}$	Bias for update gate in GRU cell
$\boldsymbol{W}_r \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for reset gate in GRU cell
$\boldsymbol{U}_r \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for reset gate input in GRU cell
$\boldsymbol{b}_r \in \mathbb{R}^{n_h}$	Bias for reset gate in GRU cell
$\boldsymbol{W}_{\tilde{h}} \in \mathbb{R}^{n_h \times n_h}$	Weight matrix for candidate hidden state in GRU cell
$\boldsymbol{U}_{\tilde{h}} \in \mathbb{R}^{n_h \times n_x}$	Weight matrix for candidate hidden state input in GRU cell
$\boldsymbol{b}_{\tilde{h}} \in \mathbb{R}^{n_h}$	Bias for candidate hidden state in GRU cell
$d_k \in \mathbb{N}$	Dimension of the keys
$\boldsymbol{Q} \in \mathbb{R}^{n_q \times d_k}$	A set of query vectors
$\boldsymbol{K} \in \mathbb{R}^{n_k \times d_k}$	A set of key vectors
$\boldsymbol{V} \in \mathbb{R}^{n_v \times d_k}$	A set of value vectors
$\boldsymbol{X} \in \mathbb{R}^{n_x \times d_x}$	Input matrix (sequence of embeddings)
$\boldsymbol{W}^Q \in \mathbb{R}^{d_x \times d_k}$	Weight matrix for queries
$\boldsymbol{W}^K \in \mathbb{R}^{d_x \times d_k}$	Weight matrix for keys
$\boldsymbol{W}^V \in \mathbb{R}^{d_x \times d_k}$	Weight matrix for values
$\boldsymbol{A} \in \mathbb{R}^{n_q \times d_v}$	Attention output
$\boldsymbol{Q}_i \in \mathbb{R}^{n_q \times d_k}$	Query matrix for the i-th attention head
$\boldsymbol{K}_i \in \mathbb{R}^{n_k \times d_k}$	Key matrix for the i-th attention head
$\boldsymbol{V}_i \in \mathbb{R}^{n_v \times d_k}$	Value matrix for the i-th attention head
$\boldsymbol{W}_i^Q \in \mathbb{R}^{d_x \times d_k}$	Weight matrix for queries in the i-th attention head
$\boldsymbol{W}_i^K \in \mathbb{R}^{d_x \times d_k}$	Weight matrix for keys in the i-th attention head
$\boldsymbol{W}_i^V \in \mathbb{R}^{d_x \times d_k}$	Weight matrix for values in the i-th attention head
$\boldsymbol{A}_i \in \mathbb{R}^{n_q \times d_v}$	Attention output for the i-th attention head

TABLE 1.1: List of mathematical symbols and variables in section 1.5.



(A) Vanilla RNN (B) LSTM (C) GRU

FIGURE 1.8: Various types of RNN cells.

1.5.2.2 LSTM

Hochreiter and Schmidhuber (1997) introduced the LSTM cell as a solution to address the issue of long-term dependencies and to mitigate the challenge of interdependencies among successive steps Hochreiter and Schmidhuber (1997). LSTM architecture incorporates a distinct component known as the cell state $\boldsymbol{c}_t \in \mathbb{R}^{n_h}$, illustrated in Figure 1.8b. Analogous to a freeway, this cell state facilitates the smooth flow of information, ensuring that it can readily traverse without undergoing significant alterations.

Gers et al. (2000) made modifications to the initial LSTM architecture by incorporating a forget gate within the cell structure. The mathematical expressions describing this modified LSTM cell are derived from its inner connections. Hence, the LSTM cell can be formally represented based on the depicted interconnections as follows.

- Forget gate decides what information should be thrown away or kept from the cell state with the equation

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{h}_{t-1} + \boldsymbol{U}_f \boldsymbol{x}_t + \boldsymbol{b}_f), \tag{1.2}$$

  where $\boldsymbol{f}_t \in \mathbb{R}^{n_h}$ is the output of the forget gate and σ is the sigmoid activation function.

- Input gate determines which new information is added to the cell state with two activation functions defined as

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{h}_{t-1} + \boldsymbol{U}_i \boldsymbol{x}_t + \boldsymbol{b}_i), \tag{1.3}$$

  where $\boldsymbol{i}_t \in \mathbb{R}^{n_h}$ is the output of the sigmoid activation function; and

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{\tilde{c}} \boldsymbol{h}_{t-1} + \boldsymbol{U}_{\tilde{c}} \boldsymbol{x}_t + \boldsymbol{b}_{\tilde{c}}), \tag{1.4}$$

where $\tilde{\boldsymbol{c}}_t \in \mathbb{R}^{n_h}$ is known as candidate value. After obtaining $\boldsymbol{i}_t$ and $\tilde{\boldsymbol{c}}_t$, we can update the cell state with

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \tag{1.5}$$

where $\boldsymbol{c}_{t-1} \in \mathbb{R}^{n_h}$ is the previous cell state and $\odot$ is the Hadamard operator.

- Output gate determines the next hidden state based on the cell state and output gate's activity

$$\boldsymbol{o_t} = \sigma(\boldsymbol{W}_o \boldsymbol{h}_{t-1} + \boldsymbol{U}_o \boldsymbol{x}_t + \boldsymbol{b}_o), \tag{1.6}$$

where $\boldsymbol{o}_t \in \mathbb{R}^{n_h}$ is the output of the output gate. Finally the updated hidden state,

$$\boldsymbol{h}_t = \tanh(\boldsymbol{c}_t) \odot \boldsymbol{o}_t \tag{1.7}$$

is fed to the next iteration.

To enable selective information retention, LSTM employs three distinct gates. The first gate, known as the forget gate, examines the previous hidden state $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{x}_t$. It generates a vector $\boldsymbol{f}_t$ containing values between 0 and 1, determining the portion of information to discard from the previous cell state $\boldsymbol{c}_{t-1}$. The second gate, referred to as the input gate, follows a similar process to the forget gate. However, instead of discarding information, it utilizes the output $\boldsymbol{i}_t$ to determine the new information to be stored in the cell state based on a candidate cell state $\tilde{\boldsymbol{c}}_t$. Lastly, the output gate employs the output $\boldsymbol{o}_t$ to filter the updated cell state $\boldsymbol{c}_t$, thereby transforming it into the new hidden state $\boldsymbol{h}_t$. The LSTM cell exhibits superior performance in retaining both long-term and short-term memory compared to the vanilla RNN cell. However, this advantage comes at the expense of increased complexity.

1.5.2.3 GRU

The LSTM cell surpasses the learning capability of the conventional recurrent cell, yet the additional number of parameters escalates the computational load. Consequently, to address this concern, Chung et al. (2014) introduced the GRU, see Figure 1.8c. GRU demonstrates comparable performance to LSTM while offering a more computationally efficient design with fewer weights. This is achieved by merging the cell state and the hidden state into "reset state" resulting in a simplified architecture. Furthermore, GRU combines the forget and input gates into an "update gate", contributing to a more streamlined computational process. For further elaboration, GRU cell incorporates two essential gates. The first gate is the reset gate, which examines the previous hidden state $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{x}_t$. It generates a vector $\boldsymbol{r}_t$ containing values between 0 and 1, determining the extent to which past information in $\boldsymbol{h}_{t-1}$ should be disregarded. The second gate is the update gate, which governs the selection of information to either retain or discard when updating the new hidden state $\boldsymbol{h}_t$, based on the value of $\boldsymbol{r}_t$.

Based on the depicted information in Figure 1.8c, the mathematical expressions governing the behavior of the GRU cell can be expressed as follows.

- Update gate decides how much of the past information needs to be passed along with

  $$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z\boldsymbol{h}_{t-1} + \boldsymbol{U}_z\boldsymbol{x}_t + \boldsymbol{b}_z), \tag{1.8}$$

  where $\boldsymbol{z}_t \in \mathbb{R}^{n_h}$ is the output of the update gate. The output of the reset gate $\boldsymbol{r}_t \in \mathbb{R}^{n_h}$ is obtained by

  $$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r\boldsymbol{h}_{t-1} + \boldsymbol{U}_r\boldsymbol{x}_t + \boldsymbol{b}_r). \tag{1.9}$$

  A candidate activation for the subsequent step is

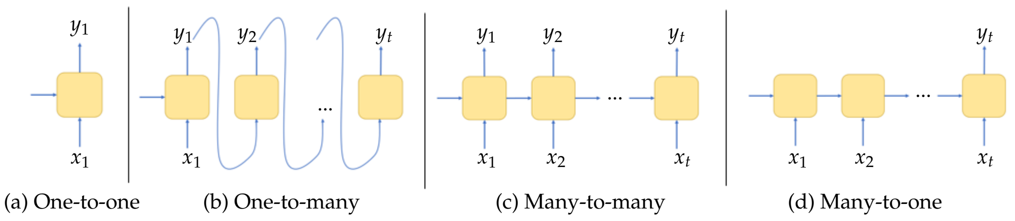
  $$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_{\tilde{h}}(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{U}_{\tilde{h}}\boldsymbol{x}_t + \boldsymbol{b}_{\tilde{h}}) \tag{1.10}$$

  where $\tilde{\boldsymbol{h}}_t \in \mathbb{R}^{n_h}$.

- The final activation is a blend of the previous hidden state and the candidate activation, weighted by the update gate, i.e.,

  $$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t + (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} \tag{1.11}$$

  where $\boldsymbol{h}_t \in \mathbb{R}^{n_h}$ is the updated hidden state. This mechanism allows the GRU to effectively retain or replace old information with new information.



(a) One-to-one (b) One-to-many (c) Many-to-many (d) Many-to-one

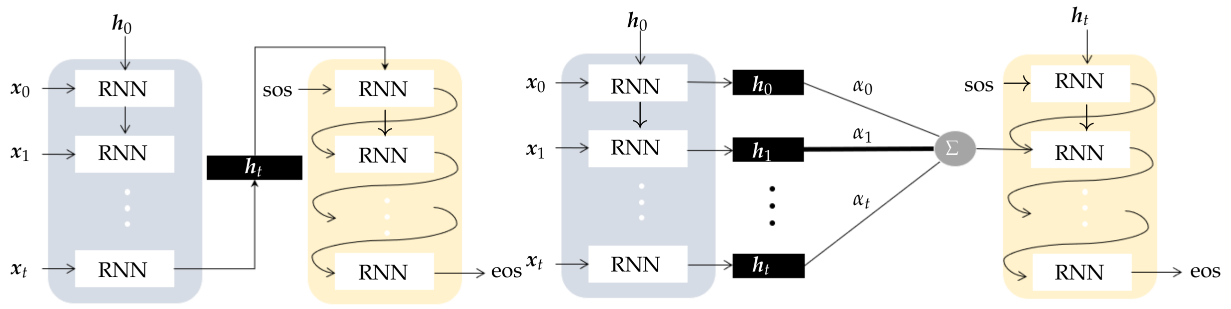
FIGURE 1.9: Types of RNN structures based on input-output pairs.

1.5.2.4 Types of RNNs

RNNs were created with an internal memory mechanism that allows them to store and use information from previous outputs. This unique trait enables RNNs to retain important contextual information over time, enabling reasoned decision-making based on past results. There are four types of popular RNN variants that each serve different purposes across a variety of applications, see Figure 1.9. For simplicity, x_i and y_i respectively represent the input and output with $i = 1, \dots, t$ in Figure 1.9.

The one-to-one is considered the simplest form of RNNs, where a single input corresponds to a single output. It operates with fixed input and output sizes, functioning similarly to a standard neural network. One-to-many represents a specific category of RNNs that is characterized by its ability to produce multiple outputs based on a single input provided to the model. This type of RNN is particularly useful in applications like image captioning, where a fixed input size results in a series of data outputs. Many-to-one RNNs merge a sequence of inputs into a single output through a series of hidden layers that learn relevant features. An illustrative instance of this RNN type is sentiment analysis, where the model analyzes a sequence of text inputs and produces a single output indicating the sentiment expressed in the text.

Many-to-many RNNs are employed to generate a sequence of output data from a sequence of input units. It can be categorized into two subcategories: equal size and unequal size. In the equal size subcategory, the input and output layers have the same size, see many-to-many architecture in Figure 1.9c. Several research efforts have emerged to tackle the limitation of the fixed-size input-output sequences in machine translation tasks, as they fail to adequately represent real-world requirements. The unequal size subcategory can handle different sizes of inputs and outputs. A practical application of the unequal size subcategory can be observed in machine translation. In this scenario, the model generates a sequence of translated text outputs based on a sequence of input sentences. Unequal size subcategory employs an encoder-decoder architecture, where the encoder adopts the many-to-one architecture, and the decoder adopts the one-to-many architecture. One notable contribution in this area was made by Kalchbrenner and Blunsom (2013), who pioneered the approach of mapping the entire input sentence to a vector. This work is related to the study conducted by Cho et al. (2014), although the latter was specifically utilized to refine hypotheses generated by a phrase-based system Sutskever et al. (2014). In this architecture, the encoder component plays a crucial role in transforming the inputs into a singular vector, commonly referred to as the context. This context vector, typically with a length of 256, 512 or 1024, encapsulates all the pertinent information detected by the encoder from the input sentence, which serves as the translation target, see Figure 1.10a. Subsequently, this vector is passed on to the decoder, which generates the corresponding output sequence. It is important to note that both the encoder and decoder components in this architecture are RNNs. Different from Figure 1.10a, Figure 1.10b gives the encoder-decoder architecture with attention which will be introduced in the next section.



(A) Without attention

(B) With attention

FIGURE 1.10: Sequence-to-sequence RNN with and without the attention mechanism. α is the attention weights vector

.

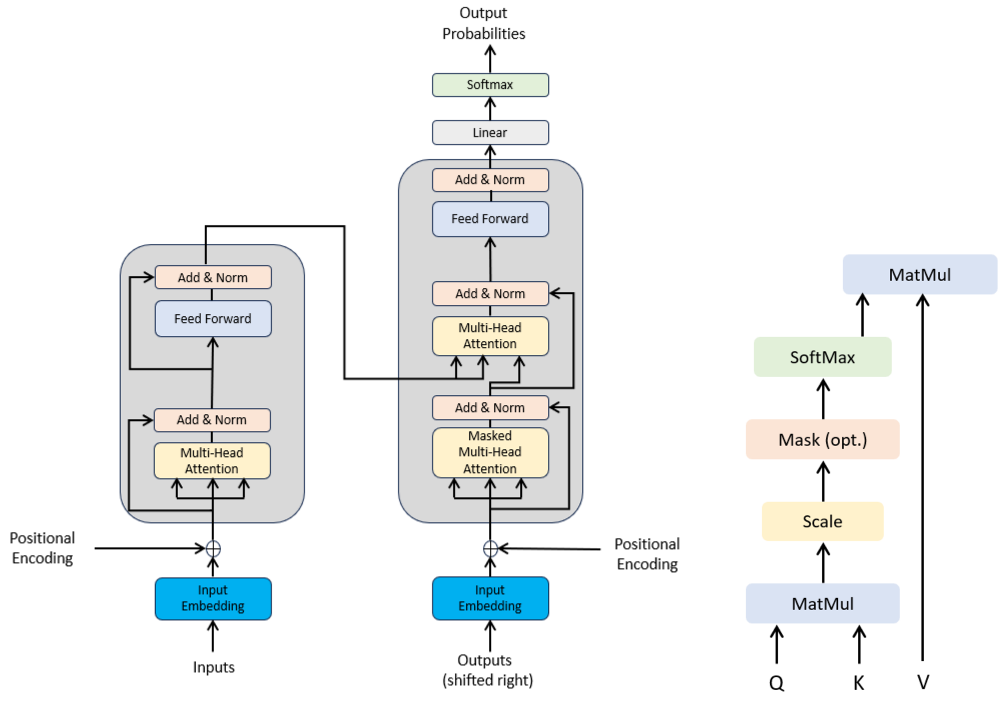
1.5.2.5 Attention

The evolution of attention mechanisms in neural networks represents a significant advancement in the field of deep learning, particularly in tasks related to NLP and machine translation. Initially introduced by Graves (2013), the concept of attention mechanisms was designed to enhance the model's ability to focus on specific parts of the input sequence when generating an output, mimicking the human ability to concentrate on particular aspects of a task. This foundational work laid the groundwork for subsequent developments in attention mechanisms, providing a mechanism for models to dynamically assign importance to different parts of the input data.

Building on Graves' initial concept, Bahdanau et al. (2014) introduced the additive attention mechanism, which was specifically designed to improve machine translation. This approach computes the attention weights through a feed-forward neural network, allowing the model to consider the entire input sequence and determine the relevance of each part when translating a segment. This additive form of attention significantly improved the performance of sequence-to-sequence models by enabling a more nuanced understanding and alignment between the input and output sequences Sutskever et al. (2014). Following this, Luong et al. (2015) proposed the multiplicative attention mechanism, also known as dot-product attention, which simplifies the computation of attention weights by calculating the dot product between the query and all keys. This method not only streamlined the attention mechanism but also offered improvements in computational efficiency and performance in various NLP tasks, marking a pivotal moment in the evolution of attention mechanisms from their inception to more sophisticated and efficient variants.

The central idea of the attention mechanism is to shift focus from the task of learning a single vector representation for each sentence. Instead, it adopts a strategy of selectively attending to particular input vectors in the input sequence, guided by assigned attention weights. This strategy enables the model to dynamically allocate its attention resources to the most pertinent segments of the sequence, thereby improving its capacity to process and comprehend the information more efficiently Brauwers and Frasincar (2021).

One possible explanation for the improvement is that the attention layer created memories associated with the context pattern rather than memories associated with the input itself, relieving pressure on the RNN model structure's weights and causing the model memory to be devoted to remembering the input rather than the context pattern Hu et al. (2018).



(A) Transformer

(B) Self-Attention

FIGURE 1.11: Transformer architecture and its self-attention mechanism (adapted from Vaswani et al. (2017)).

1.5.2.6 Self-Attention

To this point, attention mechanisms in sequence-transformation models have primarily relied on complex RNNs, featuring an encoder and a decoder, the most successful models in language translation yet. However, Vaswani et al. (2017) introduced a simple network architecture known as the Transformer, see Figure 1.11, which exclusively utilized attention mechanism, eliminating the need for RNNs. They introduced a novel attention mechanism called self-attention, which is also known as KQV-attention (Key, Query, and Value). This attention mechanism subsequently gained prominence as a central component within the Transformer architecture. The attention mechanism stands out due to its ability to provide Transformers with an extensive long-term memory. In the Transformer model, it becomes possible to focus on all previously generated tokens.

The embedding layer in a Transformer model is the initial stage where input tokens are transformed into dense vectors, capturing semantic information about each token's meaning and context within the text. These embeddings serve as the foundation for subsequent layers to process and understand the relationships between words in the input sequence Dar et al. (2022).

Self-attention is a mechanism that allows an input sequence to process itself in a way that each position in the sequence can attend to all positions within the same sequence. This mechanism is a cornerstone of the Transformer architecture, which has revolutionized NLP and beyond by enabling models to efficiently handle sequences of data with complex dependencies. The purpose of self-attention is to compute a representation of each element in a sequence by considering the entire sequence, thereby capturing the contextual relationships between elements regardless of their positional distance from each other. This ability to capture both local and global dependencies makes self-attention particularly powerful for tasks such as machine translation, text summarization, and sequence prediction, where understanding the context and the relationship between words or elements in a sequence is crucial Vaswani et al. (2017).

The mathematical formulation of self-attention involves several key steps. First, a set of query vectors $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{Q}$, a set of key vectors $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^{K}$, and a set of value vectors $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^{V}$ are calculated through linear transformations of the input sequence, where $\boldsymbol{X}$ is the input matrix representing embeddings of tokens in a sequence, and $\boldsymbol{W}^{Q}$,$\boldsymbol{W}^{K}$, and $\boldsymbol{W}^{V}$ are weight matrices for queries, keys, and values, respectively. The attention scores are then calculated by taking the dot product of the query vector with all key vectors, followed by scaling the result by the inverse square root of the dimension of the keys (say $\sqrt{d_k}$) to avoid overly large values. These scores are then passed through a softmax function to obtain the attention weights, which represent the importance of each element's contribution to the output. Finally, the output say $\boldsymbol{A}$ is computed as a weighted sum of the value vectors, i.e.,

$$\boldsymbol{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}})\boldsymbol{V}. \tag{1.12}$$

This process allows the model to dynamically focus on different parts of the input sequence, enabling the extraction of rich contextual information from the sequence.

1.5.2.7 Multi-Head-Attention

Multi-head attention is an extension of the self-attention mechanism designed to allow the model to jointly attend the information from different representation subspaces at different positions Vaswani et al. (2017). Instead of performing a single attention function, it runs the attention mechanism multiple times in parallel. The outputs of these independent attention computations are then concatenated and linearly transformed into the expected dimension. The mathematical formulation of the multi-head attention can be described in the following steps. First, for the *i*-th self-attention head, find

$$\boldsymbol{Q}_i = \boldsymbol{X}\boldsymbol{W}_i^Q, \quad \boldsymbol{K}_i = \boldsymbol{X}\boldsymbol{W}_i^K, \quad \boldsymbol{V}_i = \boldsymbol{X}\boldsymbol{W}_i^V, \tag{1.13}$$

and then compute

$$\boldsymbol{A}_i(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \text{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\top}{\sqrt{d_k}}\right)\boldsymbol{V}_i. \tag{1.14}$$

The multi-head attention is obtained by concatenating all $\boldsymbol{A}_i(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i)$.

The multi-head attention mechanism enables the model to capture different types of information from different positions of the input sequence. By processing the sequence through multiple attention "heads", the model can focus on different aspects of the sequence, such as syntactic and semantic features, simultaneously. This capability enhances the model's ability to understand and represent complex data, making multi-head attention a powerful component of Transformer-based architectures Devlin et al. (2019).

1.5.3 From Transformer to Vision Transformer

The journey from the inception of the Transformer model to the development of the ViT marks a pivotal advancement in deep learning, showcasing the adaptability of models initially designed for sequence data processing to the realm of image analysis. This transition underscores a significant shift in approach, from conventional image processing techniques to more sophisticated sequence-based methodologies.

Introduced by Vaswani et al. (2017) through the seminal paper "Attention Is All You Need", the Transformer model revolutionized NLP by leveraging self-attention mechanisms. This innovation allowed for the processing of sequences of data without the reliance on recurrent layers, facilitating unprecedented parallelization and significantly reducing training times for large datasets. The Transformer's success in NLP sparked curiosity about its potential applicability across different types of data, including images, setting the stage for a transformative adaptation.

The adaptation of Transformers for image data pivoted on a novel concept: treating images not as traditional 2D arrays of pixels but as sequences of smaller and discrete image patches. This approach, however, faced computational challenges due to the self-attention mechanism's quadratic complexity with respect to input length. The breakthrough came with the introduction of the ViT by Dosovitskiy et al. (2020), which applied the Transformer architecture directly to images, see Figure 1.12. By dividing an image into fixed-size patches and processing these patches as if they were tokens in a text sequence, ViT was able to capture complex relationships between different parts of an image using the Transformer's encoder.

The operational mechanics of ViT begin with the division of an input image into fixed-size patches, each of which is flattened and linearly transformed into a vector, effectively converting the 2D image into a 1D sequence of embeddings. To account for the lack of inherent positional awareness within the Transformer architecture, positional embeddings are added to these patch embeddings, ensuring the model retains spatial information. The sequence of embeddings is then processed through the Transformer encoder, which consists of layers of multi-head self-attention and feed-forward neural networks, allowing the model to dynamically weigh the importance of each patch relative to others for a given task.

For tasks like image classification, the output from the Transformer encoder is passed through a classification head, often utilizing a learnable "class token" appended to the sequence of patch embeddings for this purpose. The model is trained on large datasets using backpropagation and, during inference, processes images through these steps to predict their classes.

The ViT not only demonstrates exceptional performance on image classification tasks, often surpassing CNNs when trained on extensive datasets, but also highlights the Transformer architecture's capacity to capture the global context within images. Despite its advantages, ViT's reliance on substantial computational resources for training and its need for large datasets to achieve optimal performance present challenges. Nonetheless, the development of ViT signifies a significant milestone in the application of sequence processing models to the field of computer vision, opening new avenues for research and practical applications.

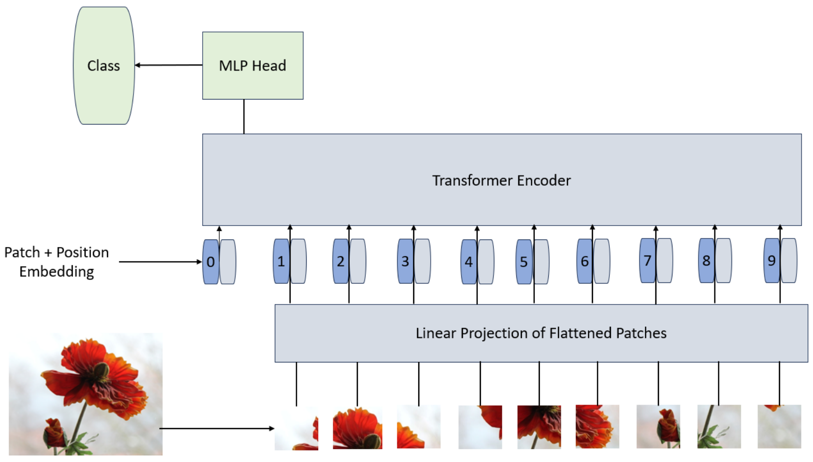


FIGURE 1.12: The ViT architecture (adapted from Dosovitskiy et al. (2020)).

The original ViT, designed for static image processing, divides images into patches and interprets these as sequences, leveraging the Transformer's self-attention mechanism to understand complex spatial relationships. Extending this model to action recognition involves adapting it to analyze video frames sequentially to capture both spatial and temporal relationships. Several works attempted to adapt ViT in action recognition task using different methods as below.

Temporal dimension integration. The integration of the temporal dimension is a fundamental step in adapting ViT for action recognition. Traditional ViT models process images as a series of patches, treating them essentially as sequences for the self-attention mechanism to analyze spatial relationships. By extending this concept to include the temporal dimension, the models can now treat videos as sequences of frame patches over time. This allows the models to capture the evolution of actions across frames. The work by Bertasius et al. (2021) highlights the potential of incorporating temporal information into Transformers, marking a significant advancement in video analysis capabilities.

Spatio-temporal embeddings. To effectively capture the dynamics of actions within videos, adapted ViT models generate spatio-temporal embeddings. This involves extending the traditional positional embeddings used in ViTs to also include temporal positions, thereby creating embeddings that account for both spatial and temporal information within video sequences. The discussion by Arnab et al. (2021) on the creation of these spatio-temporal embeddings showcases the method's effectiveness in enhancing the model's understanding of action dynamics across both space and time.

Multi-head self-attention across time. The extension of self-attention mechanisms to analyze relationships between patches not just within individual frames but also across different frames is crucial for recognizing actions over time. This approach enables the model to identify relevant features and changes across the video sequences, facilitating a deeper understanding of motion and the progression of actions. The exploration by Bertasius et al. (2021) of this concept demonstrates how Transformers can be effectively adapted to capture the temporal dynamics of actions, a key aspect of video analysis.

1.6 Research Questions

- What is the impact of shape features on image classification and semantic segmentation, and how can data augmentation be utilized to investigate this effectiveness?
- What is known about the data augmentation techniques in image classification and semantic segmentation, and what further research is needed?
- Can single modality-based (i.e. RGB modality) HAR model outperform multi modality-based models?
- Does Image segmentation tends to segment based on texture or shape features?
- How to use image segmentation in shape-transfer learning instead of texture-transfer learning?
- How to instruct the CNNs model to learn shape over texture features?
- How to create a HAR model that provide the transfer learning ability with being accurate, less complex and fast learning?
- What is known about the attention-based architectures in HAR, and what further research is needed?
- What is the impact of designing a HAR model that integrates the capabilities of CNNs and ViTs?

1.7 Research Objectives

The aim of this research thesis is to support the academic community by advancing the field of HAR through the development and evaluation of novel computational methodologies. By leveraging advanced machine learning techniques and state-of-the-art deep learning architectures, this study seeks to enhance the accuracy and robustness of HAR systems. Specifically, the thesis focuses on addressing existing challenges in the recognition such as overfitting, temporal analysis, feature extraction and models' complexity. By systematically reviewing computer vision enhancing techniques (e.g. data augmentation), analyzing deep learning architectures (e.g. 2D

CNNs, 3D CNNs, RNNs, transformers, Vits), investigating computer vision tasks (e.g. image classification and semantic segmentation) and their application across various images and videos datasets and scenarios, the research endeavors to provide comprehensive insights into the most effective practices for training and deploying HAR systems.

The First Objective: The research aims to review and compare data augmentation techniques (i.e. traditional and deep learning techniques) utilized in computer vision tasks, particularly focusing on image classification and semantic segmentation, respectively.

The Second Objective: The research aims to utilize data augmentation techniques as a methodological tool to examine the effectiveness of various visual features, specifically texture and shape, on the performance of CNNs when employed in computer vision tasks such as image classification and semantic segmentation.

The Third Objective: The research aims to employ the computer vision technique of semantic segmentation as a task to facilitate the transfer learning of shape feature. This approach instructs CNN models to prioritize learning shape features over texture features.

The Fourth Objective: The research aims to develop models capable of directly extracting shape features from RGB frames without the need for synthetic images of human semantic segmentation.

The Fifth Objective: The research aims to develop a HAR model that combine the advantages of two HAR common architectures (i.e. 2D CNN-RNN and 3D CNNs) and simultaneously avoid their drawbacks.

The Sixth Objective: The research aims to systematically review various architectures used for sequential data processing in HAR, starting from simple RNNs to advanced Vision Transformers. This study aims to understand the evolution of these models, identifying their inherent difficulties and specific applications within HAR. By reviewing the applications, benefits, and difficulties of these architectures, this exploration aims to provide comprehensive insights into the most effective sequential data processing models for enhancing HAR systems.

To achieve the research objectives, three chapters (i.e., previously papers) have been employed. Chapter 2 addresses the first and second objectives. Chapter 3 fulfills the third, fourth, and fifth objectives. Lastly, Chapter 4 meets the sixth objective.

1.8 Conducted Studies

1.8.1 The First Work

Chapter 2, aims to address the significant challenges posed by the requirement of large datasets for training deep learning models, particularly CNNs. The study provides a comprehensive survey of existing data augmentation techniques used in computer vision tasks, such as image classification and segmentation. Additionally, the Work offers a detailed taxonomy of data augmentation methods, categorizing them based on their nature and application. The Work emphasize the importance of data augmentation as a method to artificially expand and diversify training datasets, thus mitigating issues related to data scarcity and overfitting, which are common hurdles in developing robust and generalizable models.

Chapter 2 introduces a novel data augmentation strategy named random local rotation (RLR). This technique involves randomly selecting circular regions within an image and rotating them by random angles. The primary purpose of RLR is to distort the shape features in the image while keeping the texture features intact. The goal is to retain the core information of the image while enhancing its diversity without introducing artifacts commonly associated with traditional rotation techniques, such as black boundaries. The study details the implementation of RLR and compares its performance with traditional data augmentation methods through extensive experimental evaluations.

Thus, the first work endeavours to answer the following two main research questions:

- What is known about the data augmentation techniques in image classification and semantic segmentation, and what further research is needed?
- What is the impact of shape features on image classification and semantic segmentation, and how can data augmentation be utilized to investigate this effectiveness?

Expected Results: Chapter 2 anticipates that the proposed RLR method will outperform traditional data augmentation techniques in both image classification and segmentation tasks. The expected results include a significant reduction in overfitting and an improvement in model generalization. By introducing local variations while maintaining the overall integrity of the image, specifically, preserving the texture features RLR is expected to enhance the robustness of CNNs, making them more effective in real-world applications where data variability is high.

This approach is designed to demonstrate the effectiveness of diversifying texture features while also measuring the impact of shape feature distortion on computer vision tasks. Through a series of experiments, the study revealed that the shape distortion introduced by RLR does not significantly affect, and may even enhance, the accuracy of image classification tasks. This finding underscores the reliance of image classification on texture features. Conversely, RLR has a noticeable negative impact on semantic segmentation tasks, particularly those heavily dependent on shape segmentation, such as human semantic segmentation.

Expected Contribution: This research contributes to the field of computer vision and deep learning in several significant ways. Firstly, it offers a detailed survey of existing data augmentation techniques, providing a valuable resource for researchers and practitioners. The survey categorizes and evaluates different methods, highlighting their strengths and weaknesses in various contexts, thereby aiding in the selection of appropriate techniques for specific applications.

Secondly, the introduction of the RLR technique represents a novel contribution to data augmentation strategies. By focusing on local transformations that distort shape features while preserving texture features, the method addresses common issues associated with global transformations, such as boundary artifacts, and improves the diversity of training datasets without losing critical information. This approach not only enhances model performance but also provides a new direction for future research in data augmentation.

Furthermore, the experimental results and comparisons presented in Chapter 2 are expected to establish RLR as a reliable and effective data augmentation technique. The study anticipates demonstrating that image classification performance remains stable or improves despite shape distortion, highlighting the critical role of texture features in this task. Conversely, the research is expected to show that semantic segmentation tasks, particularly those reliant on shape recognition, experience a decline in performance with shape distortion. This finding will underscore the importance of shape features in segmentation tasks and demonstrate the nuanced impact of data augmentation techniques across different types of computer vision applications.

In summary, Chapter 2 is poised to make a substantial impact on the field by providing a thorough survey of current practices, introducing a novel augmentation technique, and offering empirical evidence of its benefits. The contributions are expected to enhance the understanding and implementation of data augmentation in deep learning, ultimately leading to more robust and generalizable models capable of performing well in diverse and challenging real-world scenarios.

1.8.2 The Second Work

Chapter 3 introduces an innovative deep learning architecture designed to address the complexities and limitations of current HAR models. HAR is a significant area of research in computer vision, with applications ranging from surveillance to human-computer interaction. Traditional HAR models often struggle with complex structures, lengthy training times, and the need for extensive datasets. TransNet aims to simplify these models by decomposing the traditional 3D CNNs into more manageable 2D and 1D CNN components. This decomposition allows the model to efficiently extract spatial features with 2D-CNNs and temporal patterns with 1D-CNNs. By leveraging the principles of transfer learning, TransNet is designed to be compatible with any state-of-the-art pretrained 2D-CNN models, enhancing its flexibility and efficiency.

Thus, Chapter 3 endeavours to answer the following five main research questions:

- Can single modality-based (i.e. RGB modality) HAR model outperform multi modality-based models?
- Does Image segmentation tends to segment based on texture or shape features?
- How to use image segmentation in shape-transfer learning instead of texture-transfer learning?
- How to instruct the CNNs model to learn shape over texture features?
- How to create a HAR model that provide the transfer learning ability with being accurate, less complex and fast learning?

Expected Results: TransNet is anticipated to demonstrate superior performance compared to existing HAR models in several key areas. Firstly, due to its simplified architecture, TransNet is expected to significantly reduce training times while maintaining high accuracy. The use of transfer learning is expected to enhance the model's ability to generalize from smaller datasets, addressing the common issue of data scarcity in HAR. The experimental results are predicted to show that TransNet, when integrated with pretrained models such as MobileNet or VGG16, will outperform traditional models both in terms of classification accuracy and processing speed. Additionally, the novel approach of using autoencoders in TransNet+ is expected to further improve the model's ability to capture relevant features for HAR, particularly in tasks requiring detailed human shape recognition.

Through a series of rigorous experiments, the study aims to validate these expectations. The performance of TransNet is evaluated on several benchmark datasets, including KTH, UCF101, and HMDB51, which are standard in the field of HAR. The datasets selected for evaluation—KTH, UCF101, and HMDB51—are particularly challenging due to their relatively small sizes, making them ideal for testing the validity of the transfer learning capability of the TransNet model. These datasets provide a stringent testbed for assessing the model's performance in real-world scenarios where data is limited. The results are anticipated to demonstrate that TransNet not only matches but exceeds the performance of state-of-the-art HAR models, particularly in terms of flexibility, model complexity, training speed, and classification accuracy.

Expected Contributions: The contributions of Chapter 3 are multifaceted and significant to the field of computer vision and HAR. Firstly, the introduction of TransNet provides a new, efficient architecture that simplifies the complex structures of existing HAR models. By decomposing 3D-CNNs into 2D and 1D components, Chapter 3 offers a novel way to manage spatial and temporal features separately yet effectively, leading to reduced computational complexity and faster training times.

Secondly, the utilization of transfer learning in TransNet is a substantial advancement. By making the model compatible with pretrained 2D-CNNs, TransNet leverages existing advancements in other domains (i.e. image classification and human semantic segmentation) of computer vision, thus enhancing the efficiency and effectiveness of HAR tasks. This approach not only reduces the need for large training datasets but also improves the generalization capabilities of the model, making it more robust in diverse and real-world applications.

Furthermore, Chapter 3 introduces TransNet+, which incorporates autoencoders to enhance the 2D component of the model. This strategy allows for the extraction of specific features, such as human shapes, by pretraining the autoencoder on related tasks like human semantic segmentation. This innovative use of autoencoders is expected to significantly improve the model's performance in recognizing human actions, particularly in scenarios with complex backgrounds and occlusions.

Lastly, the extensive experimental validation provided in Chapter 3 is a key contribution. By comparing TransNet with current state-of-the-art models across multiple benchmark datasets, Chapter 3 not only demonstrates the superior performance of the proposed architecture but also provides a comprehensive evaluation framework for future research. The findings are expected to offer valuable insights into the application of transfer learning and model decomposition in HAR, paving the way for more efficient and effective models in this field.

In summary, Chapter 3 makes substantial contributions by presenting a simplified yet powerful HAR model, leveraging transfer learning to enhance performance, introducing innovative use of autoencoders, and providing thorough experimental validation. These advancements are poised to significantly impact the development of future HAR models, making them more accessible, efficient, and effective for a wide range of applications.

1.8.3 The Third Work

Chapter 4 provides a comprehensive survey of the existing methodologies in HAR, focusing particularly on CNNs, RNNs, and ViTs. Chapter 4 explores the evolution of these models, highlighting the emerging trend of attention-based architectures, starting from simple RNNs and progressing to the sophisticated ViT architecture. By tracing this evolution, Chapter 4 shed light on how these models have increasingly improved the ability to capture both spatial and temporal dynamics in video data, which is crucial for accurately recognizing human actions. Recognizing the growing importance and effectiveness of hybrid models that integrate the strengths of different neural network architectures, Chapter 4 proposes a novel hybrid model combining CNNs and ViTs. This model seeks to leverage the spatial feature extraction capabilities of CNNs and the global context awareness provided by ViTs, aiming to enhance the performance and efficiency of HAR systems.

Thus, Chapter 4 endeavours to answer the following two main research questions:

- What is known about the attention-based architectures in HAR, and what further research is needed?
- What is the impact of designing a HAR model that integrates the capabilities of CNNs and ViTs?

Expected Results: Chapter 4 anticipates that the proposed hybrid model will outperform traditional HAR models that rely solely on CNNs, RNNs, or ViTs. The hybrid approach is expected to provide superior accuracy in action recognition tasks by effectively capturing both local and global features within video sequences. The inclusion of ViTs is anticipated to significantly improve the model's ability to handle long-range dependencies and complex temporal relationships, which are crucial in accurately recognizing and categorizing human actions. The expected results from the experiments include improved recognition accuracy on the KTH dataset, demonstrating the model's robustness across different scenarios.

Expected Contributions: Chapter 4 makes several key contributions to the field of HAR. Firstly, it provides an exhaustive review of the current state-of-the-art models, including CNNs, RNNs, and ViTs, and critically examines their evolution, strengths, and limitations in the context of action recognition. This survey serves as a valuable resource for researchers, offering a clear understanding of the progress and challenges in HAR, and setting the stage for future advancements.

Secondly, the introduction of the hybrid CNN-ViT model represents a significant innovation in HAR. By combining the complementary strengths of CNNs and ViTs, the model offers a new approach to handling the complex spatial-temporal dynamics inherent in video data. This hybrid architecture not only enhances the interpretability and robustness of HAR systems but also provides a scalable solution that can be adapted to different tasks and datasets.

1.8.4 Publications

- Published
  - Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging*, 9 (2):46, 2023.
  - Khaled Alomar and Xiaohao Cai. Transnet: A transfer learning-based network for human action recognition. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1825–1832. IEEE, 2023.
- Under Review/Revision
  - Khaled Alomar, Halil Ibrahim Aysel and Xiaohao Cai. RNNs, CNNs and Transformers in Human Action Recognition: A Survey and a Hybrid Model. Under review in *Artificial Intelligence Review*.

Chapter 2

Data Augmentation in Classification and Segmentation: A Survey and New Strategies

In the past decade, deep neural networks, particularly convolutional neural networks, have revolutionised computer vision. However, all deep learning models may require a large amount of data so as to achieve satisfying results. Unfortunately, the availability of sufficient amounts of data for real-world problems is not always possible, and it is well recognised that a paucity of data easily results in overfitting. This issue may be addressed through several approaches, one of which is data augmentation. In this chapter, we survey the existing data augmentation techniques in computer vision tasks, including segmentation and classification, and suggest new strategies. In particular, we introduce a way of implementing data augmentation by using local information in images. We propose a parameter-free and easy to implement strategy, the random local rotation strategy, which involves randomly selecting the location and size of circular regions in the image and rotating them with random angles. It can be used as an alternative to the traditional rotation strategy, which generally suffers from irregular image boundaries. It can also complement other techniques in data augmentation. Extensive experimental results and comparisons demonstrated that the new strategy consistently outperformed its traditional counterparts in, for example, image classification.

2.1 Introduction

Deep neural networks, like convolutional neural networks (CNNs), have been used in computer vision with numerous research applications, such as action recognition

Yudistira and Kurita (2017); Papakostas et al. (2016), object detection and localisation Milyaev and Laptev (2017); Zhou et al. (2017b), face recognition Ranjan et al. (2018), and image characterisation Druzhkov and Kustikova (2016). They have achieved superior performance against conventional approaches in many challenging computer vision tasks Rajnoha et al. (2018). Nevertheless, their shortcomings, such as large-scale data requirements, long training time, overfitting, and performance slumps upon data scarcity, may hinder their generalisation and effectiveness Zhong et al. (2020); Joshi et al. (2019).

The fruitful results presented by the CNN models encourage researchers to pursue higher accuracy models. These results are generally achieved by building more complex architectures Shorten and Khoshgoftaar (2019). Note that model complexity is often described by the number of trainable parameters. The more trainable parameters a model has, the more complex it is. More specifically, model complexity may also be defined in terms of the number of layers (i.e., non-linearity) and the number of neurons (e.g., filters) in individual layers. On the other hand, in supervised learning, data complexity can be determined according to the inter-class multiplicity (i.e., different classes) in addition to the intra-class differences. In general, the complex of the data and the model needed is proportional. If the training data is insufficient, complex models may be susceptible to the issue of memorising the training data. It is also well known that deep neural networks prevail partly because of the availability of high volume data. The networks can easily memorise data points due to their complex structure. However, the increasing complexity of the model architectures with insufficient data could exacerbate the shortcomings of CNN models Zoph et al. (2020). One of the most apparent issues when adopting complex CNN models is the overfitting problem Brownlee (2018), which can be described as the performance difference between the training and validation/test stages, where the model loses its ability to generalise. Overfitting generally occurs when a model is either too complex for the data or the data itself is insufficient Guo et al. (2018). Figure 2.1 shows an example of the loss curve of an overfit model. Although the training accuracy and validation accuracy improved concurrently during the early stages of training, they diverged after a certain point, where the model started losing its generalisation ability Brownlee (2018). Strategies like reducing the model complexity, applying regularisation, and/or acquiring more extensive data volumes have been considered to mitigate the overfitting issue in deep learning models, see Figure 2.2.

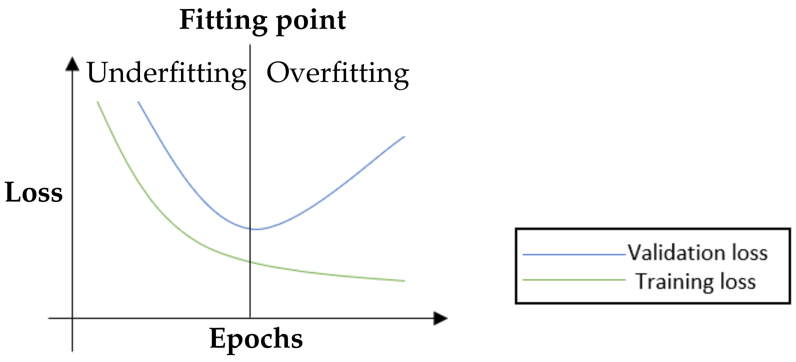


FIGURE 2.1: An illustration of the training and validation loss curves. Training and validation losses decrease simultaneously until the fitting point. After that, the validation loss begins to rise while the training loss is still decreasing, i.e., the so-called overfitting. Overfitting is associated with good performance on the training data but poor generalisation to the validation/test data (cf. underfitting is associated with poor performance on the training data and poor generalisation to the validation data) Brownlee (2018).

Regularisation techniques are implemented at the model architectural level Wan et al. (2013); Kang et al. (2017), such as dropout Srivastava et al. (2014), ridge regression (ℓ_2 regularisation) Farebrother (1976), and Lasso regression (ℓ_1 regularisation) Ranstam and Cook (2018). The main objective of these techniques is to reduce the complexity of a neural network model during training, which is considered the main reason behind overfitting, especially when the model is trained on small datasets. Other techniques, like batch normalisation and transfer learning, may speed up the training process and also have an impact on preventing overfitting Ioffe and Szegedy (2015); Pan and Yang (2009). These techniques could be regarded as byproducts of the constant competition in the pursuit of higher performance by innovating new complex deep neural architectures, such as VGG-16 Simonyan and Zisserman (2014b), ResNet He et al. (2016), Inception-V3 Szegedy et al. (2016) and DenseNet Huang et al. (2017). These models, in fact, aim to achieve higher accuracy on large datasets like Imagenet Deng et al. (2009), which has over 14 million images Deng et al. (2009). However, when applying these models to small-scale applications with small datasets, they usually suffer from poor generalisation and overfitting, indicating the necessity of developing methods to reduce their complexity.

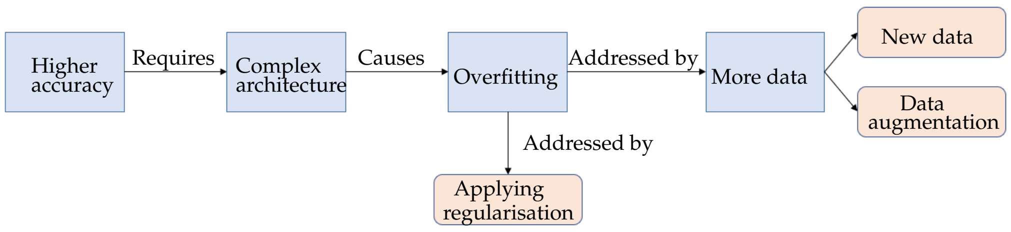


FIGURE 2.2: Diagram illustrating the overfitting problem and its well-known solutions.

Data augmentation methodology encompasses a broader range of techniques that function at the data level, rather than at the model architectural level. It can help deep learning models perform better by artificially creating different and diverse samples with balanced classes for the training dataset. When the dataset is sufficient in terms of quantity and quality, a deep learning model performs better and more accurately. In other words, the training data must fulfil two requirements, i.e., adequate diversity and size, both of which can be achieved by data augmentation Yang et al. (2022c).

Data augmentation can be categorised based on the intended purpose of applying it (i.e., increasing training dataset size and/or diversity), or it can be categorised based on the problems. The following are examples of the latter: the random erasing technique was proposed to address the occlusion problem Zhong et al. (2020); rotation and flipping were supposed to partially resolve the viewpoint problem Divon and Tal (2018); Ning et al. (2020); Massa et al. (2016); brightness was used to address the change in lighting Liu et al. (2021); and cropping and zooming were used to address the scaling and background issues. In particular, the most popular categorisation of data augmentation divides it into deep learning-based data augmentation and traditional data augmentation Shorten and Khoshgoftaar (2019), which is further divided into geometric, photometric, and noise data augmentation, see Figure 2.3. For reviews on the deep learning approaches for data augmentation, see e.g., Chlap et al. (2021); Lindner et al. (2019).

Several studies evaluating the efficacy of data augmentation have utilised standard academic image datasets to assess results. For example, MNIST, CIFAR-10, CIFAR-100 and ImageNet are four commonly used datasets Cubuk et al. (2020); Shorten and Khoshgoftaar (2019); Shijie et al. (2017); Lee et al. (2019). Note that some of these datasets, especially ImageNet, are considered “big data” Denton et al. (2021) and may not require data augmentation techniques to further increase their size. To simulate data scarcity challenges, many experiments testing data augmentation techniques limit themselves to small subsets of the original large datasets Shijie et al. (2017). It is worth emphasising that data augmentation techniques may also be used to improve the data diversity, except for the data quantity.

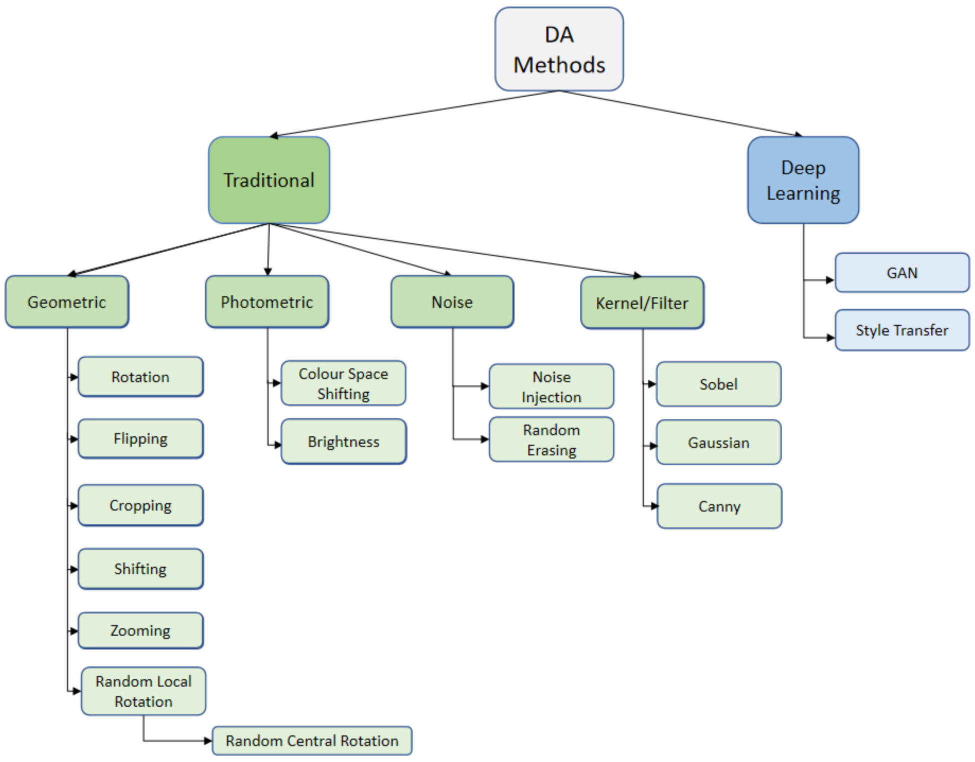
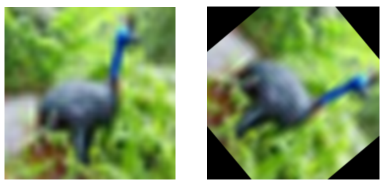


FIGURE 2.3: Data augmentation (DA) taxonomy.

This survey mainly focused on recent articles that used data augmentation techniques in image classification and segmentation, regardless of the data augmentation category, models, or datasets used in the studies. To the best of our knowledge, there are few surveys in the fields of data augmentation in image classification and segmentation. Another main contribution of this article is that we propose a new geometric data augmentation technique, which can complement the current data augmentation strategies. It is well known that traditional rotation is one of the most commonly used geometric data augmentation techniques, see Figure 2.4. It, however, has drawbacks; for example, the loss of a significant amount of pixel information when rotating. It is noticeable that rotating a square-shaped image in a circular trajectory produces black patches at the boundaries, which do not accurately reflect the original data and may affect the final augmentation performance. Filling these black patches with modified pixel values via the wrap, constant, reflection, and/or nearest rotation techniques was a common solution to this issue (see Figure 2.5). In this study, we suggest exploiting local information in images and propose conducting rotation randomly and locally to address the limitations of the traditional rotation. We named our method *"random local rotation "* (RLR). RLR rotates an image's internal circular region by selecting random location, area and angle, which is easy to implement. Rotation performed in a local manner avoids forming black regions near image boundaries. Moreover, this method could also improve the data diversity.

Extensive experiments demonstrated its superior performance compared to its counterpart, i.e., the traditional rotation technique.

The remainder of this article is organised as follows. Sections 2.2 and 2.3 recall the most common traditional data augmentation methods and the most common deep learning-based data augmentation methods, respectively. Section 2.4 reviews some recent research in image classification and segmentation utilising data augmentation for performance enhancement. Sections 2.5 and 2.6 present our proposed data augmentation method and the experimental results validating its promising performance. We conclude our study in Section 2.7.



Given image Rotated image

FIGURE 2.4: Traditional rotation. Left and right: the given image (from the CIFAR-10 dataset) and the rotated image with a randomly rotated angle. Black areas appear in the corners of the rotated image and the corners in the given image are cut off in the rotated image.

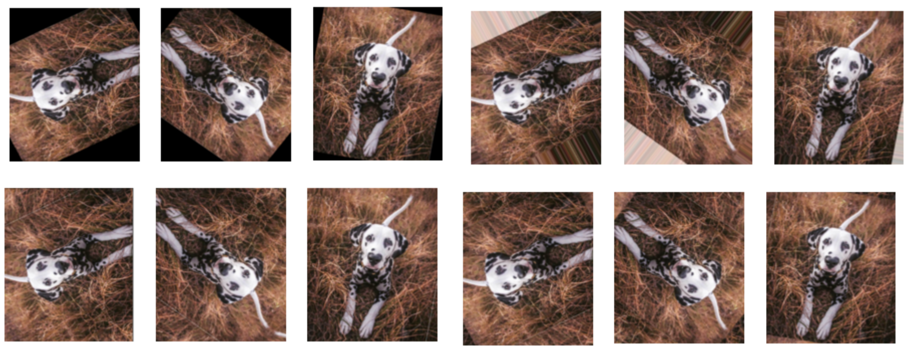


FIGURE 2.5: Data augmentation by different types of rotation techniques. The left three figures in the first row show the ”constant” technique, i.e., the traditional rotation (TR), resulting in black areas around the boundary. The RNR technique is used in the right three figures of the first row. The first three figures of the second row give the results of filling up the black areas by using RRR, and the right three figures use RWR. For each technique, three random angles were selected for rotation.

2.2 Traditional Data Augmentation Techniques

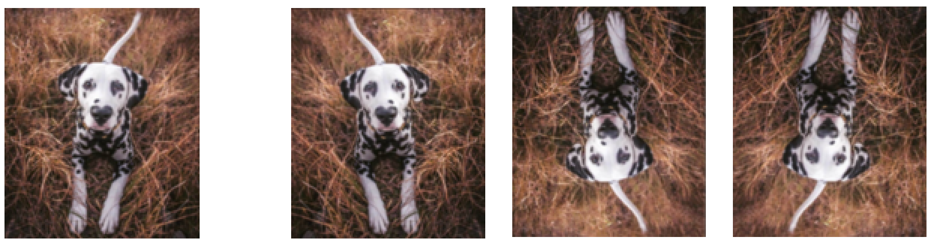
This section briefly recalls the most commonly used traditional data augmentation approaches.

2.2.1 Geometric Transformations

Basic geometric operations. like flipping, cropping and random rotation, are still sought-after techniques to augment data. They generally increase the data size to improve data diversity, and are fairly easy to apply, see below for more detailed description.

Flipping . The term flipping refers to the process of flipping images horizontally or vertically or both, see Figure 2.6. The most commonly used flipping is horizonal flipping, since it is more realistic. For example, a cat versus dog dataset may include all the dog images heading to the left from the spectator view. Not surprisingly, the trained model may suffer from misclassifying dogs heading to the right. The best way to alleviate this problem is to collect more training images that include as many different views as possible. When collecting more images is difficult, flipping may directly solve this type of problem.

Flipping is one of the most intuitive strategies to increase data size or diversity. However, it may be inappropriate when the data has unique properties. For example, considering the concept of label safety, discussed in Shorten and Khoshgoftaar (2019), asymmetric or direction sensitive data, such as letters or digit numbers, cannot use the flipping strategy since it results in inaccurate labels, or even opposite labels.



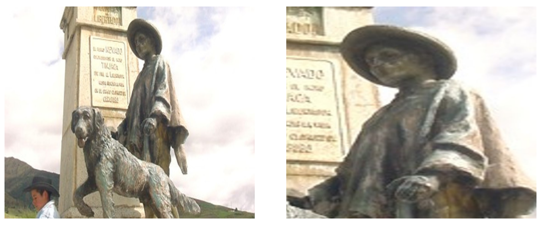
Given image Horizontal Vertical Both

FIGURE 2.6: Data augmentation by flipping. Images from left to right represent the given image, horizontally flipped image, vertically flipped image and the image flipped horizontally and vertically, respectively.

Cropping. Cropping is a basic augmentation technique that randomly crops a part of the given image and then resizes the cropped part back to a certain size. As training data may include samples of different sizes, cropping images to a certain size is a widely used step before training Lu et al. (2019); Shorten and Khoshgoftaar (2019).

It is worth mentioning that cropping may generate samples with incorrect labels. For example, images containing more than one object, which are labelled according to the object with dominant size, may experience a problem when using the cropping technique. In such a case, it is possible to crop an area of the given image that has more details of the accompanying object, rather than the dominant object, see

Figure 2.7. The conventional strategy for training modern state-of-the-art architectures is to crop patches as small as 8% of the given image and label them the same as the given image Bagherinezhad et al. (2018). This frequently results in incorrect labelling in the augmented data, as in the example shown in Figure 2.7.



Given image Cropped patch

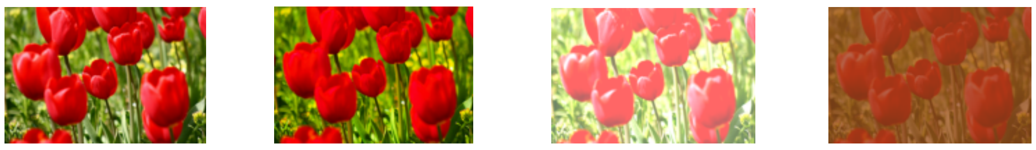
FIGURE 2.7: Data augmentation by cropping. Left and right: the given image (from ImageNet) labelled as "Dog" and the cropped patch. It is clear that the "Dog" is no longer visible in the cropped patch.

Rotation. Rotation is a simple geometric data augmentation technique. The images are rotated by a specified angle, and the newly created images are used alongside the originals as training samples. The disadvantage of rotation is that it may result in information loss at the image boundary, see Figure 2.4 and the first row in Figure 2.5. There are several possible solutions, e.g., random nearest neighbor rotation (RNR), random reflect rotation (RRR) and random wrap rotation (RWR), to fix the boundary problem of the rotated images. In particular, the RNR technique repeats the nearest pixel values to fill in the black areas, while the RRR technique employs a mirror-based approach and the RWR technique uses the periodic boundary strategy to fill in the gaps; see Figure 2.5 for an example.

These geometric data augmentation techniques have been shown to be highly effective in improving diversity and increasing data quantity. For example, Masi et al. (2016) used a fine-grained dataset of ten classes to test a variety of geometric augmentation methods for the task of aircraft classification. Cropping, rotating, rescaling, polygon occlusion, and a combination of these techniques were all tested. The cropping technique combined with occlusion achieved the highest improvement, i.e., increasing the task performance by 9% against the benchmark result. Their study, however, did not examine photometric data augmentation strategies (see below).

2.2.2 Photometric Transformations

A different type of traditional transformation is to change pixels' values rather than their positions. This approach includes different techniques, such as changing brightness, contrast, and/or colours.



(A) Given image

(B) Saturation

(C) Brightness

(D) Contrast

FIGURE 2.8: Data augmentation by colour jittering. (**a–d**) represent the given image, and the augmented images by manipulating the colour saturation, brightness and contrast, respectively.

Typically, a digital image is encoded as a tensor of three dimensions, i.e., height × width × colour channels. The difference between different colour representation schemes lies in the channel part of the tensor. For example, the RGB colour representation scheme uses a combination of three colour channels (i.e., red, green and blue) to represent individual pixels. Manipulating these individual colour channels is a very basic technique in colour augmentation Shorten and Khoshgoftaar (2019). For example, an image can be swiftly transformed into its representation in one colour channel if the others are set to black.

In addition to the RGB colour space, there are many other colour spaces. For example, the HSL colour representation scheme combines hue, saturation and lightness to represent individual pixels Ibraheem et al. (2012). A hue is a single pigment that has no tint or shade. Saturation refers to colour intensity and lightness refers to how light a colour is. HSL is user-friendly since it is convenient to see how a particular colour appears using different values for these three attributes. Please refer to e.g., Ibraheem et al. (2012); Cai et al. (2017) for different colour spaces. Transferring from one colour space to another can be a useful technique for data augmentation.

Colour jittering is a photometric data augmentation technique that employs either random colour manipulation Wu et al. (2015a) or predetermined colour adjustments Sharif Razavian et al. (2014), such as randomly changing the brightness, contrast or colour properties of an image, see Figure 2.8.

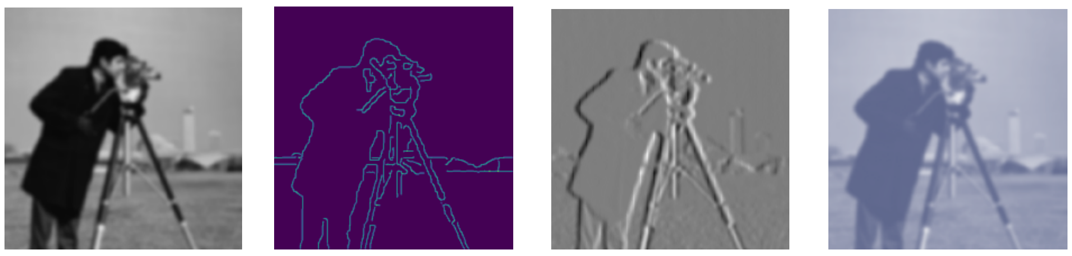
Traditional photometric techniques in augmenting data may have limitations, e.g., high memory and computation requirements. In addition, they may result in crucial image information loss, particularly when the feature is a colorimetric feature capable of differentiating different dataset categories Khalifa et al. (2021).

2.2.3 Kernel / Filter

Kernel plays an important role in deep learning. It can extract certain features from given images as a filter by sliding a window across the images. CNN models can learn features from images by automatically updating their kernel values according to the back-propagation process. Similarly, kernels with distinct values can also be used to conduct data augmentation and generate specific images containing specific features Shorten and Khoshgoftaar (2019).

In computer vision, filters can be used for edge detection (e.g., using the Sobel Kanopoulos et al. (1988) or Canny (1986) filters), sharpening (e.g., using high-contrast vertical or horizontal edge filters), and blurring (e.g., using the Gaussian filter). In particular, edge enhancement that improves object edges within images can be used for data augmentation. It is hypothesised that using training images with augmented edges could improve CNN performance, since the learned kernels in CNN could detect objects' shapes more easily Taylor and Nitschke (2018). Analogously, blurring images can also be utilised for data augmentation and could make models more resistant to blur or noise. Figure 2.9 shows an example of using different kernels/filers to augment images.

Using filters for data augmentation is a relatively unexplored field, even though the idea is straightforward. Its application in areas, such as action recognition, could be advantageous. For instance, edge detection filters may aid in recognising the human shape, thereby enabling the inference of its action. Motion blur may be used to augment data so as to improve models' resistance to blurring in action recognition Guo and Lai (2014); Wu et al. (2014).



(A) Given image

(B) Canny filter

(C) Sobel filter

(D) Gaussian filter

FIGURE 2.9: Data augmentation by using kernels/filters.

2.2.4 Noise Transformations

Noise is commonly defined as a random variation in brightness or colour information Ravishankar et al. (2017). It is frequently caused by technical limitations of the image capture sensor or poor environmental conditions. Unfortunately, these issues are often unavoidable in actual situations, making image noise a prevalent problem to address.

Noise in data may appear to be a problem for neural networks in particular. Real-world data is rarely perfect Nazaré et al. (2017). When neural networks are evaluated on real-world data, noise can impair their accuracy and cause them to perform poorly in generalisation. At the very least, the data used to test deep learning models may not be as clean as the data used to train them. This may account for why deep neural network models frequently perform poorly in tests. Their robustness could be improved by augmenting data with different types of noise. Gaussian, salt and pepper, and speckle noise are three well-known forms of noise that can be used to augment image data Boonprong et al. (2018), e.g., see Figure 2.10.

Gaussian noise is statistical noise with a probability density function equal to the normal distribution. The distribution of Gaussian noise is uniform throughout the signal Boyat and Joshi (2015). Since it is additive noise, the pixels in a noisy image are made up of the sum of their original pixel values plus random Gaussian noise values. It is also independent at each pixel, and independent of the signal magnitude. Salt-and-pepper noise is also known as “spike noise” or “impulsive noise”. It causes white and black pixels to appear at random points in the image. This type of noise is mainly created by data transfer errors Chen et al. (2009). Speckle noise is multiplicative. It is generated by multiplying random values with different image pixels Boyat and Joshi (2015). These different types of noise described above are generally dispersed over the image level. When they are used to augment data, deep learning models could be resistant to data that contains certain types of noise.

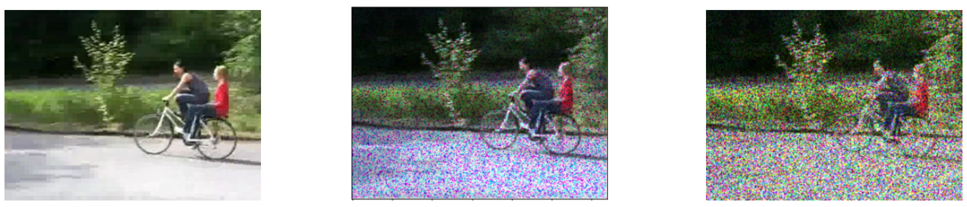

(A) Given image (B) Salt and Pepper (C) Speckle

FIGURE 2.10: Data augmentation by using noise transformation.

2.2.5 Random Erasing

Random erasing Zhong et al. (2020) is a data augmentation technique which does not attempt to change individual image pixel values in general. Instead, it replaces the values of the pixels within a random size rectangle in an image by a random value, see Figure 2.11 for example. We could regard random erasing as a kind of noise technique focusing on local areas rather than individual pixels. It intends to make the model resistant to occlusion of objects in images (e.g., the datasets CIFAR-10, CIFAR-100, and ImageNet) and, thus, to reduce the possibility of overfitting. It enhances the data diversity holistically without increasing the data size, which is different from the other aforementioned data augmentation methods.

Since the random erasing technique selects a rectangular area (i.e., occlusion region) randomly, it may entirely erase the object information to be classified in the image.

Therefore, it may not be recommended in categorising sensitive data which cannot withstand the deletion of a randomly generated local area in images, such as the cases of categorising licence plate numbers and letters.

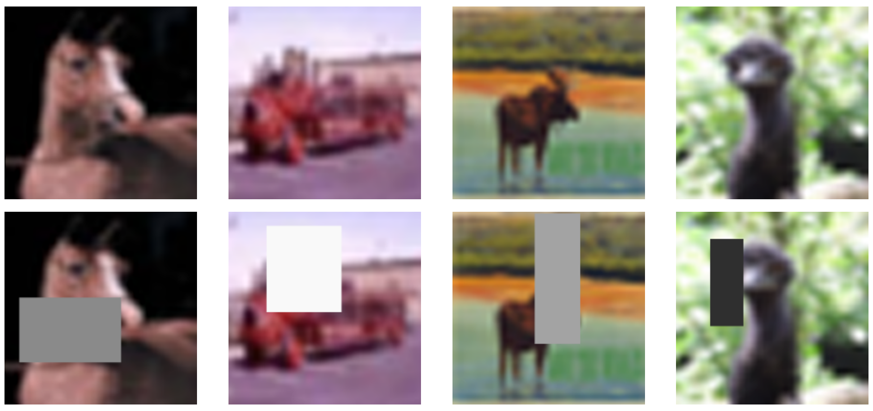


FIGURE 2.11: Data augmentation by the random erasing technique. The first and second rows represent the given images (from CIFAR-10) and the images after random erasing, respectively.

2.3 Deep Learning-Based Data Augmentation Techniques

This section briefly recalls the most commonly used deep learning-based data augmentation approaches.

2.3.1 Texture Transfer

Texture transfer Efros and Freeman (2001) aims to generate textures from source images while maintaining control over the semantic content of the source images. It allows the generation of new images with given textures, while preserving the original images' visual characteristics, such as contours, shading, lines, strokes and areas. The study in Geirhos et al. (2018) demonstrated that CNNs are biased towards objects' texture rather than shape, indicating that employing texture transfer may make a model more texture resistant.

The majority of traditional texture transfer methods resample textures into each particular content image Gatys et al. (2016). For example, image quilting Efros and Freeman (2001) creates a new image by stitching together small patches of other images. The work in Hertzmann et al. (2001) developed an image analogue technique, using pixel resampling to transfer textures from one image to another Mikołajczyk and Grochowski (2018). The newly generated images could be added into the training dataset to enlarge the data size and enhance its diversity.

2.3.2 Adversarial Training

Adversarial examples, also known as machine illusion, have attracted considerable attention in the deep learning community. Adversarial examples can also be seen as members of the noise injection data augmentation family. By injecting a systematic noise into a given image, the CNN model outputs a completely different prediction, even though the human eye cannot detect the difference, see Figure 2.12. For example, the work in Su et al. (2019) created adversarial examples by changing a single pixel per image. Adversarial training is where these examples are added to the training set to make the model robust against attacks. As adversarial examples can detect weak points in a trained model, this way of augmenting data can be seen as an effective data augmentation approach.

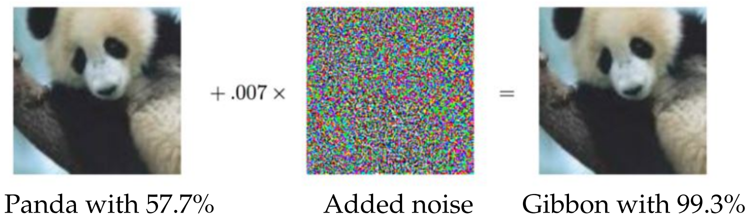


FIGURE 2.12: An adversarial example taken from Goodfellow et al. (2014b). Even though the given image and the image after adversarial noise added look exactly the same to the human eye, the noise fools the model successfully, i.e., the model labels the two images as different classes.

2.3.3 Generative Adversarial Networks for Data Augmentation

Inspired by adversarial examples, the generative adversarial network (GAN), proposed in Goodfellow et al. (2014a), has been widely used for data augmentation. Synthetic images created by GANs, which even humans find difficult to distinguish from the real images, help models significantly increase their robustness. GAN consists of two networks, i.e., a generator, which creates new images, and a discriminator, which tries to detect if the generated images are real or fake. For the variants of GANs, please refer to e.g., DCGAN Radford et al. (2015), progressively growing GANs Karras et al. (2017) and CycleGANs Zhu et al. (2018).

2.4 Data Augmentation in Image Classification and Segmentation

Data augmentations performed using traditional transformation techniques is still the most popular among academics, due to their simplicity Antoniou et al. (2017); Shorten and Khoshgoftaar (2019). Often, traditional and deep learning-based augmentation approaches are used either separately or in tandem. Image classification and image segmentation are two common, yet important, research areas in computer vision, which typically use data augmentation approaches. In this section, we discuss recent research, mostly within the past five years, in these two areas that leverage data augmentation for performance enhancement.

2.4.1 Data Augmentation on Image Classification

Lots of works have used data augmentation in image classification tasks, and their results vary, depending on aspects such as, models, data and applications. See Table 2.1 for a brief survey in this respect.

TABLE 2.1: Survey of data augmentation techniques in recent image classification works.

Papers	Dataset	Aug. Techniques	Model	Task	Findings
Shijie et al. (2017)	CIFAR10; ImageNet (10 categories)	GAN/WGAN, flipping, cropping, shifting, PCA jittering, colour jittering, noise, rotation.	AlexNet	Image classification	Four methods (i.e., cropping, flipping, WGAN, and rotation) perform generally better than other augmentation methods, and some appropriate combination methods are slightly more effective than the individuals.
Perez and Wang (2017)	A small subset of ImageNet; MNIST	Neural augmentation, CycleGAN, GANs, cropping, rotating, and flipping	SmallNet	Image classification	GANs do not perform better than traditional techniques.
Hussain et al. (2017)	Digital Database for Screening Mammography (DDSM)	Flipping, cropping, noise, Gaussian filters, principal component analysis (PCA)	VGG-16	Medical images classification	The flipping and Gaussian filter techniques are better than noise transformation.
Pawara et al. (2017)	Folio, AgrilPlant, and the Swedish Leaf datasets	Rotation, blur, contrast, scaling, illumination, and projective transformation	AlexNet; GoogleNet	Plant image classification	CNN models trained from scratch benefit significantly from data augmentation.
Inoue (2018)	ILSVRC2012; CIFAR-10	SamplePairing, flipping, distorting, noise, and cropping	GoogLeNet	Image classification	Developed a new technique known as SamplePairing.

Continued on next page...

TABLE 2.1: (continued)

Papers	Dataset	Aug. Techniques	Model	Task	Findings
Li et al. (2018a)	Indian Pines and Salinas datasets	Pixel-block pair, flipping, rotation, and noise	PBP-CNN	Hyperspectral imagery classification	A threefold increase in sample size is often sufficient to reach the upper bound.
Frid-Adar et al. (2018)	Liver lesions dataset	Translation, rotation, scaling, flipping and shearing, and GAN-based synthetic images	Customised small CNN architecture	Medical image classification	Combining traditional data augmentation with GAN-based synthetic images improves small datasets.
Pham et al. (2018)	Skin lesion dataset (ISBI Challenge)	Geometric augmentation and colour augmentation	InceptionV4	Skin cancer image classification	Skin cancer and medical image classifiers could benefit from data augmentation.
Motlagh et al. (2018)	Tissue Micro Array; Breast Cancer Histopathological Images (BreaKHis)	Random resizing, rotating, cropping, and flipping	ResNet50	Breast cancer image classification	Traditional data augmentation techniques are adequate for obtaining distinct samples of various types of cancer.
Zheng et al. (2019)	Caltech 101; Caltech 256	Neural style transfer, rotation, and flipping	VGG16	Image classification	Neural style transfer can be utilised as a deep-learning data augmentation technique.
Ismael et al. (2020)	Brain tumor dataset	Horizontal and vertical flips, rotating, shifting, zooming, shearing, and brightness alteration	ResNet	MRI image classification (Brain Cancer)	The effectiveness of traditional augmentation methods varied among classes.
Gour et al. (2020)	BreaKHis dataset	Stain normalisation, image patch generation, and affine transformation	ResHist model	Breast cancer histopathological image classification	The model performance for classifying histopathology images is better with data augmentation than with pre-trained networks.
Nanni et al. (2021)	Virus, a bark, a portrait, and a LIGO glitches datasets	Kernel filters, colour space transforms, geometric transformations, and random erasing	ResNet50	Image classification	Introduced the discrete wavelet transform and the constant-Q Gabor transform as two new methods for data augmentation.
Anwar and Zakir (2021)	Customised image based ECG signals	Flipping, cropping, contrast and Gamma distortion	EfficientNet B3	ECG images classification	In the experiment with images of ECG signal, traditional data augmentation did not improve the performance of neural networks.

Continued on next page...

TABLE 2.1: (continued)

Papers	Dataset	Aug. Techniques	Model	Task	Findings
Kandel and Castelli (2021)	MURA dataset	Horizontal flip, vertical flip, rotation, and zooming	VGG19; ResNet50; InceptionV3; Xception; DenseNet121	X-ray images classification	Augmentation was found to significantly enhance classification performance.
Bird et al. (2022)	Public lemon image dataset (2690 images)	Conditional Generative Adversarial Networks (CGANs) for synthetic image generation	VGG16	Fruit quality classification	CGANs improved classification accuracy from 83.77% to 88.75%. Model pruning reduced model size by 50% while maintaining 81.16% accuracy.
Goceri (2023)	Multiple datasets across imaging modalities (MRI, CT, mammography, fundoscopy)	Traditional augmentations (rotation, flipping, scaling) and GAN-based augmentations	CNN architectures (varied)	Classification and segmentation of diseases (brain tumors, lung nodules, breast lesions, eye conditions)	GAN-based augmentations yielded improved performance for rare image datasets, with effectiveness varying based on modality and task.
Naveed et al. (2024)	Various datasets (CIFAR-10, CIFAR-100, ImageNet)	Image mixing and deleting methods: Cut and Delete, Cut and Mix, Mix and Up	WideResNet-28-10, ResNet-50, PyramidNet	Image classification, object detection, fine-grained recognition	Image mixing and deleting improve generalization, robustness, and calibration while addressing overfitting and data scarcity challenges.
Farhan et al. (2025)	Four brain MRI datasets: Dataset A (3264 images), Dataset B (4292 slices), Dataset C (3064 images), Dataset D (253 images)	Oriented Combination MRI (OCMRI): combining images with MSE-guided thresholds	PRCnet	Brain tumor classification	OCMRI improved classification accuracy: Dataset A (85.19% → 92.7%), Dataset B (90.12% → 95.37%), Dataset C (94.77% → 96.51%), Dataset D (90% → 98%).

In 2017, the work in Perez and Wang (2017) suggested that deep learning-based augmentation methods, like GANs, do not perform significantly better than traditional techniques, but consume nearly three times more computational cost. Moreover, in Perez and Wang (2017), a model called "SmallNet" was trained using traditional augmentation techniques and style transfer with CycleGAN Zhu et al. (2017). It was observed that combining deep learning-based methods with traditional techniques could achieve better results. Hussain et al. (2017) evaluated various augmentation strategies on a medical image dataset using VGG-16. They demonstrated that the flipping and Gaussian filter augmentation techniques yielded superior outcomes compared to the other ones, particularly when adding noise, which gave the lowest accuracy. Pawara et al. (2017) applied data augmentation techniques, such as rotation, blur, contrast, scaling, illumination, projective transformation, and multiple combinations of these techniques to enhance plant image classification performance. In this challenge, pre-trained and untrained AlexNet and GoogleNet models were used. It was observed that CNN models trained from scratch benefited significantly from data augmentation, whereas pre-trained CNN models did not. In addition, it was discovered that combinations of data augmentation techniques like rotation and varied illuminations could contribute most for CNN models trained from scratch in achieving excellent performance.

In 2018, Inoue (2018) developed a new technique, known as SamplePairing, in which a new sample was synthesised from one image by overlaying another image randomly selected from the training data, i.e., taking an average of two images. Li et al. (2018a) found that traditional data augmentation techniques were not cumulative, and that a threefold increase in sample size was often sufficient to reach the upper bound. In addition, the PBP technique proposed by the authors significantly increased the number of samples, and was proved to be effective for hyperspectral imagery classification. Frid-Adar et al. (2018) classified liver lesions using a small customised CNN architecture. In order to accommodate small datasets and input sizes, they suggested that CNN designs should often contain fewer convolutional layers. By combining traditional data augmentation techniques with GAN-based synthetic images, more accurate results from a small dataset were obtained. Pham et al. (2018) discussed how to solve the challenges of skin lesion classification and limited data in medical images by applying image data augmentation techniques, such as geometric augmentation and colour augmentation. The effects of a different number of augmented samples were evaluated on the performance of different classifiers, and it was concluded that the performance of skin cancer classifiers and medical image classifiers could be improved by utilising data augmentation. Motlagh et al. (2018) classified several forms of cancer using 6402 tissue microarrays (TMAs) as training samples and utilising transfer learning and deep neural networks. Data augmentation techniques, such as random scaling, rotation, cropping, and flipping, were used to obtain sufficiently different samples, and the results showed that 99.8 percent of the four cancer types, including breast, bladder, lung and lymphoma, were correctly classified using the ResNet50 pre-trained model.

In 2019, Zheng et al. (2019) assessed the efficacy of neural style transfer using VGG16 on the Caltech 101 and Caltech 256 datasets, and the results demonstrated a two-percent gain in accuracy. Recent research has demonstrated that neural style transfer algorithms can apply the artistic style of one image to another image without altering the latter's high-level semantic content, showing that neural style transfer can be used for data augmentation to add more variation to the training dataset.

In 2020, Ismael et al. (2020) employed data augmentation to solve the problem of insufficient training data and imbalanced classes in the MRI image classification task for brain cancer. Various augmentation techniques, including horizontal and vertical flipping, rotating, shifting, zooming, shearing and brightness alteration, were utilised. They observed that each augmentation technique had different effects on the performance of distinct classes. For instance, manipulation of brightness yielded 96 percent accuracy for class one, whereas the rotation technique yielded 98 percent accuracy for the same class. For class two, these two techniques achieved a score of 99 percent with brightness and 98 percent with rotation. By combining all of the previously mentioned augmentation techniques, they were able to attain 99 percent overall accuracy, i.e., 4 percent improvement against the results obtained without data augmentation. Additionally, Gour et al. (2020) developed ResHist, a 152-layer CNN based on residual learning, for breast cancer histopathological image classification. A data augmentation strategy was devised, based on stain normalisation, image patch generation and affine transformation, to improve the model performance. Experimental results demonstrated that with the help of data augmentation the model performance for classifying histopathology images was better than the pre-trained networks, including AlexNet, VGG16, VGG19, GoogleNet, Inception-v3, ResNet50 and ResNet152.

In 2021, Kandel and Castelli (2021) examined the impact of test time augmentation (TTA) on X-ray images for bone fracture detection using the MURA dataset. It was observed that TTA could dramatically improve classification performance, especially for models with a low score, by comparing the performance of nine different augmentation techniques with five state-of-the-art CNN models. Nanni et al. (2021) investigated the performance of over ten different kinds of data augmentation techniques, including kernel filters, colour space transforms, geometric transformations, random erasing/cutting and image mixing, and proposed two approaches: the discrete wavelet transform and the constant-Q Gbor transform. Using the aforementioned data augmentation techniques, the performance of several ResNet50 networks was evaluated on four benchmark image datasets (i.e., a virus dataset, a bark dataset, a portrait dataset, and a LIGO glitches dataset), representing diverse problems and different scales, indicating the efficacy of data augmentation techniques in enhancing model performance. In addition, the work in Anwar and Zakir (2021) investigated the impact of augmenting ECG images for COVID-19 and cardiac disease classification using deep learning. They argued that traditional data augmentation did not improve the performance of neural networks in their experiments with ECG signal images.

2.4.2 Data Augmentation on Image Segmentation

Image segmentation is also an important field in computer vision. It involves grouping an image into different parts where each part may share certain features and characteristics. It has a close relationship to image classification. For example, image segmentation, in some sense, could be achieved by classifying individual pixels in an image into different groups. A great deal of emphasis has been placed on data augmentation in order to achieve better segmentation results, particularly when working with small training datasets. For practical semantic segmentation applications, collecting and annotating sufficient training data for deep neural networks is notoriously difficult. Therefore, data augmentation techniques are of great importance. We, below, survey a number of studies that have involved data augmentation in image segmentation tasks. See Table 2.2 for a summary of relevant literature.

In 2018, the work in Benson et al. (2018) used an encoder–decoder structure, adapted from an hourglass network, prevalent in the field of human-pose estimation Newell et al. (2016), in order to classify and segment brain tumours in MRI scans for the BraTS 2018 challenge Menze and Jakab (2015); Bakas et al. (2017b); Simpson et al. (2019); Bakas et al. (2018). Two data augmentation techniques were utilised: vertical flipping, which matches up to the naturally symmetrical shape of the brain, and random intensity variation, used because the intensity between MRI scans varies significantly. The network was trained with, and without, data augmentation. It was discovered that data augmentation appeared to provide a small increase in accuracy for the Dice coefficient and a significant improvement in Hausdorff accuracy.

TABLE 2.2: Survey of data augmentation techniques in recent image segmentation works.

Papers	Dataset	Aug. Techniques	Model	Task	Findings
Benson et al. (2018)	BraTS 2018	Vertical flipping and random intensity variation.	Hourglass Network Newell et al. (2016)	Brain tumor segmentation	Data augmentation methods appear to have a different impact on the Dice coefficient and Hausdorff accuracy.

Continued on next page

TABLE 2.2: (Continued)

Papers	Dataset	Aug. Techniques	Model	Task	Findings
Casado-García et al. (2019)	ISBI challenge Arganda-Carreras et al. (2015)	Automated image augmentation tool.	U-Net architecture with four different models	Semantic segmentation	Introduced an approach that enables researchers to use image augmentation techniques automatically to the challenges of object classification, localisation, detection, semantic segmentation, and instance segmentation.
Ma et al. (2019)	Sheep segmentation dataset (SSG)	Colour transformation, flipping, cropping, projection transformation, local copy, and JPEG compression.	DeepLabv3+ Chen et al. (2018)	Semantic segmentation	A combination of augmentation methods could achieve the best performance, while excessive augmentation could degrade the performance.
Qiao et al. (2020)	Cattle segmentation dataset	Random image cropping and patching.	Bonnet Milioto and Stachniss (2019)	Semantic segmentation	The proposed method of randomly cropping and patching images to increase the number of training images improves segmentation performance.
Khryashchev and Larionov (2020)	Planet, and the Resurs datasets	Random chromatic distortion, rotation, and shifting.	U-Net with the ResNet34 as encoder	Semantic segmentation	The application of random chromatic distortion in HSV colour format improves the robustness of deep learning algorithms for images with noise, such as small clouds and glare from reflective surfaces.
Chen and Jung (2020)	Tongue image dataset	Cropping, rotation, flipping, and colour transformations.	U-Net with 15 different CNN models as encoders	Semantic segmentation	Geometric transformations can achieve higher performance than colour transformations, and segmentation accuracy can be increased by 5 to 20% compared to no augmentation.
Qin et al. (2020)	Kidney Tumour dataset	An automatic deep reinforcement learning based augmentation method	An end-to-end augmentation segmentation architecture	Medical image segmentation	Conventional augmentation techniques (e.g., rotation, cropping, etc.) are random and sometimes damaging to the image segmentation task.

Continued on next page

TABLE 2.2: (Continued)

Papers	Dataset	Aug. Techniques	Model	Task	Findings
Cirillo et al. (2021)	BraTS2020 Dataset	Flipping, rotation, scaling, brightness adjustment, and elastic deformation.	3D U-Net Ronneberger et al. (2015)	3D brain tumor segmentation	Conventional data augmentation significantly improves the validation performance of brain tumour segmentation.
Su et al. (2021)	Narrabri and Bonn datasets	Random image cropping and patching (RICAP) method.	Bonnet	Semantic segmentation	The RICAP technique increases the mean accuracy and mean intersection over union (IOU) of the CNNs with the traditional data augmentation.
Zhang et al. (2021a)	PASCAL VOC 2012; Cityscapes; CRAG dataset	Object-level augmentation method.	MobileNet based DeepLab V3+	Semantic segmentation	ObjectAug can easily be integrated with existing image-level augmentation techniques to further improve the segmentation performance. ObjectAug supports category-aware augmentation that gives objects in each category a variety of options.
Mallios and Cai (2021)	VoxTox Burnet et al. (2017)	GAN-based synthetic images.	RS-FCN	Rectum segmentation	Demonstrated the viability of producing synthetic data and subsequently incorporating it into the training samples in order to get satisfactory outcomes.
Chen et al. (2022)	Two public MRI datasets: cardiac and prostate images	AdvChain: adversarial photometric and geometric chained transformations	U-Net-based segmentation networks	Cardiac and prostate MRI segmentation	AdvChain significantly improved segmentation accuracy in low-data and semi-supervised settings, outperforming existing methods like RandAugment.
Zhang et al. (2023)	Four datasets: ATLAS (chronic stroke), in-house acute ischemic stroke, in-house whole brain tumor, MSSEG (multiple sclerosis lesions)	CarveMix: lesion-aware ROI carving and mixing, harmonization, and mass effect modeling	nnU-Net	Brain lesion segmentation (chronic stroke, ischemic stroke, tumors, multiple sclerosis)	CarveMix improves segmentation performance with Dice coefficients outperforming other augmentation methods, especially under limited training data conditions.

Continued on next page

TABLE 2.2: (Continued)

Papers	Dataset	Aug. Techniques	Model	Task	Findings
He et al. (2024)	Three datasets: Head and Neck (HNC, HNPETCT, 140 CT scans), SegTHOR (60 CT scans), Pancreas-CT (BTCV, 50 CT scans)	Statistical deformation-based augmentation with inter- and intra-patient deformations	3DAttU-Net with coarse-to-fine segmentation framework	Segmentation of organs-at-risk (OARs) in CT scans	Achieved superior segmentation performance, improving DSC by 1.56% and reducing 95% HD on average compared to state-of-the-art methods.
Sun et al. (2025)	Five datasets: ISIC 2017 (skin lesions), GlaS (glands), MoNuSeg (nuclei), Synapse (organs in CT), BraTS2018 (brain tumors in MRI)	HSMix: Superpixel-based CutMix (hard) and Mixup (soft)	UNet, TransUnet, DeepLabv3+, HiFormer, UNeXt	Medical image segmentation	Improved Dice similarity coefficients (e.g., ISIC: 80.92% → 83.50%, BraTS2018: 77.33% → 78.97%) and reduced Hausdorff distances with minimal computational overhead.

In 2019, Casado-García et al. (2019) presented a versatile method which was implemented in the open-source package CLoDSA, dedicated to classification, semantic segmentation, instance segmentation, localisation and detection. Three different datasets were used to demonstrate the benefits of applying data augmentation. Ma et al. (2019) created the SSG dataset, i.e., a small-scale and open-source sheep segmentation dataset containing hundreds of images. To find the best technique for this small semantic segmentation dataset, they evaluated seven data augmentation methods, including colour transformation, flipping, cropping, projection transformation, local copy, a proposed technique named ”JPEG compression” and their combinations. Experimental results showed that the combination of compression, cropping and local shift could achieve the best augmentation performance for their AI-Ranch application. However, they also found that excessive augmentation could degrade performance.

In 2020, Qiao et al. (2020) introduced a data augmentation technique where images were randomly cropped into distinct regions and then patched together to form a new one. Experimental results on their acquired cattle dataset showed that this data augmentation technique, together with an open-source semantic segmentation CNN architecture, “Bonnet” Milioto and Stachniss (2019), achieved 99.5 percent mean accuracy and 97.3 percent mean intersection of unions. In Khryashchev and Larionov

(2020), an U-Net neural network with the ResNet34 encoder was used for automated wildfire detection on high-resolution aerial photos using two small satellite RGB image datasets. To overcome the small data size challenge, data augmentation techniques, such as rotation, shifting and random chromatic distortion in HSV colour format, were used to increase the robustness of the deep learning algorithm for noisy images, such as small clouds and glare from reflective surfaces. The experimental results showed that data augmentation methods led to better results on test datasets for all metrics used in the experiments. Qin et al. (2020) argued that the data generated by conventional augmentation techniques (e.g., rotation, cropping, etc.) was random and sometimes detrimental to the image segmentation process. In light of this, an automatic learning-based data augmentation technique was developed for CT kidney tumor segmentation.

The work in Chen and Jung (2020) focused on automatic tongue segmentation using 15 different pre-trained network models (such as VGG, ResNet, ResNext, DenseNet, EfficientNet, inceptionV3, SE-ResNet, inception, ResNetV2, etc.). They utilised multiple label-preserving transformations to increase the size and diversity of the training dataset. Their findings indicated that geometric transformations could achieve greater performance than colour transformations, and that the segmentation accuracy could be improved by 5 to 20 percent compared to no augmentation.

In 2021, the work in Zhang et al. (2021a) proposed a data augmentation technique, named ObjectAug, for image segmentation. The ObjectAug technique operates at the object level by first decoupling the image into individual objects and the background using semantic labelling and, then, each object is individually augmented using conventional augmentation techniques (e.g., scaling, shifting and rotation), followed by image inpainting, which is utilised to further restore the pixel artefacts introduced by object augmentation. The final step is integrating the augmented objects and background into an augmented image. Extensive experiments on both normal and medical image datasets demonstrated that the ObjectAug technique outperformed conventional augmentation techniques and improved segmentation performance. Cirillo et al. (2021) examined how augmentation techniques, such as flipping, rotation, scaling, brightness adjustment and elastic deformation, affected the learning process when training a standard 3D U-Net Ronneberger et al. (2015) on the BraTS dataset Menze et al. (2014); Bakas et al. (2017b,a). In multiple cases, their findings indicated that data augmentation significantly improved validation performance. They presumed that the reason why data augmentation had not been thoroughly investigated for brain tumour segmentation was because the BraTS training set was quite large and several works Lyksborg et al. (2015); Havaei et al. (2017) suggested that data augmentation would not be of much assistance.

Mallios and Cai (2021) investigated image-guided radiation therapy Delaney et al. (2005); Burnet et al. (2017), which is one of the most prevalent methods for treating numerous types of cancer. Their study included the development of deep learning approaches for segmenting the organs-at-risk in CT images during radiation therapy. It was observed that the scarcity of annotated data, stemming from the difficulty and time consuming nature of manual annotation in this area, hindered research development for medical applications. In order to compensate for the shortage of labelled real-world data required to train very deep models, like FCN architecture Long et al. (2015), cGAN Mirza and Osindero (2014) was used to generate synthetic images. The experimental results illustrated the superior performance of the proposed segmentation methods for the rectum under the help of deep learning-based data augmentation. In Su et al. (2021), a framework for augmenting data for semantic segmentation, based on the random image cropping and patching (RICAP) method, was presented. Experiments on two datasets using Bonnet architecture Milioto and Stachniss (2019) showed that the developed framework improved segmentation performance in terms of accuracy and mean intersection over union.

2.5 Proposed Strategy for Data Augmentation

In this section, we propose a new data augmentation technique, belonging to the traditional data augmentation category. It was inspired by techniques focusing on local areas in images, e.g., the random erasing technique.

Let $\mathcal{D}$ be the training dataset. Let $C_{x,y,r}$ be a circular area in an image $\boldsymbol{I} \in \mathcal{D}$, with centre location (x, y) and radius r. Let $\theta \in [0, 2\pi]$ be an angle for rotation.

The main procedure of the proposed augmentation technique is given below. Firstly, $\forall \boldsymbol{I} \in \mathcal{D}$, we select a circular area $C_{x,y,r}$ within image $\boldsymbol{I}$, with a randomly generated centre (x, y) and radius r. Then, the image content within the circular area $C_{x,y,r}$ is rotated with a randomly generated angle $\theta \in [0, 2\pi]$, while the image content outside the circular area $C_{x,y,r}$ is kept, and we call this newly generated image $\tilde{\boldsymbol{I}}$. Finally, image $\tilde{\boldsymbol{I}}$ is used to augment the original training dateset $\mathcal{D}$. Here we suggest two ways. The first one is to use the generated image $\tilde{\boldsymbol{I}}$ to replace the original image $\boldsymbol{I} \in \mathcal{D}$. This way does not change the size of the dataset $\mathcal{D}$, but may change the data diversity. The other way is to add image $\tilde{\boldsymbol{I}}$ into the dataset $\mathcal{D}$, which increases the dataset size and enhances the data diversity. We call the above technique random local rotation (RLR), see Figure 2.13 for the diagram showing the conducting of the RLR data augmentation strategy and Algorithm 1 for the summary of RLR.

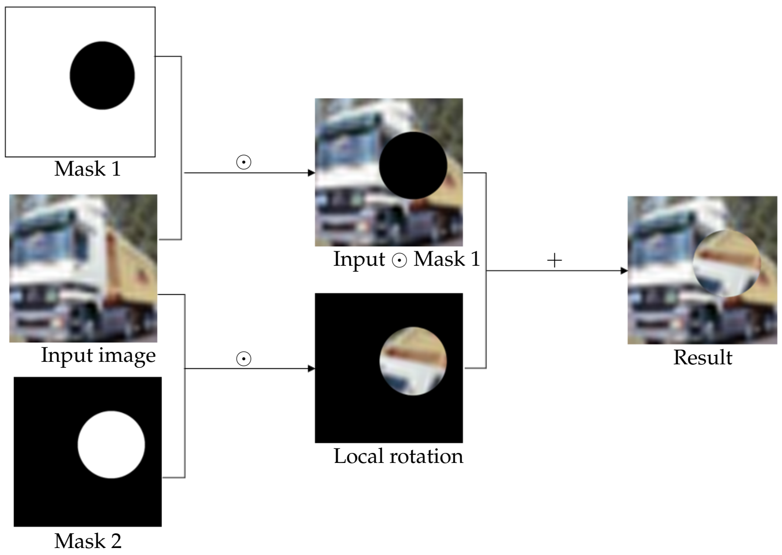


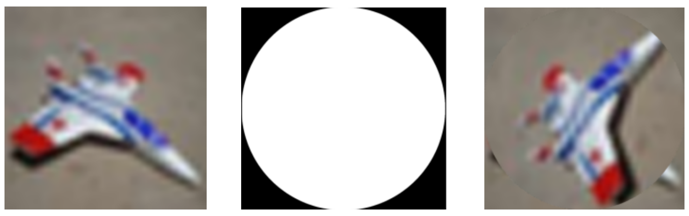
FIGURE 2.13: The proposed random local rotation data augmentation strategy. Symbol ⊙ represents pointwise multiplication.

Algorithm 1 Random Local Rotation

```
Input: The training dataset $\mathcal{D}$
Output: The augmented training dataset $\tilde{\mathcal{D}}$
Create a subset say $\mathcal{D}' \subseteq \mathcal{D}$ by randomly selecting $N$ images from $\mathcal{D}$
Create an empty set say $\mathcal{D}^*$
for $\forall \boldsymbol{I} \in \mathcal{D}'$ do
    Randomly generate centre $(x, y)$ and radius $r$ within image $\boldsymbol{I}$
    Form a circular area $C_{x,y,r}$ within image $\boldsymbol{I}$ with centre $(x, y)$ and radius $r$
    Randomly generate angle $\theta \in [0, 2\pi]$
    Form image $\tilde{\boldsymbol{I}}$ by rotating the area within $C_{x,y,r}$ in image $\boldsymbol{I}$ with angle $\theta$
    Add $\tilde{\boldsymbol{I}}$ into $\mathcal{D}^*$
end for
Way 1: $\tilde{\mathcal{D}} \leftarrow \mathcal{D}^* \cup \mathcal{D} \setminus \mathcal{D}'$
Way 2: $\tilde{\mathcal{D}} \leftarrow \mathcal{D}^* \cup \mathcal{D}$
```

A special case of RLR uses the largest possible circular rotation area in the image centre, see Figure 2.14. In the rest of this article, we call this special case random centre rotation (RCR). RCR could be applied as a direct replacement of the traditional rotation technique for data augmentation.

An obvious advantage of RLR against the traditional rotation is that it avoids the black boundary caused by traditional rotation, as shown in Figure 2.4. Moreover, the local area information distortion brought by RLR could improve the data diversity, without removing much information from the given images, like other augmentation techniques, e.g., image cropping, random erasing, etc. Detailed validation of RLR is presented in the next section.



(**a**) Given image (**b**) Rotation mask (**c**) Rotated image

FIGURE 2.14: Random local rotation data augmentation technique using the largest possible circular rotation area in the image centre.

2.6 Experiments

To validate the proposed RLR agumentation technique, we employed three state-of-the-art CNN models, i.e., ResNet50 He et al. (2016), MobileNet Howard et al. (2017) and InceptionV3 Szegedy et al. (2016), which were all trained from scratch. We conducted experiments in both classification and segmentation tasks and mainly compared with the traditional rotation technique (shortened to TR) with randomly selected rotation angles. The quantitative results reported below with standard deviation were obtained by repeating the experiments five times.

2.6.1 Classification Experiment

The CIFAR-10 dataset was selected for conducting experiments regarding the classification task. It contained 60,000 coloured images, where every image was of size 32 × 32, A total of 50,000 images were for training and 10,000 images were for testing. CIFAR-10 consisted of ten classes, each with 6000 images. To simulate the scenario of data scarcity, we reduced the original training data size to 2%, 4% and 6%, forming three subsets with numbers of samples of 1000, 2000 and 3000, respectively, and used the original test set for testing.

For each subset, three extra copies were created by the TR, RCR and RLR data augmentation techniques. Each augmented copy was twice as large as its corresponding original subset. The data balance between the classes was also taken into consideration when constructing these subsets. Additionally, the image resolution was adjusted to fit the default input shape of each CNN model used in the experiments, i.e., 299 × 299 for InceptionV3, and 244 × 244 for MobileNet and ResNet50. According to the constructed datasets, each model was subjected to a total of 12 tests, (i.e., number of subsets × number of techniques).

For fair comparison, same hyperparameters were kept for each model. Models were trained for 50 epochs with the Adam optimiser and categorical cross entropy loss function. Test accuracy was selected as the monitoring metric. The Spyder platform was utilised to train and evaluate the models.

2.6.1.1 Classification Results

Table 2.3 gives the classification accuracy of the CNN models (i.e., ResNet50, MobileNet and InceptionV3) with the TR, RCR and RLR data augmentation techniques on the three subsets, including the comparison with the baseline results (i.e., the ones obtained on the subsets without using data augmentation).

The results in Table 2.3 show that our proposed data augmentation technique RLR constantly achieved the best classification accuracy with all the three CNN models on all the subsets, indicating its excellent performance. In contrast, the traditional rotation technique did not improve performance, and, in many cases, degraded the results, compared with the baseline results. This might be because of the aforementioned limitations of the traditional rotation technique, i.e., the black and irregular boundary it introduces. As for the performance of the RCR technique, its results were slightly better than the TR results and were comparable to the baseline results. This was what we expected, since RCR is quite similar to TR. Yet, the images generated by RCR did not suffer from the black and irregular boundaries, and, therefore, performed slightly better than TR.

For further performance evaluation, we also reported the comparison of the RLR method with the mostly used traditional data augmentation techniques, see Table 2.4. The smallest subset of CIFAR-10 (i.e., the one with 1000 samples) was employed with data augmentation techniques, including RLR, RNR, RWR, RRR, flipping, shifting, zooming, and brightness. The results in Table 2.4 show that, generally, data augmentation techniques could indeed enhance the performance of different models. It again demonstrated the great performance of the proposed RLR method; for example, RLR achieved the best accuracy when the ResNet model was used. The results in Table 2.4 also show that the performance of the augmentation techniques might differ for different models, which is worth investigating further in the future.

TABLE 2.3: Classification accuracy comparison between the TR, RCR and RLR data augmentation techniques. CNN models, i.e., MobileNet, ResNet and InceptionV3, with the data augmentation techniques, were applied on three different CIFAR-10 subsets, with numbers of samples of 1000, 2000 and 3000, respectively. The results indicated the superior performance of the proposed RLR technique.

Model	Subset	Baseline	RLR	TR	RCR
MobileNet	1000	41.69 ± 0.29	**42.24** ± 0.44	40.57 ± 0.22	39.51 ± 0.52
	2000	50.62 ± 0.43	**51.76** ± 0.56	48.77 ± 0.51	50.6 ± 0.64
	3000	56.95 ± 0.62	**60.96** ± 0.54	55.30 ± 0.82	58.18 ± 0.66
ResNet	1000	39.73 ± 0.64	**41.47** ± 0.39	38.11 ± 0.52	38.28 ± 0.51
	2000	50.16 ± 0.49	**51.06** ± 0.54	47.95 ± 0.84	48.84 ± 0.59
	3000	53.78 ± 0.49	**56.15** ± 0.76	53.38 ± 0.56	53.31 ± 0.56
InceptionV3	1000	42.65 ± 0.63	**45.41** ± 0.55	43.32 ± 0.47	43.15 ± 0.58
	2,000	54.63 ± 0.45	**55.71** ± 0.48	54.85 ± 0.57	53.78 ± 0.28
	3000	61.24 ± 0.37	**62.45** ± 0.86	59.72 ± 0.42	60.18 ± 0.66

TABLE 2.4: Classification accuracy comparison between RLR and other common data augmentation techniques. CNN models, i.e., MobileNet, ResNet and InceptionV3, with the data augmentation techniques, were applied on the smallest subset of CIFAR-10 (i.e., the one with 1000 samples).

Model	Baseline	RLR	RNR	Flip	Shift	Zoom	Bright	RWR	RRR
MobileNet	41.89	**42.28**	41.37	42.52	45.83	45.75	47.84	45.53	45.10
ResNet	39.40	**41.70**	40.07	40.18	41.57	40.19	39.06	41.22	41.18
InceptionV3	42.85	**45.61**	44.84	45.86	43.03	45.81	46.11	46.67	45.41

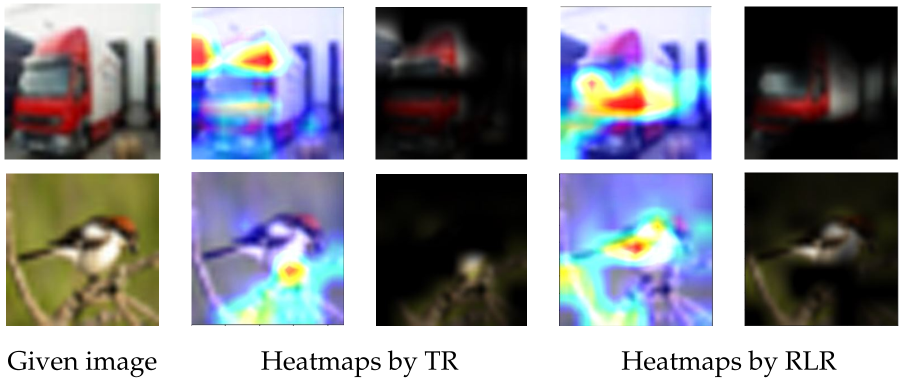


FIGURE 2.15: Data augmentation techniques evaluation by saliency map. Column 1: given images; columns 2 and 3: two types of saliency maps for TR; columns 4 and 5: two types of saliency maps for RLR. In particular, for the two types of saliency maps evaluating each data augmentation technique, the first saliency map highlights the activated area in the given image, and the second highlights the activated area using the content of the given image. The CNNs used for the test images in the first and second rows are MobileNet and ResNet50, respectively. The saliency maps created for the models, which were trained with the dataset augmented with RLR, clearly focus on the wider part of the object while for the other cases where augmentation is achieved with TR, the models focus on a smaller area of the object. The models trained with RLR output more reliable results, together with the wider focus on the target object shown in the above saliency maps, demonstrating the superior performance of RLR compared to TR.

2.6.1.2 Qualitative Comparison via Saliency Maps

To further evaluate the effectiveness of the proposed RLR technique against the traditional rotation, we employed GradCAM Selvaraju et al. (2017), one of the well-known methods illustrating the decision made by CNNs, to show the saliency maps regarding the TR and RLR techniques.

Figure 2.15 shows the saliency maps of the TR and RLR techniques on images randomly selected from the test datasets. The truck image (first row in Figure 2.15) was classified as truck with 94% (here the percentage was the probability produced by the Softmax activation function in the CNN architectures) via MobileNet model trained with the TR augmentation technique, and nearly 100% with the proposed RLR technique. The bird image (second row in Figure 2.15) was classified as bird with 95% via ResNet50 model trained with the TR augmentation technique, and nearly 100% with the proposed RLR technique. The saliency maps shown in Figure 2.15 for the TR and RLR techniques indicated that the proposed RLR technique was, indeed, more effective in terms of assisting the CNN architectures to make decisions based on more reasonable areas within the test images.

2.6.2 Segmentation Experiment

Two publicly available datasets were selected for conducting experiments regarding the segmentation task. The first dataset was the Supervisely Person supervise.ly (2018), which contained 5711 images and 6884 high-quality annotated human instances for human semantic segmentation, see e.g., Figure 2.16. The second dataset was the Nuclei images dataset Hamilton (2018), which contained 670 microscopic images with their corresponding segmentation masks, see e.g., Figure 2.17. Each augmented copy was twice as large as its corresponding original dataset. Then, each dataset copy was divided into training (90% of the data) and validation (10% of the data) subsets. Note that, in this experiment, we also considered the concept of equivariance. Equivariance implies that the output changes in proportion to the input. The concept of equivariance is important in segmentation, where the location of the object and the location of the segmented object shift proportionally, e.g., see Figures 2.16 and 2.17. In contrast, invariance refers to a change in the location of an object while the output remains unchanged, which is considered in Section 2.6.1 for the classification task.

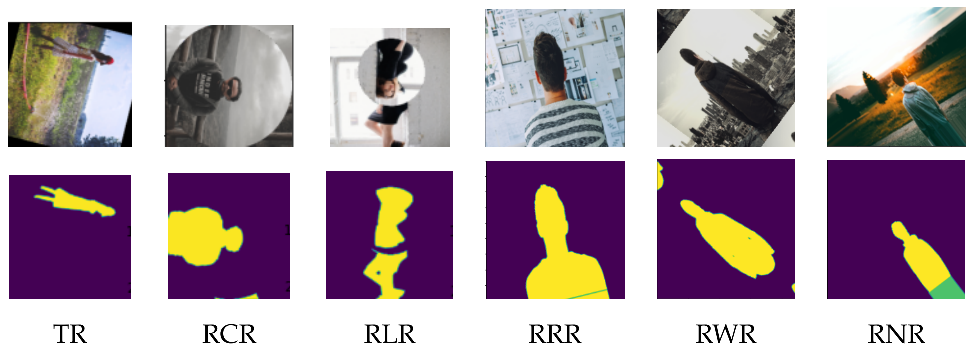


FIGURE 2.16: Samples of the Supervisely Person dataset by applying the RLR, TR, RCR, RRR, RWR and RNR augmentation techniques. Rows one and two are the augmented samples with their corresponding human body segmentation, respectively.

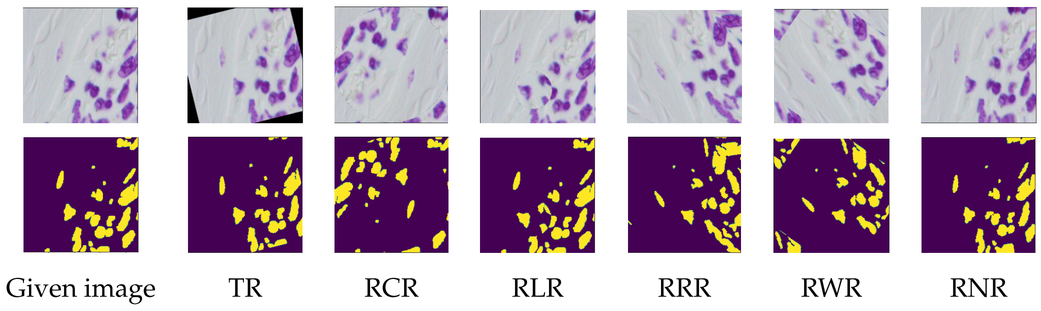


FIGURE 2.17: Samples of the Nuclei images dataset by applying the RLR, TR, RCR, RRR, RWR and RNR augmentation techniques. Rows one and two are the augmented samples with their corresponding segmentation, respectively.

Two autoencoders were used to conduct the semantic segmentation task. These two autoencoders were constructed based on two models (i.e., MobileNet and VGG16), each with a customised decoder, see Table 2.5 for the detailed architectures. Each autoencoder was subjected to a total of seven tests. They were trained for 200 epochs with the Adam optimiser and binary cross entropy loss function.

TABLE 2.5: The architectures of the decoders of the MobileNet-based and VGG16-based auto-encoders. Conv2D and Conv2DT represent the 2D convolutional layer and the transposed 2D convolutional layer, respectively.

MobileNet Decoder				**VGG16 Decoder**			
Layer	**Kernel Size**	**Filters Number**	**Activation Function**	**Layer**	**Kernel Size**	**Filters Number**	**Activation Function**
Conv2DT	(3,3)	1024	Relu	Conv2DT	(3,3)	1024	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	1024	Relu	Conv2D	(3,3)	1024	Relu
Batch Normalisation				Batch Normalisation			
Conv2DT	(3,3)	512	Relu	Conv2DT	(3,3)	512	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	512	Relu	Conv2D	(3,3)	512	Relu
Batch Normalisation				Batch Normalisation			
Conv2DT	(3,3)	256	Relu	Conv2DT	(3,3)	256	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	256	Relu	Conv2D	(3,3)	256	Relu
Batch Normalisation				Batch Normalisation			
Conv2DT	(3,3)	128	Relu	Conv2DT	(3,3)	128	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	128	Relu	Conv2D	(3,3)	128	Relu
Batch Normalisation				Batch Normalisation			
Conv2DT	(3,3)	64	Relu	Conv2DT	(3,3)	64	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	64	Relu	Conv2D	(3,3)	64	Relu
Batch Normalisation				Batch Normalisation			
Conv2DT	(3,3)	32	Relu	Conv2DT	(3,3)	32	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	32	Relu	Conv2D	(2,2)	32	Relu
Batch Normalisation				Batch Normalisation			
Conv2D	(3,3)	1	Sigmoid	Conv2D	(3,3)	1	Sigmoid

Segmentation Results

Table 2.6 gives the segmentation accuracy of the autoencoders (i.e., MobileNet-based, and VGG16-based) with the TR, RCR, RLR, RNR, RWR and RRR data augmentation techniques on the Supervisely Person dataset, including the comparison with the baseline results (i.e., the one obtained on the original samples without using data augmentation). In contrast to the classification results, the results in Table 2.6 show that all the augmentation techniques tested did not improve the segmentation performance. This might imply that using rotation alone to augment data might not be a good method for the segmentation task, particularly if the shape feature was the most important one in the dataset, as in the Supervisely Person dataset. For further investigation into the influence of different features on the performance of the augmentation techniques, we conducted an experiment using the Nuclei images dataset Hamilton (2018).

Differing from the results obtained in Table 2.6 on the Supervisely Person dataset, the results in Table 2.7 on the Nuclei images dataset demonstrated that the rotation augmentation methods could improve the segmentation performance. This performance gain might be due to the fact that the augmentation techniques did not degrade the image qualities much in the Nuclei dataset, since the shape feature was not that critical, compared to the Supervisely Person dataset. In the Nuclei Images dataset, colours and textures are likely to be more essential than the shape features. In particular, the segmentation results in Table 2.7 on the Nuclei dataset showed that RLR achieved the best performance among the rotation augmentation methods. This might be due to the information preservation ability that RLR provided, whereas RRR, RWR and RNR either lost parts of the information in the image's periphery or repeated some parts of the image, see Figure 2.18.

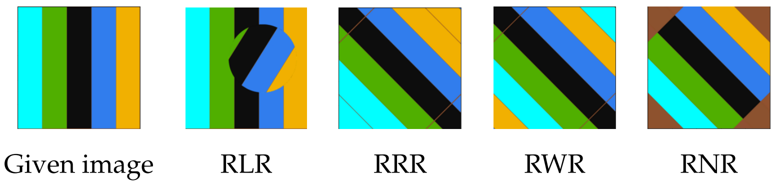

Given image RLR RRR RWR RNR

FIGURE 2.18: The effect of different rotation methods on the rotated image. The RRR and RWR expanded the central region (i.e., the black stripe) by repeating parts of it. RNR resulted in the loss of image content at the image's periphery and the creation of artificial pixel values to fill the gap. In contrast, RLR manipulated the content of the image while preserving the information around the image's periphery well.

2.6.3 Discussion

The vast majority of researchers combine many data augmentation techniques to obtain a final result. This makes it difficult to acquire an accurate evaluation for these

TABLE 2.6: Segmentation accuracy comparison between different data augmentation techniques (i.e., TR, RCR, RLR, RNR, RWR, and RRR). MobileNet-based and VGG16-based autoencoders were applied on the Supervisely Person dataset. The results indicated that using rotation solely to augment data might not be a good for the segmentation task in this case.

Model	**Baseline**	**RLR**	**TR**	**RCR**	**RNR**	**RWR**	**RRR**
MobileNet	75.13 ± 0.25	**72.12** ± 0.21	73.75 ± 0.15	72.72 ± 0.18	71.12 ± 0.17	71.27 ± 0.10	71.09 ± 0.12
VGG16	75.42 ± 0.16	**72.56** ± 0.19	73.16 ± 0.22	72.66 ± 0.24	71.31 ± 0.08	71.22 ± 0.14	71.06 ± 0.15

TABLE 2.7: Segmentation accuracy comparison between different data augmentation techniques (i.e., TR, RCR, RLR, RNR, RWR, and RRR). MobileNet-based and VGG16-based autoencoders were applied on the Nuclei images dataset. The results indicated that using rotation to augment data could enhance the segmentation performance in this case.

Model	**Baseline**	**RLR**	**TR**	**RCR**	**RNR**	**RWR**	**RRR**
MobileNet	94.25 ± 0.05	**97.6** ± 0.08	95.37 ± 0.11	94.50 ± 0.28	95.28 ± 0.19	95.24 ± 0.14	95.43 ± 0.13
VGG16	94.24 ± 0.12	**97.81** ± 0.09	95.36 ± 0.21	94.74 ± 0.17	95.08 ± 0.18	94.91 ± 0.15	95.32 ± 0.16

techniques individually. In this study, we chose the random rotation technique and examined it in more detail, along with its impact on two significant tasks (i.e., classification and segmentation), in order to make a contribution to the data augmentation regime in general. Segmentation and classification are two distinct tasks. The notion that both rely on the same features to attain their desired outcomes may not be accurate. Our results in the previous section showed that the rotation augmentation techniques could enhance methods' performance for the classification task, but not the segmentation task. It was observed that the segmentation task naturally relied on shape features Bajcsy et al. (1990). Geirhos et al. (2018) conducted a quantitative experiment demonstrating that CNNs trained with ImageNet had a strong inclination to classify texture over shape. This feature distinction might account for the disparity between classification and segmentation results when the rotation augmentation techniques were applied. In particular, in the segmentation experiment, the RLR method distorted the shape of the human body the most, yielding a slightly poorer result than that of the TR method, which did not distort the shape of the human body. The distortion of the shape feature might explain the deterioration of the segmentation results when applying the rotation augmentation techniques. In contrast, for the classification task, the rotation augmentation techniques altered the object shape but not the overall texture, which benefited the performance enhancement for the classification task.

2.7 Conclusions

In this Chapter, we explored the role of data augmentation techniques in improving feature extraction for classification and segmentation tasks. One of the key findings was that using data augmentation techniques that distort shape features negatively impacts segmentation tasks, unlike classification tasks. This finding highlights the critical role of human segmentation in the overall research, as segmentation relies heavily on preserving object shape information. To exploit this finding, an idea of integrating semantic segmentation with another strategy, such as transfer learning, needs to be explored. The importance of transfer learning in enhancing model performance is particularly evident in scenarios where labeled data is scarce or when domain adaptation is required. Transfer learning enables models to leverage prior knowledge from large-scale datasets, leading to improved generalizability and robustness in downstream tasks. This is especially beneficial in HAR, where data variability and complexity pose significant challenges to effective model training.

Given that segmentation models inherently focus on shape-based features, an interesting direction to explore is whether transfer learning from human segmentation tasks can improve HAR performance. Since shape features are vital for segmentation and also play a significant role in action recognition, leveraging pre-trained segmentation models for HAR may provide a more structured and informative feature representation. Building on these insights, Chapter 3 introduces TransNet, a novel HAR architecture that integrates transfer learning from human segmentation models. By utilizing autoencoders trained on segmentation tasks, TransNet aims to refine feature extraction for HAR, enhancing both efficiency and classification accuracy. This chapter investigates how segmentation-driven transfer learning can enhance HAR, bridging the gap between semantic segmentation and action recognition through a structured learning approach.

Chapter 3

TransNet: A Transfer Learning-Based Network for Human Action Recognition

Human action recognition (HAR) is a high-level and significant research area in computer vision due to its ubiquitous applications. The main limitations of the current HAR models are their complex structures and lengthy training time. In this chapter, we propose a simple yet versatile and effective end-to-end deep learning architecture, coined as *TransNet*, for HAR. TransNet decomposes the complex 3D-CNNs into 2D- and 1D-CNNs, where the 2D- and 1D-CNN components extract spatial features and temporal patterns in videos, respectively. Benefiting from its concise architecture, TransNet is ideally compatible with any pretrained state-of-the-art 2D-CNN models in other fields, being transferred to serve the HAR task. In other words, it naturally leverages the power and success of transfer learning for HAR, bringing huge advantages in terms of efficiency and effectiveness. Extensive experimental results and the comparison with the state-of-the-art models demonstrate the superior performance of the proposed TransNet in HAR in terms of flexibility, model complexity, training speed and classification accuracy.

3.1 Introduction

The computer vision community has studied video analysis for decades, including action recognition Tran et al. (2015) and activity understanding Kitani et al. (2012). Human action recognition (HAR) analyses and detects actions from unknown video sequences. Due to the rising demand for automated behaviour interpretation, HAR has gained dramatic attention from academics and industry and is crucial for many applications Paul and Singh (2014).

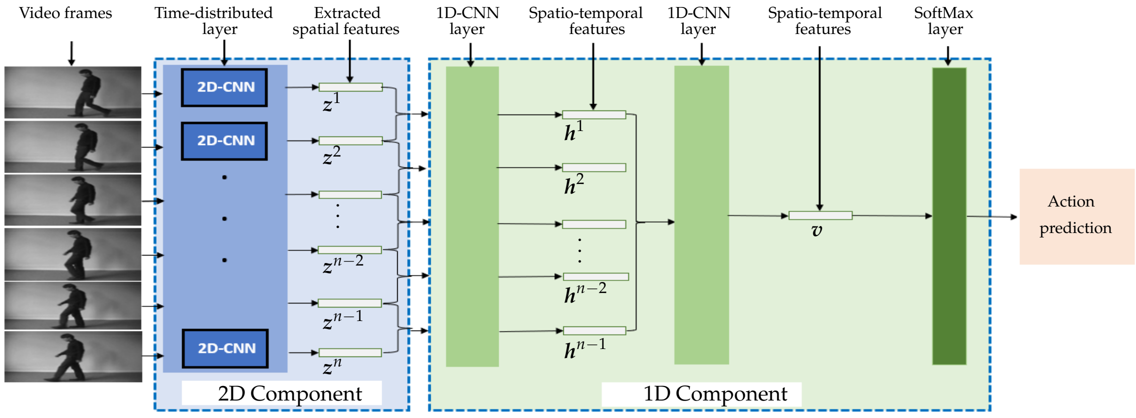


FIGURE 3.1: TransNet architecture for HAR. The given video frames are input into the time-distributed layer, which employs a 2D-CNN model (e.g., MobileNet, MobileNetV2, VGG16, or VGG19) several times based on the number of video frames, allowing the architecture to analyse multiple frames without expanding in size. Then the spatial features corresponding to the individual input frames are generated, which are subsequently analysed by the 1D-CNN layers, extracting the spatio-temporal features. The SoftMax layer finally classifies the action according to the spatio-temporal pattern.

Good action recognition requires extracting spatial features from the sequenced frames (images) of a video and then establishing the temporal correlation (i.e., temporal features) between these spatial features. Thus, action recognition models analyse two types of features, establish their relationship, and classify complex patterns. This makes these models vulnerable to a number of significant challenges, including i) the limited ability to transfer learning exploiting advanced models from other fields in computer vision, ii) the need for large volumes of data due to the model complexity, iii) the need for accurate temporal analysis of spatial features, and iv) the overlap of moving object data with cluttered background data Jegham et al. (2020).

The improvement process across generations of these models is inconsistent Simonyan and Zisserman (2014a). This results in a wide range of works that may face difficulty of transferring learning ability between generations, especially when these models are constructed differently and/or developed in different fields for extracting specific spatial features in HAR.

Temporal modeling presents a big challenge in action recognition. To address this, researchers often employ 3D-CNN models, which excel at interpreting spatio-temporal characteristics but suffer from much larger size compared to 2D-CNN models Yang et al. (2019). Moreover, optimising 3D networks becomes difficult when dealing with insufficient data Kong et al. (2021), since training a 3D convolutional filter necessitates a substantial dataset encompassing diverse video content and action categories Hu et al. (2021). Unlike recurrent neural networks (RNNs) that emphasise temporal patterns Narang et al. (2017), 3D networks analyse videos as 3D images, potentially compromising the sequential analysis of temporal data. Both 3D-CNNs

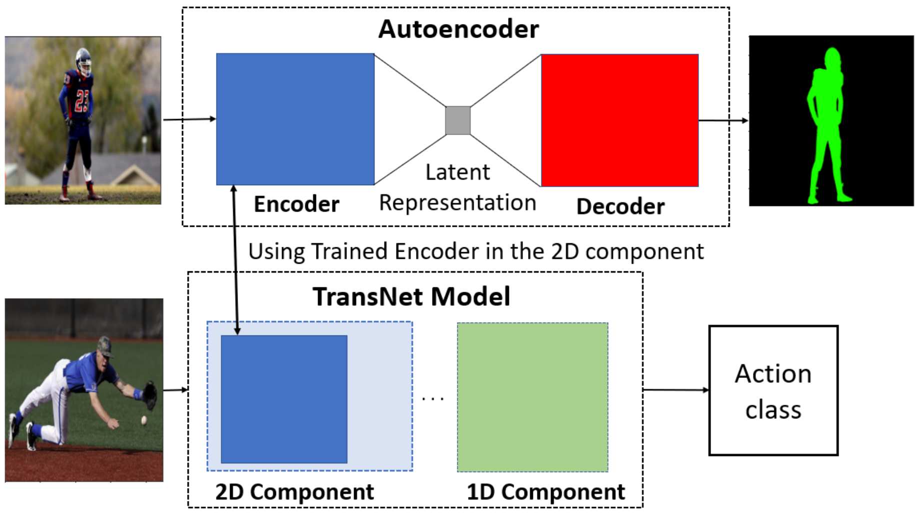


FIGURE 3.2: An illustration of TransNet+ for HAR. TransNet+ inherits the architecture of TransNet. It uses the autoencoder's encoder to form the TransNet's 2D component.

and RNNs are challenged by the increased model size and lengthy training time Stamoulakatos et al. (2021).

The presence of cluttered backgrounds presents another challenge in HAR. Indoor environments with static and constant backgrounds are typically assumed to yield high performance for HAR models, whereas performance could significantly diminish in outdoor contexts Liu et al. (2015); Wu et al. (2011). Cluttered backgrounds introduce interruptions and background noise, encoding problematic information in the extraction of global features and leading to a notable decline in performance. To address this challenge, a practical approach is to design models that focus on the human object rather than the background. Scholarly literature consistently indicates that incorporating multiple input modalities, including optical flow and body part segmentation, shows promise in enhancing HAR performance. This conclusion is substantiated by a range of survey studies conducted in the field of action recognition, providing robust evidence for the effectiveness of leveraging diverse input modalities Beddiar et al. (2020); Kong and Fu (2022); Sun et al. (2022).

However, there are several issues with these types of models, including their various modelling steps, preprocessing stages, lengthy training time, and significant demands on resources such as memory and processing power. These models are also difficult to implement in real-world applications.

In this chapter, we propose an end-to-end deep learning architecture called *TransNet* for HAR, see Figure 3.1. Rather than using complex 3D-CNNs, TransNet consists of 2D- and 1D-CNNs that extract spatial features and temporal patterns in videos, respectively. TransNet offers the following multiple benefits: i) a single network stream using only RGB frames; ii) transfer learning ability and flexibility because of its compatibility with any pretrained state-of-the-art 2D-CNN models for spatial feature

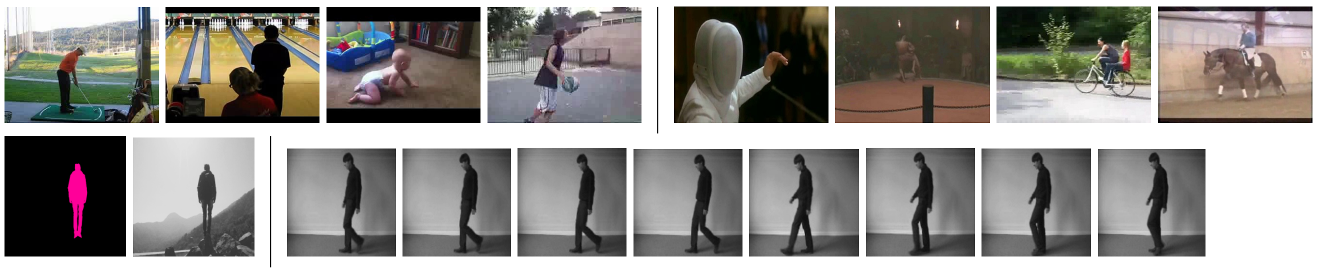


FIGURE 3.3: Data samples. First row: samples of UCF101 actions (left) and HMDB51 actions (right); second row: samples of the supervisely person dataset (left) and a frame sequence of the action class "walking" from the KTH dataset (right).

extraction; iii) a customisable and simpler architecture compared to existing 3D-CNN and RNN models; and iv) fast learning speed and state-of-the-art performance in HAR. These benefits allow TransNet to leverage the power and success of transfer learning for HAR, bringing huge advantages in terms of efficiency and effectiveness.

An additional contribution in this chapter is that we introduce the strategy of utilising autoencoders to form the TransNet's 2D component, i.e., named *TransNet+*, see Figure 3.2. TransNet+ employs the encoder part of the autoencoder trained on computer vision tasks like human semantic segmentation (HSS) to conduct HAR. Extensive experimental results and the comparison with the state-of-the-art models demonstrate the superior performance of the proposed TransNet/TransNet+ in HAR.

3.2 Related Work

3.2.1 HAR with Background Subtraction

Most research on HAR focuses on human detection and motion tracking Jaouedi et al. (2020). Background subtraction has been suggested in a number of methods and proven to be viable for HAR. For example, a background updating model based on a dynamic optimisation threshold method was developed in Zhang and Liang (2010) to detect more complete features of the moving object. The work in Kim et al. (2018a) introduced a basic framework for detecting and recognising moving objects in outdoor CCTV video data using background subtraction and CNNs. Jaouedi et al. (2020) employed a Gaussian mixture model and Kalman filter Liu et al. (2007) techniques to detect human motion by subtracting the background.

3.2.2 HAR with Multimodality

Since video comprehension requires motion information, researchers have integrated several input modalities in addition to RGB frames to capture the correlation between frames in an effort to enhance model performance.

Optical Flow: Optical flow Horn (1981), which effectively describes object or scene motion flow, is one of the earliest attempts to capture temporal patterns in videos. In comparison to RGB images, optical flow may successfully remove the static background from scenes, resulting in a simpler learning problem than using RGB images as input Diamantas and Alexis (2020); Wang et al. (2018b). Simonyan and Zisserman (2014a) began the trend of using multiple input modalities, including optical flow, with CNNs. However, when compared to the latest deep learning techniques, optical flow has a number of disadvantages, including being computationally complex and highly noise-sensitive Bovik (2009); Li and Wang (1998), which make its use in real-time applications less feasible.

Semantic Segmentation: Semantic segmentation is a technique that may be used to separate either the entire body or particular body parts from the background Minaee et al. (2021). It is a pixel-wise labelling of a 2D image, offering spatial features describing the shape of the object of interest Ulku and Akagündüz (2022). Zolfaghari et al. (2017) presented a chained multi-stream model that pre-computes and integrates appearance, optical flow, and human body part segmentation to achieve better action recognition and localisation. Benitez-Garcia et al. (2021) offered an alternative to the costly optical flow estimates used in multimodal hand gesture recognition methods. It was built using RGB frames and hand segmentation masks, with better results achieved.

Although semantic segmentation approaches have shown promising outcomes in action recognition, the majority of them are computationally demanding. In fact, real-world action recognition methods involving semantic segmentation of video content are still in their infancy phase Zhang et al. (2019).

In sum, most of the aforementioned research focused on creating synthetic images that reflect different input modalities and then analysing them using action recognition models. Pre-computing multiple input modalities such as optical flow, body part segmentation, and semantic segmentation can be computationally and storage-intensive, making them unsuitable for large-scale training and real-time deployment. Since research in the subject of semantic segmentation may still be in its early stage, one of the objectives of this study is to enhance its potential in HAR.

3.2.3 3D-CNNs Decomposition

Video can be conceptually simplified by viewing it as a 3D tensor with two spatial dimensions and one time dimension. As a result, 3D-CNNs are adopted to model the spatial and temporal data in video as a processing unit Yao et al. (2019); Ji et al. (2012); Kalfaoglu et al. (2020). Ji et al. (2012) proposed the pioneer work in the application of

3D-CNNs in action recognition. Although the model's performance is encouraging, the network's depth is insufficient to demonstrate its potential. Tran et al. (2015) extended the work in Ji et al. (2012) to a 3D network with more depth, called C3D. C3D adopts the modular architecture, which can be viewed as a 3D version of the VGG16 network.

It is worth noting that training a sufficiently deep 3D-CNN from scratch will result in much higher computational cost and memory requirements compared to 2D-CNNs. Furthermore, 3D networks are complex and difficult to optimise Kong et al. (2021); therefore, a big dataset with diverse video data and activity categories is required to train a 3D-CNN effectively. In addition, it is not straightforward for 3D-CNNs to transfer learning from state-of-the-art pretrained 2D-CNN models since kernel shapes are completely different. Carreira and Zisserman (2017) proposed I3D, a 3D-CNN architecture that circumvents the dilemma that 3D-CNNs must be trained from scratch. A strategy was employed to transform the weights of pretrained 2D models, e.g. on ImageNet, to their 3D counterparts. To understand this intuitively, they repeated the weights of the trained 2D kernels along the time dimension of the 3D kernels. Although I3D was successful in overcoming the challenge of spatial transfer-learning, its 3D kernels require enormous quantities of action recognition datasets to capture temporal features. Moreover, the way that I3D stretches 2D-CNN models into 3D-CNNs remains computationally expensive.

P3D by Qiu et al. (2017) and R2+1D by Tran et al. (2018) investigate the concept of decomposing the 3D CNN's kernels into 2D and 1D kernels. They differ in their arrangement of the two factorised operations and their formulation of each residual block. This kind of approach to 3D network decomposition acts at the kernel level. The notion of kernel-level factorisation restricts the ability to switch models (e.g., ResNet50 and VGG16) based on implementation requirements and hinders transfer learning from the current state-of-the-art models.

3.3 Proposed TransNet

In this section, we first present our motivations and then introduce the proposed TransNet and its variants.

3.3.1 Preliminary

Video data analysis in deep learning commonly involves two types of approaches: 2D-CNN-RNN Yang et al. (2020); Fan et al. (2016); Abdullah et al. (2020); Rangasamy et al. (2020) and 3D-CNN Diba et al. (2016); Hegde et al. (2018); Hou et al. (2019). The

CNN-RNN approach comprises a spatial component based on 2D-CNN and a temporal component based on RNN, offering customisation in the 2D-CNN part. However, it often requires longer training time due to the complexity of RNN compared to CNN Prokhorov et al. (2002). On the other hand, 3D-CNN is faster and simpler to implement but struggles with convergence and generalisation when dealing with limited datasets Wang et al. (2020). Alternatively, the implementation of 1D-CNN in temporal data analysis holds promise for developing more accurate and efficient models Martin et al. (2021); Liu et al. (2021).

The other main motivation is transfer learning, applying well-designed and well-trained models learnt from one task (i.e., the source task, generally with large data available) to another (i.e., the target task, generally with limited data available) for performance enhancement Zhang et al. (2017b). The underlying essential assumption is that the source and target tasks are sufficiently similar Zhang et al. (2017b); Taylor and Stone (2009). In the data scarcity scenario, models may be prone to overfitting, and data augmentation may not be enough to resolve the issue Zhang et al. (2021d). Therefore, transfer learning could play a key role in this regard.

Since HAR requires 3D data analysis, obtaining an optimised model requires training on a large amount of data compared to 2D data Habibie et al. (2019); Hu et al. (2021). This calls for the use of transfer learning, e.g., pre-training state-of-the-art models first to classify 2D images using large datasets such as ImageNet. However, it is important to study and verify the assumption that the HAR task shares sufficient similarities with the image classification task. Previous research in Geirhos et al. (2018) has shown disparities between CNNs trained on ImageNet and human observers in terms of shape and texture cues, with CNNs exhibiting a strong preference for texture over shape. Similar findings have been reported in other studies, such as Baker et al. (2018); Brendel and Bethge (2019). Additionally, several studies suggest that object shape representations hold greater importance in action recognition tasks Hirota and Komuro (2021); Zhang et al. (2021b); Dhiman and Vishwakarma (2020); El-Ghaish et al. (2018).

3.3.2 Methodology

TransNet: We propose to construct a paradigm of utilising the synergy of 2D- and 1D-CNNs, see Figure 3.1 for the end-to-end *TransNet* architecture. TransNet provides flexibility to the 2D-CNN component in terms of model customisability (i.e., using different state-of-the-art 2D-CNN models) and transferability (i.e., involving transfer learning); moreover, it benefits from the 1D-CNN component supporting the development of faster and less complex action recognition models.

TransNet includes the time-distributed layer wrapping the 2D-CNN model. In particular, the 2D component is customisable, and any sought-after 2D-CNN models (e.g., MobileNet, MobileNetV2, VGG16 or VGG19) can be utilised. The time-distributed layer is followed by three 1D convolutional layers for spatio-temporal analysis. In detail, the first one's kernels process the feature map vectors over $(n-1)$ steps, where each kernel has a size of 2, capturing the correlations between a frame and its neighbour, and n is the number of frames in a video clip; the second one's kernels have a size of $(n-1)$, analysing all feature vectors in one step to capture the whole temporal pattern of the frame-sequence; and the third one uses the SoftMax function for the final classification, followed by the flatten layer. More details are given below.

We first define the symbols used for semantic segmentation. Let $\boldsymbol{X}$ represent the input image, and $\boldsymbol{z} = p_\theta(\boldsymbol{X}) \in \mathbb{R}^L$ be the output vector (i.e., latent representation) of the encoder function p_θ (e.g. MobileNet or VGG16) with parameters θ. The decoder function is defined analogously. The formed autoencoder can then be trained with the ground truth images.

Let $\mathcal{X}$ be a collection of n frames $\mathcal{X} = \{\boldsymbol{X}^i\}_{i=1}^n$, which is fed into the 2D component (spatial component) of the TransNet architecture in Figure 3.1. The trained encoder p_θ is then used n times to process $\mathcal{X}$ frame by frame and create a set of n spatial feature vectors $\mathcal{Z} = \{\boldsymbol{z}^i\}_{i=1}^n$, where $\boldsymbol{z}^i = p_\theta(\boldsymbol{X}^i)$. Let $\{\boldsymbol{w}^{j,1}, \boldsymbol{w}^{j,2}\}_{j=1}^K$ be a set of weights, where $\boldsymbol{w}^{j,1}, \boldsymbol{w}^{j,2} \in \mathbb{R}^L$. The first of the three 1D layers (i.e., the temporal component) processes every two adjacent spatial vectors of $\mathcal{Z}$, i.e., $\{\boldsymbol{z}^i, \boldsymbol{z}^{i+1}\}$, to generate the corresponding spatio-temporal feature vectors $\boldsymbol{h}^i = (h_1^i, \cdots, h_K^i) \in \mathbb{R}^K, i = 1, \ldots, n-1$, where

$$h_j^i = f(\sum_{l=1}^{L}\sum_{k=i}^{i+1} z_l^k w_l^{j,k-i+1} + b_i^j), \;\; j = 1, \ldots, K,$$

b_i^j are the biases and f is the activation function (i.e., Relu $f(x) = \max(0, x)$ is used here). Let $\{\hat{\boldsymbol{w}}^{j,1}, \hat{\boldsymbol{w}}^{j,2}, \cdots, \hat{\boldsymbol{w}}^{j,n-1}\}_{j=1}^C$ be another set of weights, with $\hat{\boldsymbol{w}}^{j,k} \in \mathbb{R}^K, k = 1, \ldots, n-1$. The second 1D layer processes the set of spatio-temporal vectors $\{\boldsymbol{h}^i\}_{i=1}^{n-1}$ to generate a single spatio-temporal vector $\boldsymbol{v} = (v_1, \cdots, v_C) \in \mathbb{R}^C$, where

$$v_j = f(\sum_{l=1}^{K}\sum_{k=1}^{n-1} h_l^k \hat{w}_l^{j,k} + \hat{b}^j), \;\; j = 1, \ldots, C,$$

and $\hat{b}^j$ are the biases. Finally, the Softmax layer is used on $\boldsymbol{v}$ to classify action classes.

TransNet+. Except for using a sought-after 2D-CNN for TransNet's 2D component, below we present a way of leveraging transfer learning for it. To do so, we construct an autoencoder where TransNet's 2D component serves as its encoder. The autoencoder is then trained on a specific computer vision task such as HSS to extract specific features such as human shape, e.g., see the left of the second row in Figure 3.3.

TABLE 3.1: TransNet's model complexity. The last column gives the total number of parameters for each setting. The 2D component column reflects the model size of the time-distributed layer, which is invariant against the number of frames in the input video clip. The two values in the 1D component column show the number of kernels in the first and second 1D-CNN layers, respectively. In our experiments, we equipped TransNet with MobileNetV1 and 64 filters.

2D model	2D component	1D component	Total
MobileNetV1	6,444,288	64; 6	6,449,416
	9,655,616	128; 6	9,673,992
MobileNetV2	6,277,248	64; 6	6,282,376
	10,291,392	128; 6	10,309,768
VGG16	16,322,432	64; 6	16,327,560
	17,928,128	128; 6	17,946,504
VGG19	21,632,128	64; 6	21,637,256
	23,237,824	128; 6	23,256,200

After training, the encoder's parameters become saturated with weights that are capable of describing the features of the required task, such as HAR, see Figure 3.2. In this way, the features like object shape that TransNet's 2D component needs to learn can be predetermined by training the autoencoder. We name this way of executing TransNet as *TransNet+*.

Note that autoencoders have been used in action recognition challenges e.g. Zolfaghari et al. (2017). However, there are a number of disadvantages in their use of autoencoders, including the multiplicity of modelling steps, the need for a large amount of memory, and the lengthy training time due to the high computational cost of training the autoencoder network and action recognition network.

In contrast, TransNet+ is a huge step further in contributing to the development of an end-to-end HAR model with potential in real-time implementation, since it simplifies the process by just integrating the trained encoder rather than the entire autoencoder in TransNet, with the premise that the trained encoder carries weights capable of describing important features (see Figure 3.2).

On the whole, the traditional method of using autoencoders in HAR differs from TransNet+ in that the traditional method uses the entire autoencoder and its output as the next stage's input, whereas TransNet+ just employs the trained encoder of the autoencoder for spatial feature extraction.

Model Complexity: The proposed TransNet model is customisable, and thus its size varies depending on the 2D-CNN model being used in the spatial component. In particular, it is quite cost-effective since it uses a time-distributed layer, allowing the 2D-CNN to be used repeatedly without expanding in size. Table 3.1 gives the number of parameters regarding different choices of the 2D-CNN models.

TABLE 3.2: Results of TransNet with different backbones and different pretraining strategies on the KTH dataset.

TransNet backbone	Pre-training	Acc.
MobileNet	None	94.35 ± 0.33
	ImageNet	**100.00** ± 0.00
	HSS	**100.00** ± 0.00
MobileNetV2	None	88.31 ± 0.54
	ImageNet	95.86 ± 0.41
	HSS	**96.40** ± 0.34
VGG16	None	90.12 ± 0.38
	ImageNet	96.25 ± 0.43
	HSS	**98.01** ± 0.48
VGG19	None	80.06 ± 0.72
	ImageNet	88.26 ± 0.51
	HSS	**94.39** ± 0.26

3.4 Data

In our study, we use two primary groups of benchmark datasets. The first consists of ImageNet and the supervisely person dataset used for transfer learning, while the second consists of the KTH, HMDB51 and UCF101 datasets used for method evaluation (with a split ratio of 80% and 20% for training and test, respectively); see below Figure 3.3 for a brief description and for some samples from these datasets.

3.4.1 Transfer Leaning Datasets

ImageNet: ImageNet Deng et al. (2009) is a famous database consisting of 14,197,122 images with 1000 categories. Since 2010, it has been used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

Supervisely Person Dataset: This dataset supervise.ly (2018) is publicly available for human semantic segmentation, containing 5,711 images and 6,884 high-quality annotated human instances. It is available for use in academic research with the purpose of training machines to segment human bodies.

3.4.2 Human Action Recognition Datasets

KTH: In 2004, the Royal Institute of Technology introduced KTH, a non-trivial and publicly available dataset for action recognition Schuldt et al. (2004). It is one of the most standard datasets, including six actions (i.e., walking, jogging, running, boxing, hand-waving, and hand-clapping). Twenty-five different people did each activity, allowing for variation in performance; moreover, the setting was systematically changed for each actor's action, i.e., outdoors, outdoors with scale variation, outdoors with varied clothing, and indoors. KTH includes 2,391 sequences. All sequences were captured using a stationary camera with 25 fps over homogeneous backgrounds.

UCF101: In 2012, UCF101 was introduced as a follow-up to the earlier UCF50 dataset. It is a realistic (not staged by actors) HAR dataset, containing 13,320 YouTube videos representing 101 human actions. It provides a high level of diversity in terms of object appearance, significant variations in camera viewpoint, object scale, illumination conditions, a cluttered background, etc. These video clips are, in total, over 27 hours in duration. All videos have a fixed frame rate of 25 fps at a resolution of 320 × 240 Soomro et al. (2012).

HMDB51: HMDB51 was released in 2011 as a realistic HAR dataset. It was primarily gathered from movies, with a small portion coming from public sources such as YouTube and Google videos. It comprises 6,849 videos organised into 51 action categories, each with at least 101 videos Kuehne et al. (2011).

3.5 Experimental Results

3.5.1 Settings

Our model is built using Python 3.6 with the deep learning library Keras, the image processing library OpenCV, matplotlib, and the scikit-learn library. A computer with an Intel Core i7 processor, an NVidia RTX 2070, and 64GB of RAM is used for training and test.

Four CNN models with small sizes (i.e., MobileNet, MobileNetV2, VGG16, and VGG19) are selected as the backbones of TransNet/TransNet+, with parameter numbers of 3,228,864, 2,258,984, 14,714,688, and 20,024,388 (without the classification layers), respectively.

TransNet with each different backbone is implemented in three different ways: i) untrained; ii) being trained on ImageNet; and iii) being trained on HSS using the supervisely person datasetas as encoders. Note that the last way is described in TransNet+. For ease of reference, we drop the '+' sign in the following. The number of 200 epochs is used to train all autoencoders, with a batch size of 24. The models are first trained and evaluated on the KTH dataset. Then the one with the best performance is selected to be evaluated on all the datasets, and compared with the current state-of-the-art HAR models. Each video clip consists of a sequence of 12 frames, and the input modality is RGB with a size of 224 × 224 × 3.

TABLE 3.3: Results comparison between TransNet and the state-of-the-art methods on the KTH dataset. TransNet model is pretrained in ImageNet.

Method	Model	Venue	Acc.
Grushin et al. (2013)	LSTM	IJCNN '13	90.70
Shu et al. (2014)	Spiking neural networks (SNN)	IJCNN '14	92.30
Geng and Song (2016)	2D-CNN	ICCSAE '15	92.49
Veeriah et al. (2015)	LSTM	ICCV '15	93.96
Zhang et al. (2017a)	Hybrid feature approach	ISMEMS '17	95.00
Arunnehru et al. (2018)	3D-CNN	RoSMa '18	94.90
Abdelbaky and Aly (2020)	PCA-based filters	ITCE '20	87.52
Jaouedi et al. (2020)	Gated recurrent neural network	KSUCI journal '20	96.30
Sahoo et al. (2020)	Deep Bidirectional LSTM with 2D-CNN	TETCI '20	97.67
Ramya and Rajeswari (2021)	Feed-Forward Neural Networks	MTA journal '21	91.40
Lee et al. (2021)	ConvLSTMs with 3D-CNN	CVF '21	89.40
Basha et al. (2022)	3D-CNN with LSTM	MTA journal '22	96.53
Picco et al. (2023)	HOG with PCA	NN journal '23	90.83
Ye and Bilodeau (2023)	2D-CNN	CVF '23	90.90
TransNet	3D-CNN	-	**100.00**

TABLE 3.4: Results comparison between TransNet and the state-of-the-art methods on the UCF101 dataset.

Model	Pre-training	Backbone	Venue	Acc.
DeepVideo Karpathy et al. (2014)	ImageNet	AlexNet	CVPR '14	65.40
Two-stream Simonyan and Zisserman (2014a)	ImageNet	CNN-M	NeurIPS '14	88.00
LRCN Zhu et al. (2020b)	ImageNet	CNN-M	CVPR '15	82.30
C3D Tran et al. (2015)	ImageNet	VGG16-like	ICCV '15	82.30
Fusion Feichtenhofer et al. (2016)	ImageNet	VGG16	CVPR '16	92.50
TSN Zhu et al. (2020b)	ImageNet	BN-Inception	ECCV '16	94.00
I3D Carreira and Zisserman (2017)	ImageNet Kinetics400	BN-Inception-like	CVPR '17	95.60
P3D Zhu et al. (2020b)	Sports1M	ResNet50-like	ICCV '17	88.60
ResNet3D Zhu et al. (2020b)	Kinetics400	ResNeXt101-like	CVPR '18	94.50
HAR-Depth Sahoo et al. (2020)	-	BiLSTM+AlexNet	TETCI '20	92.97
MemDPC Zhu et al. (2020b)	Kinetics400	R-2D3D	ECCV '20	86.10
TEA Li et al. (2020c)	Imagenet	ResNet50-like	CVPR '20	96.90
CVRL Zhu et al. (2020b)	Kinetics600	R3D-50	CVPR '21	93.40
TDN Wang et al. (2021a)	Kinetics400 ImageNet	ResNet	CVF '21	97.40
Multi-Domain Omi et al. (2022)	-	MDL	IEICE '22	94.82
MEST Zhang (2022)	Imagenet	2D-CNN	Sensors '22	96.80
STA-TSN Yang et al. (2022a)	Imagenet	ResNet-LSTM	PloS One '22	92.80
FSAN Chen et al. (2023)	Imagenet	2D-CNN	Sensors '23	95.68
TransNet	ImageNet	MobileNet V1	-	**98.32**

3.5.2 Results and Discussion

In a nutshell, we conduct experiments below with three main objectives: i) determining whether or not the proposed TransNet architecture can offer a reliable mechanism by leveraging transfer learning; ii) evaluating if the HSS-trained TransNet provides superior spatio-temporal characteristics for HAR than the ImageNet-trained TransNet; and iii) validating if the performance of the TransNet architecture can achieve state-of-the-art performance in comparison to current state-of-the-art methods in HAR.

Initially, we subject TransNet to an evaluation using the KTH dataset, which serves as a suitable choice due to its primary emphasis on human action detection while

TABLE 3.5: Results comparison between TransNet and the state-of-the-art methods on the HMDB51 dataset.

Model	Pre-training	Backbone	Venue	Acc.
C3D Tran et al. (2015)	ImageNet	VGG16-like	ICCV '15	56.80
CBT Zhu et al. (2020b)	ImageNet	S3D	arXiv '19	44.60
DPC Zhu et al. (2020b)	Kinetics400	R-2D3D	ICCVW '19	35.70
SpeedNet	Kinetics400	S3D-G	CVPR '20	48.80
MemDPC Zhu et al. (2020b)	Kinetics400	R-2D3D	ECCV '20	54.50
CoCLR Zhu et al. (2020b)	Kinetics400	S3D	NeurIPS '20	54.60
HAR-Depth Sahoo et al. (2020)	-	BiLSTM+AlexNet	TETCI '20	69.74
STA-TSN Kwon et al. (2020)	Imagenet	ResNet50-like	ECCV '20	77.40
TEA Li et al. (2020c)	Imagenet	ResNet50-like	CVPR '20	73.30
TDN Wang et al. (2021a)	Kinetics400 ImageNet	ResNet	CVF '21	76.30
Multi-Domain Omi et al. (2022)	-	MDL	IEICE '22	71.57
MEST Zhang (2022)	Imagenet	2D-CNN	Sensors '22	73.40
STA-TSN Yang et al. (2022a)	Imagenet	ResNet-LSTM	PloS One '22	68.60
FSAN Chen et al. (2023)	Imagenet	2D-CNN	Sensors '23	72.60
TransNet	ImageNet	MobileNet V1	-	**97.93**

excluding the presence of additional objects in the background, in contrast to the UCF101 and HMDB51 datasets. The purpose of this evaluation is to validate the viability of employing HSS as a means of pretraining to improve the performance of the model in similar tasks.

The results presented in Table 3.2 demonstrate the superior performance of the TransNet model which was trained using HSS in comparison to its untrained and ImageNet-trained counterparts. Specifically, the untrained MobileNet, MobileNetV2, VGG16, and VGG19-based TransNet models achieved an average accuracy of 88.21%, and the ImageNet-trained models achieved an average accuracy of 95.09%. In contrast, the HSS-trained TransNet models achieved an average accuracy of 97.20%, indicating a significant improvement of $\sim$ 8.99% and $\sim$ 2.11% over the untrained and ImageNet-trained models, respectively. These findings underscore the effectiveness of the pretraining strategy employing autoencoders in enhancing the performance of the TransNet model. Additionally, the findings show the significance of incorporating transfer learning as a means of enhancing performance, thereby bestowing a substantial advantage to the 2D-1D-CNN architecture and enabling us to leverage transfer learning within the 2D-CNN component.

Tables 3.3, 3.4 and 3.5 present the quantitative comparisons between TransNet and the current state-of-the-art methods being applied to the HAR datasets, i.e., KTH, UCF101 and HMDB51. In these comparisons, a MobileNet-based TransNet pretrained on ImageNet is used. The findings demonstrate the exceptional performance achieved by the proposed TransNet, surpassing the existing state-of-the-art results by a considerable margin. Additionally, these findings solidify the 2D-1D-CNN architecture as a highly effective approach for HAR.

3.6 Conclusion

The development and evaluation of TransNet have demonstrated the effectiveness of transfer learning and autoencoder-based feature extraction in enhancing HAR performance. By leveraging pre-trained models from segmentation tasks, TransNet successfully refined feature extraction, leading to improved classification accuracy and model generalization. This approach not only mitigated the limitations posed by data scarcity but also emphasized the importance of leveraging prior knowledge to enhance HAR systems. However, despite these advancements, several fundamental questions remain regarding the optimal architectural choices for HAR models.

The evolving landscape of deep learning presents a diverse range of architectures, each with distinct advantages and limitations. Convolutional Neural Networks (CNNs) excel in extracting spatial features from images, making them highly effective for capturing local patterns in action recognition. RNNs, particularly LSTM networks, have demonstrated strong capabilities in modeling temporal dependencies, allowing HAR models to capture sequential motion patterns. More recently, ViTs have emerged as a powerful alternative, leveraging self-attention mechanisms to model long-range dependencies with greater flexibility and efficiency. Given these developments, it becomes crucial to systematically examine how each of these architectures contributes to HAR and whether hybrid models can further enhance recognition capabilities.

In the next chapter 4, we conduct an extensive survey of CNN, RNN, and Transformer-based HAR models, tracing their evolution and analyzing their comparative strengths and weaknesses. Building on these insights, we propose a hybrid CNN-ViT architecture that integrates the spatial feature extraction capabilities of CNNs with the global self-attention mechanisms of ViTs. This hybrid approach aims to address the limitations of individual models while enhancing the overall robustness and efficiency of HAR systems. Through this investigation, Chapter 4 seeks to bridge the gap between existing methodologies and future directions in deep learning-based HAR.

Chapter 4

CNNs, RNNs and Transformers in Human Action Recognition: A Survey and a Hybrid Model

Human action recognition encompasses the task of monitoring human activities across various domains, including but not limited to medical, educational, entertainment, visual surveillance, video retrieval, and the identification of anomalous activities. Over the past decade, the field of HAR has witnessed substantial progress by leveraging CNNs and RNNs to effectively extract and comprehend intricate information, thereby enhancing the overall performance of HAR systems. Recently, the domain of computer vision has witnessed the emergence of ViTs as a potent solution. The efficacy of Transformer architecture has been validated beyond the confines of image analysis, extending their applicability to diverse video-related tasks. Notably, within this landscape, the research community has shown keen interest in HAR, acknowledging its manifold utility and widespread adoption across various domains. This article aims to present an encompassing survey that focuses on CNNs and the evolution of RNNs to ViTs given their importance in the domain of HAR. By conducting a thorough examination of existing literature and exploring emerging trends, this study undertakes a critical analysis and synthesis of the accumulated knowledge in this field. Additionally, it investigates the ongoing efforts to develop hybrid approaches. Following this direction, this article presents a novel hybrid model that seeks to integrate the inherent strengths of CNNs and ViTs.

4.1 Introduction

Human action recognition (HAR) focuses on the classification of the specific actions exhibited within a given video. On the other hand, action detection and segmentation focus on the precise localization or extraction of individual instances of actions from video content Ulhaq et al. (2022). The capacity of deep learning models to effectively capture the spatial and temporal complexities inherent in video representations plays a vital role in the recognition and understanding of actions.

Over the preceding decade, a considerable amount of research has been dedicated to the thorough investigation of action recognition, resulting in an extensive collection of review articles and survey papers addressing the topic Pareek and Thakkar (2021); Sun et al. (2022); Kong and Fu (2022). However, it is worth noting that a predominant focus of these scholarly works has been placed on the examination and evaluation of CNNs and traditional machine learning models within the realm of action recognition.

The advent of Transformer architecture Vaswani et al. (2017) has sparked a paradigm shift in deep learning. By employing a multi-head self-attention layer, the Transformer model computes sequence representations by effectively aligning words within the sequence with other words in the same sequence Ulhaq et al. (2022). This approach outperforms traditional convolutional and recursive operations in terms of representation quality while utilizing fewer computational resources. As a consequence, the Transformer architecture diverges from conventional convolutional and recursive methods, favoring a more focused utilization of multiple processing nodes. The incorporation of multi-head attention allows the Transformer model to collectively learn a range of representations from diverse perspectives through the collaboration of multiple attention layers. Inspired by Transformers, many natural language processing (NLP) tasks have achieved remarkable performance, reaching human-level capabilities, as exemplified by models such as GPT Brown et al. (2020) and BERT Devlin et al. (2018).

The remarkable achievements of Transformers in handling sequential data, particularly in the domain of NLP, have prompted the exploration and advancement of Vision Transformer (ViT) Dosovitskiy et al. (2020) (a special Transformer for computer vision tasks). ViTs have demonstrated comparable or even superior performance compared to CNNs in the context of image recognition tasks, especially when operating on vast datasets such as ImageNet Han et al. (2022); Lin et al. (2022); Khan et al. (2022). This observation signifies a noteworthy shift in the field, wherein ViTs possess the potential to supplant the established dominance of CNNs in computer vision, mirroring the displacement witnessed in the case of recurrent neural networks (RNNs) Ulhaq et al. (2022). The achievements of Transformer models have engendered considerable scholarly interest within the computer vision research community, prompting rigorous exploration of their efficacy in pure computer vision tasks.

The natural progression in the advancement of ViTs has led to the logical exploration of video recognition tasks. Unlike image recognition, video recognition focuses on the complex challenge of identifying and understanding events within video sequences, including the recognition of human actions. Consequently, there is a compelling need for a recent review that comprehensively examines the state-of-the-art research including ViTs and hybrid models in addition to CNNs and RNNs for HAR. Such a review would serve as a crucial guiding resource to shape the future research directions with Transformer and CNN-Transformer hybrid architectures beside CNNs which previously were seen as unique and influential models for HAR. The main contributions of this chapter is as follows.

- We present a thorough review of the CNNs, RNNs and ViTs. This review examines the evolution from traditional methods to the latest advancements in neural network architectures.
- We present an extensive examination of existing literature related to HAR.
- We propose a novel hybrid model integrating the strengths of CNNs and ViTs. In addition, we provide a detailed performance comparison of the proposed hybrid model against existing models. The analysis highlights the model's efficacy in handling complex HAR tasks with improved accuracy and efficiency.
- We also discuss emerging trends and the future direction of HAR technologies, emphasizing the importance of hybrid models in enhancing the interpretability and robustness of HAR systems.

These contributions enrich the understanding of the current state and future prospects of HAR, proposing innovative approaches and highlighting the importance of integrating different neural network architectures to advance the field.

The chapter is structured as follows. Section 1.5 delves into the background, covering foundational concepts and technologies crucial to HAR, including CNNs, RNNs and ViTs, highlighting the chronological evolution of HAR deep learning technologies. Section 4.2 thoroughly reviews related HAR works with a brief discussion. A novel hybrid model combining CNNs and ViTs is proposed in Section 4.3, including the details of the experimental setup and the results. Finally, Section 4.4 concludes the chapter.

4.2 Literature Review

This section briefly recalls the most commonly used deep learning-based HAR approaches.

4.2.1 CNN-Based Approaches in HAR

This section recalls the most prominent CNN-based approaches in HAR based on the model type (i.e., the two-stream CNN, 3D CNN, and RNNs with CNNs), organized chronologically.

Deep learning was still in its early stages in 2012, and CNNs or RNNs had not yet gained significant popularity in the field of HAR. The focus was primarily on traditional machine learning approaches, such as support vector machines Cortes and Vapnik (1995), and handcrafted features, such as histogram of oriented gradients Dalal and Triggs (2005) and histogram of optical flow Barron et al. (1994). A few studies did, nevertheless, start looking into neural networks for action recognition.

In 2014, the use of CNNs in action recognition was at a pivotal stage, marking a shift from hand-crafted feature-based methods to deep learning approaches. The key points of the use of CNNs in action recognition at that period of time are the following. (I) *Emergence of deep learning:* deep learning, particularly CNNs, had started to dominate image classification tasks, thanks to their ability to learn feature representations directly from raw pixel data. This success in static images paved the way for applying CNNs to video data for action recognition. (II) *Challenges in video data:* unlike 2D images, videos incorporate a third dimension which represents the temporal patterns, making action recognition more complex. CNNs had to be adapted to not only recognize spatial patterns but also capture motion information over time dimension. (III) *Datasets and benchmarks:* the adoption of large-scale video datasets like UCF-101 Soomro et al. (2012) and HMDB-51 Kuehne et al. (2011) became more common. These datasets provided diverse sets of actions and were large enough to train deep networks. The performance on these benchmarks has been becoming a key measure of progress for action recognition models. (IV) *Transfer learning:* due to the computational expense of training CNNs from scratch and the relatively smaller size of video datasets compared to image datasets, transfer learning became a popular strategy. Networks pre-trained on large image datasets like ImageNet Deng et al. (2009) were fine-tuned on video frames for action recognition tasks. (V) *Computational constraints:* despite the promise of CNNs, computational constraints were a significant challenge. Training deep networks required significant GPU power, and processing video data with CNNs was resource-intensive. This limited the complexity of the models that could be trained and the size of the datasets that could be used.

4.2.1.1 Two-Stream CNNs

Simonyan and Zisserman (2014a) presented an innovative approach to recognize actions in video sequences by using a two-stream CNN architecture. This approach divides the task into two distinct problems: recognizing spatial features from single frames and capturing temporal features across frames. The spatial stream CNN processes static visual information, while the temporal stream CNN handles motion by analyzing optical flow. The model was tested on benchmark datasets like UCF-101 and HMDB-51, where it achieved state-of-the-art results, showcasing the effectiveness of this two-stream method. The novelty of this work lies in the separation of motion and appearance features, which allows for more specialized networks that can better capture the complexities of video-based action recognition. The success of this model has made a significant impact on the field, influencing many future research directions in video understanding. Consequently, numerous methods have been proposed to enhance the the two-stream model Wang et al. (2015b); Feichtenhofer et al. (2016); Wang et al. (2016); Peng et al. (2018); Wang et al. (2017).

In 2016, building on the the two-stream CNN, Feichtenhofer et al. (2016) focused on improving the two-stream CNN by exploring various fusion strategies for combining spatial and temporal streams, resulting in better performance on the UCF-101 and HMDB-51 datasets. By enhancing fusion techniques, this work addressed the limitations of the initial two-stream model, leading to more effective integration of spatial and temporal information. Wang et al. (2016) introduced temporal segment networks (TSN). This work aimed to capture long-range temporal structures for action recognition, achieving significant improvements on the UCF-101 and HMDB-51 datasets by dividing videos into segments for comprehensive analysis. The introduction of TSN extended the temporal analysis capabilities of the two-stream CNN, enabling the capture of long-range dependencies.

In 2017, derived from the two-stream CNN, Cosmin Duta et al. (2017) proposed a three-stream method by using spatio-temporal vectors, with locally max-pooled features to enhance performance. Tested on the UCF-101 and HMDB-51 datasets, the approach demonstrated improved recognition accuracy by efficiently capturing spatio-temporal dynamics. In 2018, the efficient convolutional network for online video understanding (ECO) was introduced by Zolfaghari et al. (2018), combining the two-stream CNN approach with lightweight 3D CNNs, and focusing on efficiency and real-time processing, with high efficiency and competitive accuracy demonstrated on the Kinetics and UCF-101 datasets.

Feichtenhofer et al. (2019) introduced the SlowFast network which processes video data at varying frame rates to capture both spatial semantics and motion dynamics, achieving state-of-the-art results on the Kinetics-400 and Charades datasets. By introducing different temporal resolutions, this work innovated on the two-stream concept, capturing fine and coarse temporal details. Wang et al. (2018a) expanded on their previous work with TSN, developing a multi-stream approach that incorporated RGB, optical flow, and warped optical flow streams to model long-range temporal structures more effectively. This approach achieved state-of-the-art results by capturing both spatial and temporal information across various time scales. In 2021, temporal difference networks (TDN) were introduced by Wang et al. (2021a), leveraging the multi-stream CNN with a focus on capturing motion dynamics efficiently. Using the UCF-101 and HMDB-51 datasets, TDN achieved notable improvements by effectively modeling temporal differences. By emphasizing temporal differences, this work advanced the ability of the two-stream CNN to capture motion dynamics more effectively.

Table 4.1 presents the works discussed in this section that utilized two or more stream CNNs approaches.

TABLE 4.1: Two-stream CNN-based approaches in HAR.

Paper	Model	Dataset	Novelty
Simonyan and Zisserman (2014a)	Two-stream CNN	UCF-101, HMDB-51	Introduced the two-stream architecture separating spatial and temporal streams for effective action recognition.
Feichtenhofer et al. (2016)	Two-stream CNN	UCF-101, HMDB-51	Explored various fusion strategies to combine spatial and temporal streams, and improved performance.
Wang et al. (2016)	Two-stream CNN + TSN	UCF-101, HMDB-51	Introduced TSN to capture long-range temporal structures by dividing videos into segments.
Cosmin Duta et al. (2017)	Three-Stream CNN	UCF-101, HMDB-51	Proposed a three-stream method using spatio-temporal vectors with locally max-pooled features for enhanced performance.
Zolfaghari et al. (2018)	Two-stream CNN + 3D CNN	Kinetics, UCF-101	Combined the two-stream CNN with lightweight 3D CNNs for efficient real-time processing.
Feichtenhofer et al. (2019)	Two-stream CNN + SlowFast	Kinetics-400, Charades	Introduced SlowFast networks processing video data at varying frame rates to capture both spatial and motion dynamics.

Continued on next page

Table 4.1 – continued from previous page

Paper	Model	Dataset	Novelty
Wang et al. (2018a)	CNN-RNN, (Multi-stream TSN)	UCF101, HMDB51	Expanded on TSN by developing a multi-stream approach that incorporated RGB, optical flow, and warped optical flow streams to model long-range temporal structures more effectively.
Wang et al. (2021a)	Multi-stream CNN + TDN	Something-Something V1 and V2	Introduced TDN focusing on capturing motion dynamics efficiently.

4.2.1.2 3D CNN-Based Approaches

The foundational work conducted by Ji et al. (2012) introduced 3D CNNs for HAR, demonstrating their effectiveness in capturing spatio-temporal features on the KTH and UCF-101 datasets and outperforming traditional 2D CNNs. The work paved the way for further research on enhancing 3D convolutional models. Tran et al. (2015) introduced C3D, a generic 3D CNN for spatio-temporal feature learning, achieving state-of-the-art performance on the Sports-1M and UCF-101 datasets and highlighting the scalability and effectiveness of 3D convolutions. Building on the work by Ji et al. (2012), C3D demonstrated the potential of 3D CNNs across diverse datasets, influencing subsequent research in 3D CNNs. Varol et al. (2017) introduced long-term temporal convolutions to capture extended motion patterns. This work improved the accuracy on the UCF-101 and HMDB-51 datasets and emphasized the importance of long-term motion information. Moreover, this study extended the temporal scope of 3D CNNs, highlighting the need for capturing long-term motion for accurate action recognition. In the same year, Qiu et al. (2017) proposed pseudo-3D residual networks (P3D), which combined 2D and 3D convolutions to balance the accuracy and computational complexity. This work achieved competitive performance on the Kinetics and UCF-101 datasets. Moreover, P3D networks offered a more efficient approach by blending 2D and 3D convolutions, further refining the capabilities of 3D CNNs. Additionally, Carreira and Zisserman (2017) introduced I3D by inflating 2D convolutions to 3D, achieving significant improvements on the Kinetics dataset by leveraging ImageNet pre-training, thereby setting new performance benchmarks. I3D bridged the gap between 2D and 3D CNNs, demonstrating the benefits of transfer learning in 3D convolutional models.

Hara et al. (2018) evaluated the scalability of 3D CNNs with increased data and model sizes, demonstrating that deeper 3D CNNs can achieve better performance on the

Kinetics and UCF-101 datasets, paralleling the success of 2D CNNs on ImageNet. This study emphasized the need for larger datasets and deeper models in 3D convolutional research, highlighting the potential of 3D CNNs to retrace the historical success of 2D CNNs. Building on these insights, Diba et al. (2017) introduced a new temporal 3D ConvNet architecture with enhanced transfer learning capabilities, demonstrating superior performance on the UCF-101 and HMDB-51 datasets through architectural innovations and effective transfer learning. This work underscored the importance of architectural innovation and transfer learning, pushing the boundaries of 3D CNN performance and further advancing the field of action recognition. Tran et al. (2018) further contributed by conducting a comprehensive analysis of spatio-temporal convolutions, highlighting the benefits of factorizing 3D convolutions into separate spatial and temporal components, achieving state-of-the-art results on the Kinetics and UCF-101 datasets. This dissection provided insights that informed subsequent model designs and optimizations. In the same year, Xie et al. (2018) explored the trade-offs between speed and accuracy in spatio-temporal feature learning, proposing efficient 3D CNN variants that balance computational cost and recognition performance on the Kinetics and UCF-101 datasets. Their work highlighted the practical considerations of deploying 3D CNNs, emphasizing the need to balance speed and accuracy, thereby refining the approach to spatio-temporal feature learning. Additionally, Wang et al. (2018c) introduced non-local neural networks to capture long-range dependencies, demonstrating that non-local operations significantly improve the modeling of complex temporal relationships and enhance action recognition performance on the Kinetics and Something-Something datasets. By integrating non-local operations, this study advanced the ability of 3D CNNs to capture complex temporal patterns, further pushing the boundaries of spatio-temporal modeling.

Feichtenhofer et al. (2019) introduced SlowFast Networks, a novel approach that processes video at different frame rates to capture both slow and fast motion dynamics, and achieved state-of-the-art results on the Kinetics-400 and Charades datasets. This innovation highlighted the importance of capturing varied motion dynamics for improved video recognition. In the same year, Tran et al. (2019) presented channel-separated convolutional networks (CSN), which reduced computational complexity by separating convolutions by channel, demonstrating efficiency without sacrificing accuracy on the Kinetics and Sports-1M datasets. This approach contributed to the development of more computationally feasible models. Concurrently, Ghadiyaram et al. (2019) leveraged large-scale weakly-supervised pre-training on video data, significantly boosting performance on the IG-65M and Kinetics datasets and underscoring the potential of massive datasets in enhancing 3D CNN capabilities. Additionally, Kopuklu et al. (2019) proposed resource-efficient 3D CNNs using depthwise separable convolutions and achieved competitive accuracy with significantly reduced computational requirements on the Kinetics-400 and

UCF-101 datasets. This work emphasized the importance of optimizing 3D CNNs for computational efficiency, further advancing the field of action recognition.

Feichtenhofer (2020) proposed X3D, a family of efficient video models by expanding architectures along multiple axes. It achieved state-of-the-art performance with reduced model complexity on the Kinetics-400 and Charades datasets. X3D highlighted the significance of model efficiency in balancing performance and computational demands. In the same year, Li et al. (2020b) introduced an efficient 3D CNN with a temporal attention mechanism and achieved high accuracy with efficient computation by focusing on salient temporal features on the Kinetics-400 and UCF-101 datasets. This work demonstrated the potential of selectively focusing on important temporal features to enhance the efficiency and accuracy of 3D CNNs, further advancing the field of action recognition.

Table 4.2 presents the works discussed in this section that utilized 3D CNN approaches.

TABLE 4.2: 3D CNN-based approaches in HAR.

Paper	Model	Dataset	Novelty
Ji et al. (2012)	3D CNN	UCF-101, HMDB-51	Introduced 3D CNNs for HAR, effectively capturing spatio-temporal features and outperforming 2D CNNs.
Tran et al. (2015)	3D CNN	Sports-1M, UCF-101	Introduced C3D, a generic 3D CNN for spatio-temporal feature learning, and achieved state-of-the-art performance.
Varol et al. (2017)	3D CNN	UCF-101, HMDB-51	Introduced long-term temporal convolutions to capture extended motion patterns, and improved accuracy.
Qiu et al. (2017)	3D CNN	Kinetics, UCF-101	Proposed P3D networks combining 2D and 3D convolutions, balancing accuracy and computational complexity.
Carreira and Zisserman (2017)	3D CNN	Kinetics	Introduced I3D by inflating 2D convolutions to 3D, leveraging ImageNet pre-training for significant improvements.
Hara et al. (2018)	3D CNN	Kinetics, UCF-101	Evaluated the scalability of 3D CNNs with increased data and model sizes, and showed parallels to 2D CNN success.

Continued on next page

Table 4.2 – continued from previous page

Paper	Model	Dataset	Novelty
Diba et al. (2017)	3D CNN	UCF-101, HMDB-51	Introduced a new temporal 3D ConvNet architecture with enhanced transfer learning capabilities.
Tran et al. (2018)	3D CNN	Kinetics, UCF-101	Conducted a comprehensive analysis of spatio-temporal convolutions, and highlighted the benefits of factorizing 3D convolutions.
Xie et al. (2018)	3D CNN	Kinetics, UCF-101	Explored speed-accuracy trade-offs in spatio-temporal feature learning, and proposed efficient 3D CNN variants.
Wang et al. (2018c)	3D CNN	Kinetics, Something-Something	Introduced non-local operations to capture long-range dependencies, and improved modeling of complex temporal relationships.
Feichtenhofer et al. (2019)	3D CNN	Kinetics-400, Charades	Proposed SlowFast networks to process video at different frame rates, capturing both slow and fast motion dynamics.
Tran et al. (2019)	3D CNN	Kinetics, Sports-1M	Introduced CSN to reduce computational complexity without sacrificing accuracy.
Ghadiyaram et al. (2019)	3D CNN	IG-65M, Kinetics	Leveraged large-scale weakly-supervised pre-training on video data, and significantly boosted performance.
Kopuklu et al. (2019)	3D CNN	Kinetics-400, UCF-101	Proposed resource-efficient 3D CNNs using depthwise separable convolutions, and achieved competitive accuracy with reduced computational requirements.
Feichtenhofer (2020)	3D CNN	Kinetics-400, Charades	Proposed X3D, a family of efficient video models by expanding architectures along multiple axes.
Li et al. (2020b)	3D CNN	Kinetics-400, UCF-101	Introduced a temporal attention mechanism to enhance efficiency and accuracy in 3D CNNs.

4.2.1.3 CNN-RNN-Based Approaches

The integration of CNNs and RNNs for HAR was significantly advanced by the work of Donahue et al. (2015), who introduced long-term recurrent convolutional networks

(LRCN). This approach effectively combined the spatial feature extraction capabilities of CNNs with the temporal dynamics modeling of LSTMs, demonstrating substantial improvements in action recognition tasks on datasets like UCF-101 and HMDB-51. Building on this foundation, Yue-Hei Ng et al. (2015) extended the application of deep networks to video classification by integrating deep CNNs with LSTMs to handle longer video sequences. Their method, tested on the Sports-1M and UCF-101 datasets, highlighted the importance of capturing extended temporal dependencies for improved performance in complex video classification tasks. Further pushing the boundaries, Srivastava et al. (2015) explored unsupervised learning of video representations using LSTMs. By leveraging LSTMs to learn spatio-temporal features without labeled data, their approach demonstrated effective video representation learning on the UCF-101 dataset, showcasing the versatility and potential of CNN-RNN architectures in both supervised and unsupervised learning scenarios for HAR.

The development of CNN-RNN architectures for HAR saw significant advancements in 2016. Wu et al. (2015b) proposed a hybrid deep learning framework that modeled spatial-temporal clues by combining CNNs for spatial feature extraction with RNNs for temporal sequence modeling. Their approach, tested on the UCF-101 and HMDB-51 datasets, demonstrated substantial improvements in video classification accuracy. Additionally, Li et al. (2016) expanded the application of CNN-RNN architectures to real-time scenarios with their approach for online human action detection using joint classification-regression RNNs. Combining CNNs for spatial features and RNNs for temporal dynamics, their method, tested on the J-HMDB and UCF-101 datasets, achieved notable improvements in accuracy and efficiency, showcasing the practicality of CNN-RNN models in real-time action detection.

Building on these advancements, 2017 and 2018 witnessed further refinements and innovations in CNN-RNN architectures for HAR. Li et al. (2018b) introduced VideoLSTM, integrating convolutions, attention mechanisms and optical flow within a recurrent framework, and demonstrating improved performance on the UCF101 and HMDB51 datasets. Carreira and Zisserman (2017) made a significant contribution with the two-stream Inflated 3D ConvNet (I3D), which inflated 2D CNN architectures into 3D and combined them with RNNs for temporal modeling. The model was evaluated on the Kinetics dataset, as well as UCF101 and HMDB51. Ullah et al. (2017) proposed a novel architecture combining CNNs with bi-directional LSTMs, effectively utilizing both spatial and temporal information from video sequences and showing superior performance on the UCF-101 and HMDB-51 datasets. In 2020, in the realm of human activity recognition using sensor data, Xia et al. (2020) proposed an LSTM-CNN architecture that effectively captured both temporal dependencies and local feature patterns, showing improved accuracy on the WISDM, UCI HAR, and OPPORTUNITY datasets. Similarly, Mutegeki and Han (2020) developed a CNN-LSTM approach for smartphone sensor-based activity recognition, demonstrating high accuracy on the UCI HAR dataset and further validating the effectiveness of combining CNNs and RNNs for processing time-series data in activity recognition tasks.

Recent advancements in HAR have leveraged sophisticated CNN-RNN architectures to enhance performance and reduce computational complexity. Muhammad et al. (2021) introduced an attention-based LSTM network combined with dilated CNN features, and significantly improved the recognition accuracy on the UCF-101 and HMDB-51 datasets by capturing essential spatial features through dilated convolutions and temporal patterns with attention mechanisms. Building on this, Malik et al. (2023) focused on multiview HAR; utilizing a CNN-LSTM architecture to cascade pose features, they achieved high accuracy (94.4% on the MCAD dataset and 91.67% on the IXMAS dataset) while reducing the computational load by targeting pose data rather than entire images.

Table 4.3 presents the works discussed in this section that utilized CNN-RNN approaches.

TABLE 4.3: CNN-RNN-based approaches in HAR.

Paper	Model	Dataset	Novelty
Donahue et al. (2015)	CNN-RNN, (LRCN)	UCF-101, HMDB-51	Combined CNNs for spatial feature extraction with LSTMs for temporal dynamics.
Yue-Hei Ng et al. (2015)	CNN-RNN	Sports-1M, UCF-101	Integrated deep CNNs with LSTMs to handle longer video sequences, capturing extended temporal dependencies.
Srivastava et al. (2015)	CNN-RNN, (Unsupervised LSTM)	UCF-101	Explored unsupervised learning of video representations using LSTMs, leveraging spatiotemporal features.
Wu et al. (2015b)	CNN-RNN	UCF-101, HMDB-51	Modeled spatial-temporal clues by combining CNNs for spatial features with RNNs for temporal sequence modeling.
Li et al. (2016)	CNN-RNN	J-HMDB, UCF-101	Applied CNN-RNN architectures to real-time scenarios for online human action detection.

Continued on next page

Table 4.3 – continued from previous page

Paper	Model	Dataset	Novelty
Li et al. (2018b)	CNN-RNN (VideoLSTM)	UCF-101, HMDB-51	Integrated convolutions, attention mechanisms, and optical flow within a recurrent framework.
Carreira and Zisserman (2017)	3D CNN-RNN	Kinetics, UCF101, HMDB51	Inflated 2D CNN architectures into 3D, and combined them with RNNs for temporal modeling.
Ullah et al. (2017)	CNN-RNN, (CNN-BiLSTM)	UCF101, HMDB51	Combined CNNs with bi-directional LSTMs to utilize both spatial and temporal information.
Xia et al. (2020)	CNN-RNN	WISDM, UCI, OPPORTUNITY	Captured both temporal dependencies and local feature patterns for human activity recognition using sensor data.
Mutegeki and Han (2020)	CNN-RNN	UCI	Developed a CNN-LSTM approach for smartphone sensor-based activity recognition, and demonstrated high accuracy.
Muhammad et al. (2021)	CNN-RNN, (CNN-Attention-LSTM)	UCF-101, HMDB-51	Improved recognition accuracy with attention-based LSTM network combined with dilated CNN features.
Malik et al. (2023)	CNN-RNN	MCAD, IXMAS	Achieved high accuracy in multiview HAR by cascading pose features using a CNN-LSTM architecture.

4.2.2 ViT-Based Approaches in HAR

In 2020, the ViT was conceptualized and introduced in the academic domain through the paper authored by Dosovitskiy et al. (2020). The ViT marked a paradigm shift in still image recognition methodologies, applying the Transformer model, predominantly known for its success in NLP, to the realm of computer vision. The application of ViTs in action recognition, a more specific and complex task within the field of computer vision, followed the initial introduction of ViT. Specifically, in 2021 and beyond, subsequent research and publications have explored and expanded the use of ViTs for action recognition tasks, demonstrating their efficacy in capturing spatial-temporal features within video data. They employ attention mechanisms to minimize redundant information and to model interactions over long distances in both space and time Koot et al. (2021). The adaptation of ViT to action recognition signifies the model's versatility and its potential for broader applications in computer vision beyond static image analysis.

Recent advancements in action recognition have seen a significant shift towards ViT, highlighting their efficacy in video understanding tasks. Arnab et al. (2021) introduced ViViT, extending the vision Transformer architecture to handle video sequences. They demonstrated its potential on datasets like Kinetics-400 and Something-Something-V2, marking a substantial improvement in video action recognition capabilities. Building on this, Bertasius et al. (2021) proposed a space-time Transformer that models temporal information innovatively, and achieved competitive results on similar datasets. The efficiency of multiscale ViTs was further illustrated by Fan et al. (2021), who showed that such architectures could effectively capture fine-grained video details and enhance classification performance on comprehensive video datasets. Moreover, Liu et al. (2022) presented the Swin Transformer, utilizing a shifted window mechanism to model long-range dependencies more efficiently, and leading to significant improvements in action recognition accuracy. Together, these works underscore the transformative impact of ViTs in advancing the field of HAR. Additionally, Wang et al. (2021b) introduced ActionCLIP, leveraging the CLIP model for enhanced video action recognition on multiple standard video datasets, including Kinetics-400 and HMDB-51. This novel approach integrated visual and linguistic representations.

Chen and Ho (2022) introduced Mm-ViT, a multi-modal video Transformer designed for compressed video action recognition, and demonstrated high performance by leveraging multi-modal inputs on compressed video datasets such as HACS and UCF101. Sharir et al. (2021) explored the extension of ViT to video data, showing its potential in capturing temporal dynamics effectively across several standard video datasets including Kinetics-400 and HMDB-51. Furthermore, Xing et al. (2023) developed SVFormer, a semi-supervised video Transformer that leverages both labeled and unlabeled data to bridge the gap between supervised and unsupervised learning, and achieved significant improvements in action recognition tasks on various standard HAR datasets such as Kinetics-400 and UCF101. Together, these works underscore the transformative impact of ViTs in advancing the field of HAR.

Table 4.4 presents the works discussed in this section that utilized ViTs.

TABLE 4.4: ViT-based approaches in HARs.

Paper	Model	Dataset	Novelty
Arnab et al. (2021)	ViViT	Kinetics-400, Something-Something-V2	Extended ViT to video sequences.
Bertasius et al. (2021)	Space-Time Transformer	Kinetics-400	Innovative temporal information modeling.
Fan et al. (2021)	Multiscale ViT	Kinetics-400, Something-Something-V2	Efficient capture of fine-grained video details.
Liu et al. (2022)	Swin Transformer	Kinetics-400, Something-Something-V2	Shifted window mechanism for long-range dependency modeling.
Wang et al. (2021b)	ActionCLIP	Kinetics-400, HMDB-51	Leveraged CLIP for enhanced video action recognition.
Chen and Ho (2022)	Mm-ViT	HACS, UCF101	Multi-modal inputs for compressed video action recognition.
Sharir et al. (2021)	ViT	Kinetics-400, HMDB-51	Applied ViT to video data.
Xing et al. (2023)	SVFormer	Kinetics-400, UCF101	Semi-supervised learning for action recognition.

4.2.3 CNN-ViT Hybrid Architectures

The integration of ViTs with CNNs has significantly advanced HAR tasks. Zhang et al. (2021c) proposed a two-stream hybrid CNN-Transformer network (THCT-Net), which demonstrated enhanced generalization ability and convergence speed on the NTU RGB+D dataset by combining CNNs for low-level context sensitivity and Transformers for capturing global information. Following this, Jegham et al. (2022) applied a similar hybrid model to driver action recognition, leveraging multi-view data to achieve high accuracy through the integration of CNNs for spatial feature extraction and Transformers for temporal dependencies. Kalfaoglu et al. (2022) extended this approach by integrating 3D CNNs with Transformers for late temporal modeling, and achieved substantial improvements in action recognition accuracy on the HMDB-51 and UCF101 datasets. Moreover, Yu et al. (2023) proposed Swin-Fusion, which combines Swin Transformers with CNN-based feature fusion to achieve state-of-the-art performance on datasets like Kinetics-400 and Something-Something-V2, demonstrating the robustness and superior performance of hybrid models in HAR tasks.

Djenouri and Belbachir (2022) proposed a hybrid visual Transformer model that integrates CNNs and Transformers for efficient and accurate human activity recognition. They demonstrated its capability on datasets like Kinetics-400 and UCF101, and showed that the hybrid approach leverages the local feature extraction of CNNs with the global context modeling of Transformers. Following this, Surek et al. (2023) provided a comprehensive review of deep learning approaches for video-based human activity recognition, emphasizing the potential of hybrid models. This review underscored the effectiveness of such hybrid models in capturing both spatial and temporal features from video data, and evaluated on various human activity datasets including NTU RGB+D and UTD-MHAD. Ahmadabadi et al. (2023) explored the use of knowledge distillation techniques to enhance the performance of hybrid CNN-Transformer models. Their approach was validated on datasets such as HMDB-51 and Kinetics-400, showing significant improvements in HAR by effectively transferring knowledge from complex models to more efficient ones. Together, these works highlight the evolving landscape of hybrid models in human activity recognition, showcasing their robustness and efficiency in handling complex video data.

Table 4.2 presents the works discussed in this section that utilized CNN-ViT approaches.

TABLE 4.5: CNN-ViT hybrid approaches in HARs.

Paper	Model	Datase	Novelty
Zhang et al. (2021c)	The two-stream hybrid CNN-Transformer network (THCT-Net)	NTU RGB+D	Combined CNNs and Transformers for improved generalization and convergence speed.
Jegham et al. (2022)	Multi-view vision Transformer	Custom driver action datasets	Leveraged multi-view data for spatial and temporal feature integration.
Kalfaoglu et al. (2022)	3D CNN-Transformer	HMDB-51, UCF101	Integrated 3D CNNs with Transformers for late temporal modeling.
Yu et al. (2023)	Swin-Fusion	Kinetics-400, Something-Something-V2	Combined Swin Transformers with CNN-based feature fusion for state-of-the-art performance
Djenouri and Belbachir (2022)	Hybrid visual Transformer	Kinetics-400, UCF101	Efficient and accurate human activity recognition leveraging strengths of CNNs and Transformers
Surek et al. (2023)	Various deep learning models including hybrid models	NTU RGB+D, UTD-MHAD	Comprehensive review highlighting the potential of hybrid models.
Ahmadabadi et al. (2023)	Hybrid CNN-Transformer	HMDB-51, Kinetics-400	Knowledge distillation from CNN-Transformer models for enhanced performance.

4.2.4 Discussion

In the field of HAR, the choice of models – whether CNN-based, ViT-based, or a hybrid of CNN and ViT – significantly influences the outcome and efficiency of the task. CNN-based models are particularly adept at extracting local features due to their convolutional nature LeCun et al. (2015), making them highly effective in pattern recognition within images and videos. Their computational efficiency is a boon for real-time applications Howard et al. (2017), and their robustness to input variations is notable Simonyan and Zisserman (2014b). However, CNNs often struggle with global contextual understanding Szegedy et al. (2015) and are prone to overfitting. Moreover, their ability to model long-range temporal dependencies Karpathy et al. (2014), which is crucial in action recognition, is somewhat limited.

ViT-based models, in contrast, excel in capturing global dependencies Carion et al. (2020); Dosovitskiy et al. (2020), thanks to their self-attention mechanism. This attribute makes them particularly suited for understanding complex actions that require a broader view beyond local features. ViTs are scalable with data, benefiting significantly from larger datasets, and are flexible in processing inputs of various sizes Touvron et al. (2021). The adaptability in processing various input sizes is a byproduct of the patch-based approach and the global receptive field of the ViTs. However, these models are computationally more intensive and require substantial training data to achieve optimal performance Khan et al. (2022). Unlike CNNs, ViTs are not as efficient in extracting detailed local features, which can be a critical drawback in certain action recognition scenarios.

Hybrid models that combine CNNs and ViTs aim to harness the strengths of both architectures. They offer the local feature extraction capabilities of CNNs along with the global context awareness of ViTs, potentially providing a more balanced approach to action recognition. These models can be more efficient and versatile, adapting well to a range of tasks. However, this combination brings its own challenges, including increased architectural complexity, higher resource demands, and the need for careful tuning to balance the contributions of both CNN and ViT components. The choice among these models depends on the specific requirements of the action recognition task, such as the available computational resources, the nature and size of the dataset, and the types of actions that need to be recognized.

For a summary of the advantages and disadvantages of these three architectural variations, see Table 4.6.

TABLE 4.6: Capability comparison between Transformer-based, CNN-based, and hybrid models in HARs.

Criteria	ViT-based	CNN-based	Hybrid Models
Advantages			
Excel at capturing global dependencies	✓		✓
Scalable with data	✓		✓
Flexible in processing various input sizes	✓		✓
Adept at extracting local features		✓	✓
Computationally efficient		✓	
Robust to input variations		✓	✓
Efficient and versatile			✓
Adapts well to a range of tasks			✓
Disadvantages			
Computationally intensive	✓		✓
Requires substantial training data	✓		✓
Limited global contextual understanding		✓	
Prone to overfitting		✓	
Limited in modeling long-range dependencies		✓	
Architectural complexity			✓
Higher resource demands			✓
Need for careful tuning			✓
Balancing contributions of both components can be challenging			✓

4.3 Proposed CNN-ViT Hybrid Architecture

In this section, we present our proposed CNN-ViT architecture for HAR, leveraging the benefits of both approaches described in previous sections, see Figure 4.1. The architecture incorporates a TimeDistributed layer with a CNN backbone, followed by a ViT model to classify actions in video sequences.

Spatial component. Let $\mathcal{X}$ be a collection of N frames, i.e., $\mathcal{X} = \{\boldsymbol{X}_i\}_{i=1}^{N}$. The CNN backbone (i.e. MobileNet in Howard et al. 2017) in the TimeDistributed layer (see Figure 4.1) processes the individual frames $\boldsymbol{X}_i$ and outputs the spatial features vector

$\boldsymbol{v}_i = p_\theta(\boldsymbol{X}_i) \in \mathbb{R}^L$, where p_θ is the CNN model (e.g. MobileNet or VGG16) with parameters in θ wrapped by the TimeDistributed layer.

Temporal component. In the proposed hybrid CNN-ViT model, it is engineered to process the sequence of the N spatial features vectors, i.e., $\{\boldsymbol{v}_i\}_{i=1}^N$, where each $\boldsymbol{v}_i$ represents a distinct frame of the input video clip, see Figure 4.1. Afterwards, the ViT block outputs a final representation, $\boldsymbol{z}$, which is then fed into the softmax layer to classify the action in the video. In detail, the Transformer encoder is designed to process a sequence of vectors, each representing one frame, and aggregate information into a single vector for classification.

In the proposed ViT-only model in Figure 4.2 for the purpose of comparison, each vector represents a distinct patch. These vectors are first linearly projected into a high-dimensional space, facilitating the model's ability to learn complex patterns within the data. To ensure the model captures the sequential nature of the input, positional encodings are added to these embeddings. The core of the ViT consists of two layers, each comprising a multi-head self-attention mechanism and a feed-forward network. The self-attention mechanism allows the model to weigh the importance of different patches relative to each other, while the feed-forward network, utilizing an exponential linear unit (ELU) activation function, processes each position independently to capture global context. The ViT is designed to aggregate the information from all vectors and positional encodings into a single [CLS] token, which is prepended to the input sequence. The output vector associated with this [CLS] token, after propagation through the Transformer layers, serves as a comprehensive representation of the entire input, suitable for downstream classification tasks.

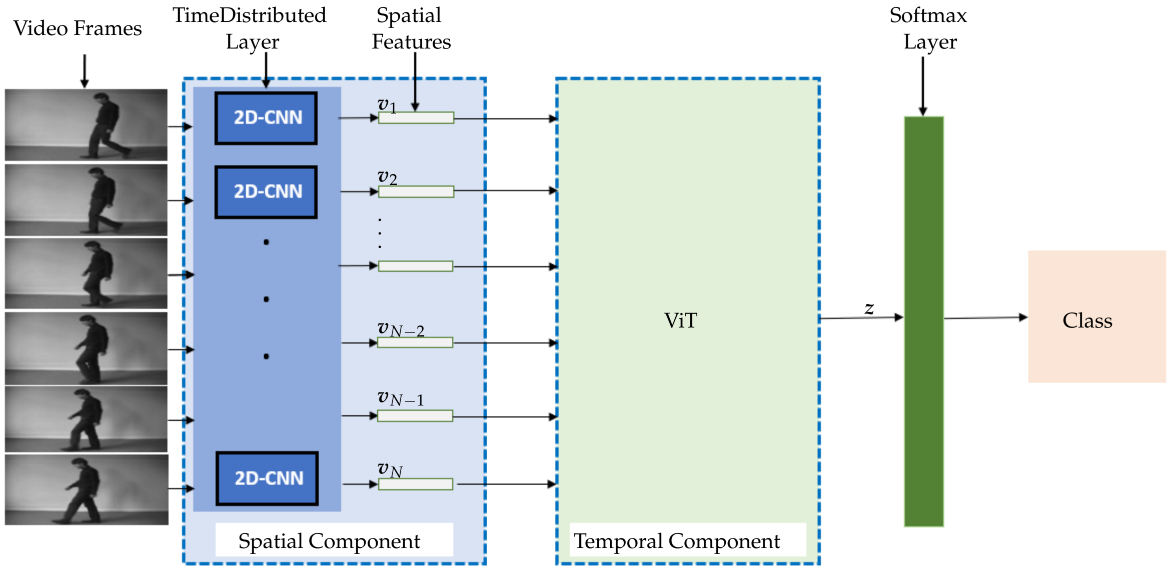


FIGURE 4.1: The hybrid CNN-ViT architecture for HARs.

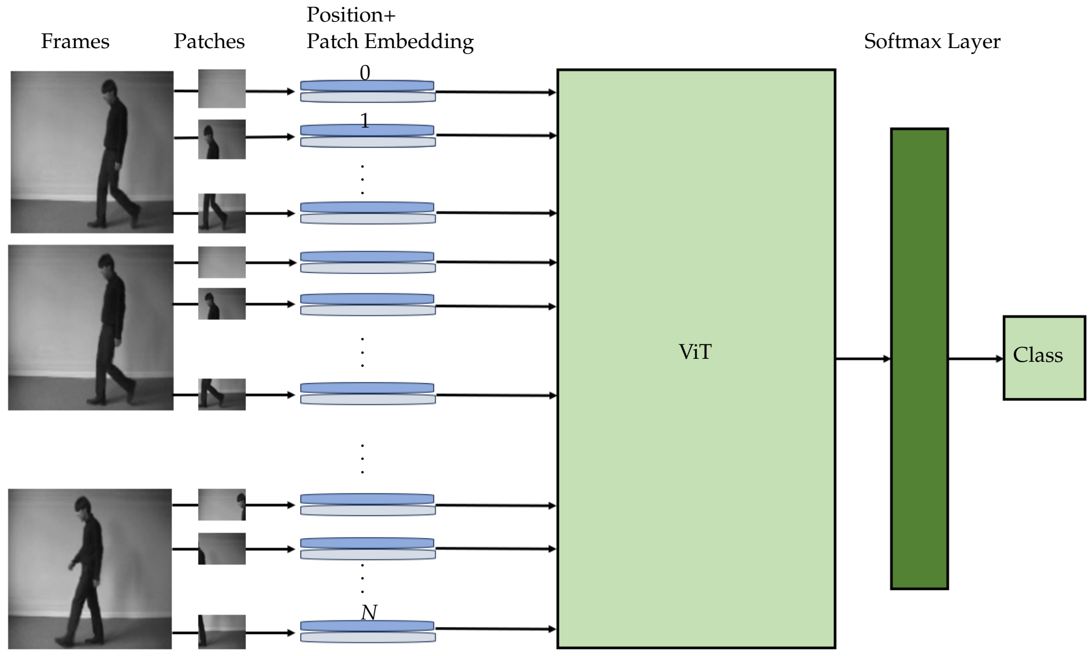


FIGURE 4.2: The ViT-only architecture for HARs.

4.3.1 Experiments

The goal of the presented experiments is not necessarily to produce a model that outperforms the state-of-the-art models in the HAR field. Rather, the aim is to conduct a comparison among the CNN, ViT-only, and hybrid models to give further insights.

The Royal Institute of Technology in 2004 unveiled the KTH dataset, a significant and publicly accessible dataset for action recognition Schuldt et al. (2004). The KTH dataset was chosen here for its balanced representation of spatial and temporal features. Renowned as a benchmark dataset, it encompasses six types of actions: walking, jogging, running, boxing, hand-waving, and hand-clapping. The dataset features performances by 25 different individuals, introducing a diversity in execution. Additionally, the environment for each participant's actions was deliberately altered, including settings such as outdoors, outdoors with scale changes, outdoors with clothing variations, and indoors. The KTH dataset comprises 2,391 video sequences, all recorded at 25 frames per second using a stationary camera against uniform backgrounds.

Nine experiments were conducted, with each of the aforementioned models trained on three different lengths of frame sequences. Care was taken to avoid pre-training in order to ensure the neutrality of the results. The TransNet model in 3 was adopted to represent the CNN model, the ViT model, and the Hybrid model were depicted in Figure 4.2 and Figure 4.1, respectively. For the spatial component of the hybrid model, we employed the spatial component of TransNet; and for the temporal component, we employed the same ViT model that we used in the ViT-only model. We constructed our model utilizing Python 3.6, incorporating the Keras deep learning framework, OpenCV for image processing, matplotlib, and the scikit-learn library. The training and test were performed on a computer equipped with an Intel Core i7 processor, an NVidia RTX 2070 graphics card, and 64GB of RAM.

4.3.1.1 Results and Discussion

TABLE 4.7: Experimental results of different models on the KTH Dataset using three different context lengths. In particular, the hybrid model was trained without pre-training, whereas Hybrid$_{\text{pre}}$ is for the hybrid model pre-trained on ImageNet. Every experiment was repeated over five runs to ensure robust statistical evaluation.

Context length	**CNN-based**	**ViT-only**	**Hybrid**	**Hybrid$_{\text{pre}}$**
12 frames	94.35 ± 0.41	92.44 ± 0.16	94.12 ± 0.05	96.34 ± 0.03
18 frames	93.91 ± 0.32	92.82 ± 0.07	94.56 ± 0.10	97.13 ± 0.04
24 frames	93.49 ± 0.24	93.69 ± 0.08	95.78 ± 0.60	97.89 ± 0.05

Table 4.7 presents the quantitative results of the three distinct models, i.e., CNN, ViT-only, and a hybrid model on the KTH dataset, focusing on three different context lengths, i.e., short (12 frames), medium (18 frames), and long (24 frames). The results from these experiments provide insightful revelations into the efficacy of each model under different temporal contexts. More details are given below.

The CNN model exhibited a decrease in accuracy as the frame length increased, recording 94.35% for 12 frames, 93.91% for 18 frames, and 93.49% for 24 frames. This descending trend suggests that CNN may struggle with processing longer sequences where temporal dynamics become more complex, potentially leading to challenges such as overfitting or difficulties in temporal feature retention over extended durations.

In contrast, the ViT model demonstrated an improvement in performance with longer sequences, achieving accuracy of 92.44% for 12 frames, 92.82% for 18 frames, and 93.69% for 24 frames. This ascending pattern supports the notion that ViT architectures, with their inherent self-attention mechanisms, are well-suited to managing longer sequences. The ability of ViTs to assign varying degrees of importance to different parts of the sequence likely contributes to their enhanced performance on longer input frames.

The hybrid CNN-ViT model showcased the highest and continuously improving accuracy rates across all frame lengths: 94.12% for 12 frames, 94.56% for 18 frames, and an impressive 95.78% for 24 frames. Moreover, the pre-trained hybrid model showcased the same trend, with the best accuracy achieved. This type of model synergistically combines CNN's robust spatial feature extraction capabilities with

ViT's efficient handling of temporal relationships via self-attention. The results from this model indicate that such a hybrid approach is particularly effective in capturing the complexities of action recognition tasks in video sequences, especially as the sequence length increases.

These findings underscore the potential advantages of hybrid neural network architectures in video-based action recognition tasks, particularly for handling longer sequences with complex interactions. The superior performance of the hybrid CNN-ViT model suggests that integrating the spatial acuity of CNNs with the temporal finesse of ViTs can lead to more accurate and reliable recognition systems. Future work could explore the scalability of these models to other datasets, their computational efficiency, and their robustness against variations in video quality and scene dynamics. Additionally, further research might investigate the optimal balance of CNN and ViT components within hybrid models to maximize both performance and efficiency.

TABLE 4.8: Comparison of the proposed hybrid model with the state-of-the-art models on the KTH dataset.

Methods	**Venue**	**Accuracy**
Geng and Song (2016)	ICCSAE '16	92.49
Arunnehru et al. (2018)	RoSMa '18	94.90
Abdelbaky and Aly (2020)	ITCE '20	87.52
Jaouedi et al. (2020)	KSUCI journal '20	96.30
Liu et al. (2020)	JAIHC '20	91.93
Sahoo et al. (2020)	TETCI '20	97.67
Lee et al. (2021)	CVF '21	89.40
Basha et al. (2022)	MTA journal '22	96.53
Ye and Bilodeau (2023)	CVF '23	90.90
Ours	-	**97.89**

To complete the comparison, Table 4.8 shows that the impressive 97.89% accuracy achieved by the presented CNN-ViT hybrid model on the KTH dataset places it prominently among state-of-the-art models for HAR. This performance is notably superior when compared to earlier benchmarks reported in the literature such as Geng and Song (2016) with 92.49% and Arunnehru et al. (2018) with 94.90%. Our model utilizes an ImageNet-pre-trained MobileNet Howard et al. (2017) as the CNN backbone in the spatial component, which enhances its robust feature extraction capabilities. Combined with the dynamic attention mechanisms of ViT, it can thereby enhance both the spatial and temporal processing of video sequences. Furthermore, our hybrid model not only surpasses other contemporary approaches like Liu et al. (2020) (91.93%) and Lee et al. (2021) (89.40%), but also shows competitive/superior performance against some of the highest accuracy in the field, such as Jaouedi et al. (2020) (96.30%) and Basha et al. (2022) (96.53%). Even in comparison to the high benchmark set by Sahoo et al. (2020) (97.67%), our hybrid model demonstrates a marginal but significant improvement, underscoring the efficacy of integrating CNN with ViT. This integration not only facilitates more nuanced feature extraction across both spatial and sequential dimensions but also adapts more dynamically to the varied contexts inherent in video data, making it a potent solution for realistic action recognition scenarios.

On the whole, the integration of CNN with ViT is particularly advantageous for enhancing feature extraction capabilities and focusing on relevant segments dynamically through the attention mechanisms of ViTs. This not only helps in improving accuracy but also in making the model more adaptable to varied video contexts, a key requirement for action recognition in realistic scenarios. This comparative advantage suggests that hybrid models are paving the way for future explorations in HAR, combining the best of convolutional and ViT-based architectures for improved performance and efficiency.

4.3.1.2 Statistical Significance Analysis

In this section, we present the statistical significance analysis used to evaluate the performance of the proposed model in comparison with benchmark models. The analysis here employs two statistical methods: the *paired-samples t-test* (see Algorithm 2 in Appendix) and the *one-sample t-test* (see Algorithm 3 in Appendix) Montgomery and Runger (2020); Devore (2000). The symbols and variables used in Algorithms 2 and 3 are summarized in Table 6.1 in Appendix.

TABLE 4.9: The t-statistic values for the models in Table 4.7 across the three contexts.

Context	**CNN-based**	**Vit-only**	**Hybrid**	**Hybrid$_{\text{pre}}$**
12 vs. 18 frames	6.52	19.75	51.12	70.19
12 vs. 24 frames	14.22	37.67	134.33	141.14
18 vs. 24 frames	6.59	58.14	301.23	73.14

TABLE 4.10: The two-tailed 5% p-value for the models in Table 4.7 across the three contexts.

Context	**CNN-based**	**ViT-only**	**Hybrid**	**Hybrid$_{\text{pre}}$**
12 vs. 18 frames	2.95×10^{-3}	3.94×10^{-5}	8.85×10^{-7}	2.47×10^{-7}
12 vs. 24 frames	1.45×10^{-4}	2.97×10^{-6}	1.84×10^{-8}	1.51×10^{-8}
18 vs. 24 frames	2.76×10^{-3}	5.24×10^{-7}	7.31×10^{-10}	2.09×10^{-7}

The paired-samples t-test algorithm evaluates different individual models in Table 4.7 whether there is a significant difference of the performance of a model between two related contexts among the total three contexts (i.e., 12, 18, and 24 frames). Applying Algorithm 2 on the quantitative results in Table 4.7, we obtain the t-statistic values t_{sp} given in Table 4.9 and the p-values p_p given in Table 4.10 for each model on paired contexts. For the paired-samples t-test, the null hypothesis (H_{p0}) posited that no significant difference exists between two contexts for each model. The alternative hypothesis (H_{p1}) suggested that a significant difference existed between two contexts for each model. The results in Table 4.10 demonstrates statistically significant differences across all contexts for each model, with p-values lower than the adjusted significance level α_a using the Bonferroni correction.

The one-sample t-test algorithm is used here to evaluate whether there is a significant difference between the performance of the proposed model and the mean performance of the benchmark models in Table 4.8. In this test, the null hypothesis (H_{o0}) assumes that the performance of the proposed model is not significantly different from the mean performance of the benchmark models in Table 4.8. The alternative hypothesis (H_{o1}) posits that a significant difference does exist. By applying Algorithm 3 on the data in Table 4.8, we obtain a p-value of $p_o = 0.0034$, which is significantly lower than the commonly accepted significance level of 0.05. As a result, we reject the null hypothesis. This finding indicates that the performance difference between the proposed model and the state-of-the-art models is statistically significant. Consequently, we can conclude with 95% confidence that the proposed model outperforms the current state-of-the-art models for the HAR task under consideration. This result highlights the effectiveness of the proposed model in advancing the field.

4.4 Conclusions

This survey provides a comprehensive overview of the current state of HAR by examining the roles and advancements of CNNs, RNNs, and ViTs. It delves into the evolution of these architectures, emphasizing their individual contributions to the field. The introduction of a hybrid model that combines the spatial processing capabilities of CNNs with the temporal understanding of ViTs represents a methodological advancement in HAR. This model aims to address the limitations of each architecture when used in isolation, proposing a unified approach that potentially enhances the accuracy and efficiency of action recognition tasks. The chapter identifies key challenges and opportunities within HAR, such as the need for models that can effectively integrate spatial and temporal information from video data. The exploration of hybrid models, as suggested, offers a pathway for future research, particularly in improving model performance on complex video datasets. The discussion encourages further investigation into optimizing these hybrid architectures and exploring their applicability across various domains. This work sets a foundation for future studies to build upon, aiming to push the boundaries of what is currently achievable in HAR and to explore new applications of these technologies in real-world scenarios.

Chapter 5

Conclusions

The conclusion of this thesis summarizes the key findings and contributions made toward advancing HAR through innovative model architectures and methodologies. This work has focused on addressing critical challenges in the field, such as improving feature extraction, enhancing transfer learning capabilities, and developing hybrid models that better capture both spatial and temporal aspects of human actions. By integrating state-of-the-art techniques such as CNNs, ViTs, and data augmentation strategies, this research has presented new solutions that offer significant improvements over traditional methods in terms of accuracy, efficiency, and generalization across diverse datasets. Section 5.1 summarises each of the three Chapters. Section 5.2 identifies the limitations of the current thesis while Section 5.3 offers suggestions for future studies. The contributions of the current thesis are presented in Section 5.4. Finally, Section 5.5 provides a concluding remark.

5.1 Summary of The Thesis

The current thesis investigates and addresses key challenges in HAR by introducing new model architectures and strategies for enhancing feature extraction, transfer learning, and data augmentation. Through a series of three interconnected studies, this research aims to improve the ability of HAR systems to generalize across complex and dynamic environments. The thesis explores the decomposition of the 3D-CNN and the integration of CNNs with ViTs to form hybrid models capable of capturing both spatial and temporal features. This integration helps overcome the limitations of traditional models that either excel in spatial or temporal analysis but struggle to balance both. The following sections provide a concise summary of each of the Chapters presented in this thesis.

5.1.1 Summary of The Second Chapter

The second Chapter 2 provides a comprehensive overview of data augmentation techniques employed in computer vision tasks, particularly in image classification and segmentation. It addresses the challenges posed by the large data requirements of deep neural networks, such as CNNs, and the issue of overfitting when data is scarce. The chapter emphasizes the importance of data augmentation, which artificially expands datasets to improve model generalization. By increasing both the size and diversity of training data, data augmentation can effectively mitigate overfitting.

The survey highlights two main categories of data augmentation: traditional methods and deep learning-based methods. Traditional techniques include geometric transformations such as flipping, rotation, cropping, and photometric transformations like brightness and contrast adjustments. These techniques are simple yet effective in improving data diversity and size. However, they come with limitations, such as boundary issues in rotation, where black patches or missing pixel information may affect the model's performance. To address such limitations, the chapter proposes a novel geometric augmentation strategy named RLR. The chapter presents experimental results that show RLR consistently outperforming traditional augmentation methods, particularly in classification tasks, where maintaining local image integrity is essential for model generalization.

The chapter extensively reviews the application of data augmentation in classification and segmentation tasks. It explores the efficacy of various augmentation methods across different classification and segmentation datasets. The review includes an analysis of traditional augmentation methods and more advanced techniques, such as GANs and texture transfer. GANs, for example, are highlighted for their ability to generate synthetic images that augment datasets, making models more robust to data scarcity.

The chapter showed that computer vision researchers frequently combine several data augmentation techniques, making it difficult to evaluate the impact of each method individually. In this chapter, the random rotation technique was examined in detail, focusing on its influence on two distinct tasks: classification and segmentation. The findings indicate that classification and segmentation rely on different features, which may not benefit equally from the same augmentation methods. For classification, random rotation proved beneficial, as it altered the object's shape while preserving texture, leading to enhanced model performance. This suggests that classification tasks are less sensitive to shape distortions caused by rotation and can benefit from the added variation introduced by the technique.

In contrast, segmentation tasks, which depend heavily on accurate shape feature extraction, showed poorer results when random rotation techniques were applied.

The RLR method, while effective in preserving some local details, distorted the global shape of objects, particularly in the segmentation of human bodies. This distortion negatively impacted the segmentation task, where shape integrity is crucial. The study highlights that while rotation-based augmentation can enhance classification performance, it may degrade results in segmentation tasks where shape features are more important than texture. These findings emphasize the need for task-specific data augmentation strategies that consider the feature priorities of each task.

In conclusion, the chapter highlights the need to tailor data augmentation techniques to the specific requirements of the computer vision task at hand. For tasks like segmentation, where shape features are crucial, augmentation techniques that preserve or enhance shape features should be prioritized.

5.1.2 Summary of The Third Chapter

The third Chapter 3 focuses on addressing the limitations of current HAR models, which often have complex structures and require lengthy training times. To overcome these challenges, the chapter proposes a simplified, end-to-end deep learning architecture called TransNet. The architecture breaks down traditional 3D-CNNs into 2D-CNN and 1D-CNN components, where the 2D-CNN extracts spatial features from video frames, and the 1D-CNN captures the temporal patterns across frames. This decomposition makes the model more efficient while maintaining strong performance in HAR tasks. The model's compatibility with pre-trained 2D-CNNs, such as MobileNet and VGG16, further enhances its flexibility and transfer learning capabilities.

One of the key motivations behind TransNet is to leverage the power of transfer learning for HAR tasks. By utilizing well-trained models from other domains, such as image classification or segmentation, TransNet can efficiently learn from limited HAR datasets without overfitting. The integration of pre-trained 2D-CNNs allows for the reuse of spatial features already learned in other tasks, improving the model's efficiency and effectiveness in recognizing human actions. The chapter also introduces TransNet+, an extension that incorporates autoencoder-based encoders trained on tasks like HSS to improve spatial feature extraction. This strategy enhances HAR by pretraining the model on specific tasks that capture important features, such as human shapes, making it highly adaptable to action recognition.

The chapter reviews related work on HAR, highlighting the challenges posed by cluttered backgrounds and the need for precise temporal modeling. Traditional methods like 3D-CNNs and RNNs are known for their ability to capture spatio-temporal features but often require large datasets and substantial computational resources. TransNet, by contrast, simplifies the model architecture without sacrificing performance, making it suitable for both limited HAR datasets and limited-resources devices. Moreover, TransNet addresses the background clutter problem by focusing on human features, offering a more practical solution for environments with complex backgrounds.

Experimental results conducted on benchmark datasets, such as KTH, UCF101, and HMDB51, demonstrate TransNet's superior performance compared to state-of-the-art HAR models. The model, especially when pre-trained using HSS, shows notable improvements in action recognition accuracy, highlighting the effectiveness of combining 2D-CNN spatial feature extraction with 1D-CNN temporal analysis. For instance, on the KTH dataset, TransNet achieved 100% accuracy, outperforming several contemporary models, which underscores its robustness and adaptability. The chapter also compares the performance of various backbone models (MobileNet, VGG16, etc.), showing that TransNet performs consistently well across different architectures.

One of the significant contributions of TransNet is its ability to perform transfer learning across diverse datasets and tasks. By pretraining the model on tasks such as HSS, the chapter demonstrates that TransNet+ can effectively transfer knowledge (i.e. Human shape features) to HAR tasks, achieving higher accuracy and efficiency. This approach reduces the dependency on large training datasets, making it an ideal solution for domains where data is limited. Additionally, the chapter provides a detailed analysis of TransNet's model complexity, highlighting its reduced computational cost compared to 3D-CNNs, making it a potential choice for real-time HAR applications.

In conclusion, the chapter presents TransNet as a versatile and efficient solution for human action recognition, leveraging transfer learning and a simplified architecture to address the limitations of traditional HAR models. Through extensive experiments and comparisons with state-of-the-art methods, the chapter demonstrates that TransNet not only improves accuracy but also offers significant advantages in terms of model complexity and training speed. The introduction of TransNet+ further enhances its capabilities, making it a promising architecture for future HAR applications in both academic research and industry.

5.1.3 Summary of The Fourth Chapter

The fourth Chapter 4 provides a detailed exploration of deep learning models used in HAR, with a specific focus on CNNs, RNNs, and ViTs. It discusses the evolution of these models and their application in understanding human actions in video sequences, emphasizing the need for capturing both spatial and temporal information for accurate recognition. The chapter underscores the limitations of each model, such as CNNs' struggle with temporal dependencies and RNNs' challenges with long-range context, while also highlighting the transformative potential of ViTs, which have shown superior performance in recent years due to their self-attention mechanisms that excel in both spatial and temporal tasks.

The survey also proposes a hybrid model that integrates CNNs and ViTs, aiming to leverage the spatial feature extraction capabilities of CNNs and the global temporal context modeling strengths of Transformers. This hybrid model is designed to overcome the limitations of traditional models by combining their respective advantages, offering a more comprehensive approach to HAR tasks. The chapter outlines the growing trend toward hybrid models, particularly for applications that require precise action recognition in complex environments, such as medical diagnostics, security surveillance, and autonomous systems. Through a review of literature and analysis of benchmark datasets, the study demonstrates that hybrid architectures hold promise for advancing the state-of-the-art in HAR, particularly when dealing with large datasets and complex action sequences.

The chapter systematically reviews the evolution of CNNs from their early use in image classification to more complex spatio-temporal models like 3D-CNNs which can process video frames by extending CNNs to the temporal dimension. It discusses the advent of RNNs, particularly LSTMs and GRUs, which were initially introduced to handle sequential data but were later found to be limited by issues such as vanishing gradients and a lack of parallelization capabilities. Transformers, and more specifically Vision Transformers, are presented as a breakthrough in overcoming these limitations, utilizing self-attention mechanisms that process sequences more efficiently and capture long-range dependencies in video data.

The chapter emphasizes the significance of data representation in HAR, explaining how CNNs are effective in capturing local spatial features, while Transformers excel at understanding global patterns across frames. However, each approach has its trade-offs: CNNs are computationally efficient but struggle with temporal relationships, while Transformers, although highly effective in capturing global dependencies, require substantial computational resources and large datasets for training. This leads to the discussion of hybrid models, which attempt to balance these strengths and weaknesses, offering improved performance and generalization in action recognition tasks.

In the experimental section, the chapter proposes and tests a novel hybrid model that integrates CNNs for spatial feature extraction with Transformers for temporal dynamics, showing that this architecture outperforms standalone CNNs, RNNs, and ViTs in action recognition tasks on a benchmark dataset. The results demonstrate that hybrid models can achieve higher accuracy and efficiency by effectively capturing both local spatial details and global temporal patterns, particularly in scenarios involving complex human actions.

In conclusion, the chapter highlights the ongoing evolution of deep learning models in human action recognition, proposing that hybrid CNN-Transformer models represent the future of HAR research. By integrating the strengths of multiple architectures, these models offer a more robust and flexible approach to tackling the challenges of real-world video data, particularly in domains where both spatial and temporal precision are critical. The study suggests that future research should focus on refining these hybrid models to further improve their efficiency and scalability for large-scale applications.

5.2 Thesis Limitation

5.2.1 Datasets Issues

One of the significant limitations in HAR is the availability and quality of datasets, which poses challenges for training deep learning models Jegham et al. (2020). HAR models, particularly those using deep architectures like CNNs, RNNs, and ViTs, require large amounts of data to generalize well to new, unseen examples Arnab et al. (2021). However, most publicly available HAR datasets are limited in size and diversity. Popular datasets such as UCF101, HMDB51, and Kinetics-400, while commonly used, may not be sufficient for capturing the wide range of human actions in real-world scenarios. These datasets are often constrained by their limited number of labeled examples, specific environments, and controlled settings, which reduces the generalizability of the models trained on them. As a result, models trained on these datasets often struggle to perform well when applied to more complex or dynamic environments, leading to reduced accuracy and robustness.

Another critical limitation of HAR datasets is the imbalance of action classes. Many datasets have an unequal distribution of action categories, where certain common actions (e.g., walking, running) have significantly more examples than rarer or more complex actions (e.g., medical activities). This class imbalance can lead to biased models that perform well on frequent actions but fail to recognize or generalize to less common activities. Additionally, certain datasets lack diversity in terms of demographic representation, background variability, and action complexity, making it difficult for models to capture the full spectrum of human behaviors. For example, datasets collected in specific regions or environments might lack diversity in terms of cultural context, clothing styles, or lighting conditions, which further limits the model's ability to adapt to new or unfamiliar contexts.

Moreover, many HAR datasets are limited in terms of temporal annotation and labeling granularity. Some datasets provide only coarse labels for the entire video sequence, without detailed information about when specific actions occur within a video. This lack of fine-grained temporal annotation makes it difficult for models to accurately detect and classify actions in real-time or within continuous video streams. Action detection, as opposed to action classification, requires more precise labeling to train models that can identify the start and end points of actions. This temporal limitation hinders the performance of HAR models, especially in applications that require real-time action recognition, such as surveillance systems, autonomous vehicles, or assistive technologies for healthcare.

Finally, there is a notable limitation in the accessibility of large-scale, annotated datasets due to privacy and ethical concerns. Collecting and annotating video data for HAR tasks often involves recording individuals performing various actions, which raises significant privacy issues. In domains like healthcare or security, where HAR could have significant impact, gathering action data in real-world, sensitive settings is often constrained by ethical considerations. This limits the ability of researchers to access diverse datasets that represent actions in natural environments. Furthermore, annotating large-scale video datasets is a labor-intensive process, often requiring human annotators to manually label actions, which adds to the cost and complexity of dataset creation. Consequently, these limitations in dataset size, diversity, temporal annotation, and privacy impact the overall progress and accuracy of HAR models in real-world applications.

5.2.2 The Publicly Available State-of-the-art Transformers

One significant limitation in HAR is the lack of publicly available state-of-the-art transformer models that are pre-trained on large-scale datasets, comparable to the widespread availability of CNNs like those trained on ImageNet. Transformers, particularly ViTs, have shown immense promise in advancing HAR by capturing both spatial and temporal dependencies comparably or more effectively than traditional CNNs or RNNs. However, while CNNs benefit from models pre-trained on massive image datasets like ImageNet, which are readily accessible for transfer learning, the same infrastructure for transformers is still emerging. This gap hinders the ability of researchers and practitioners to leverage the full potential of transformers for HAR tasks, as models often have to be trained from scratch or fine-tuned on smaller, domain-specific datasets, which is computationally expensive and time-consuming.

The absence of large-scale pre-trained transformer models for HAR tasks also presents challenges in terms of data efficiency and model generalization. Pre-trained models, such as CNNs fine-tuned from ImageNet, allow for effective transfer learning, enabling models to generalize better even when the target dataset is relatively small.

Transformers, on the other hand, are known to require significantly larger amounts of data to achieve optimal performance due to their reliance on self-attention mechanisms that capture global context. Without large pre-trained ViTs or other transformer variants available, researchers often face difficulties training transformers on limited HAR datasets, leading to overfitting and suboptimal performance. This contrasts with CNNs, where pre-trained models significantly reduce the need for large training datasets by providing robust feature representations learned from a broader context.

Furthermore, while there has been rapid progress in transformer-based architectures, their training requirements are considerably higher than those of CNNs. Transformers generally require extensive computational resources, both in terms of hardware and training time, especially when trained from scratch on large video datasets. This creates a barrier for many researchers who lack access to such resources, limiting the widespread experimentation and adoption of transformers in HAR. Pre-trained transformers on large video datasets like Kinetics or Something-Something are still relatively scarce, in contrast to pre-trained CNNs that can be readily accessed through public repositories and frameworks such as PyTorch or TensorFlow. This scarcity of pre-trained transformer models not only slows down research progress but also impedes the deployment of advanced HAR systems in practical applications.

The lack of availability of these state-of-the-art transformer models also limits the community's ability to compare new architectures and methods effectively. In CNN-based HAR, researchers often compare their models against well-established baselines such as ResNet, MobileNet, or VGG, which have been rigorously tested and pre-trained on large datasets. However, with transformers, there is a shortage of publicly available, pre-trained models that can serve as benchmarks for performance evaluation. This makes it difficult for the research community to measure progress consistently or to determine the advantages of new transformer-based models over existing approaches. Thus, the absence of pre-trained state-of-the-art transformers not only hinders practical performance improvements but also stifles innovation by making it more difficult to establish reliable benchmarks and facilitate knowledge sharing across the HAR research community.

5.3 Thesis Future Work

In the future works of this thesis, three research directions will be pursued to further advance the field of HAR. The first work will involve a comprehensive survey on the use of Graph Neural Networks (GNNs) in HAR, exploring how GNNs can model complex human actions through spatial and temporal relationships. The second work will focus on investigating different attention techniques, evaluating the latest transformer models in HAR, and developing hybrid models that combine the strengths of GNNs and CNNs with ViTs for improved performance. Finally, the third work will aim to create an Arabic sign language video dataset, filling a crucial gap in the field, while also surveying existing sign language datasets. This dataset will be used to train and benchmark the models developed in the thesis and potential future works, contributing to more inclusive and effective action recognition systems. These studies will collectively enhance the understanding and development of HAR, particularly in underrepresented areas like sign language recognition and hybrid model architectures.

5.3.1 First Paper: Survey on Graph Neural Networks and Developing a GNN-CNN Hybrid Model for HAR

The first future work will involve conducting a comprehensive survey of GNNs in HAR and a development of a GNN-CNN hybrid model. GNNs have emerged as a powerful tool for capturing complex relationships in data represented as graphs, making them highly suitable for tasks like HAR, where interactions between joints in skeleton-based data or relationships between objects can be modeled as graphs Yan et al. (2018). The survey will explore the current landscape of GNNs in HAR, and their application to benchmark datasets. The survey will provide valuable insights into the advantages and limitations of GNNs, their ability to capture spatial and temporal dynamics, and how they compare to other models traditionally used in HAR, such as CNNs and RNNs.

Additionally, this work will involve the creation of a GNN-CNN hybrid model for HAR, leveraging the strengths of both architectures. While CNNs excel at extracting local spatial features from video frames, they are limited in modeling the relationships between different regions or time steps. GNNs, on the other hand, are excellent at capturing these relationships, making the combination of both approaches highly beneficial for HAR tasks. The GNN-CNN hybrid model will be designed to utilize CNNs for initial spatial feature extraction from video frames, followed by GNNs to model the relational structure between key points or body joints over time. This hybrid approach will provide a more comprehensive representation of human actions, allowing for better recognition of complex actions involving multiple interacting objects or people.

The importance of this survey lies in its ability to synthesize the existing research on GNNs in HAR, identifying trends, gaps, and future directions for researchers. Given that GNNs are relatively new in the field of HAR compared to CNNs and RNNs, this paper will highlight their potential and limitations. It will also provide insights into the datasets commonly used for GNN-based HAR, how GNNs handle data variability, and the techniques used to improve generalization. The implementation of this survey will involve categorizing various GNN architectures, comparing their performance on benchmark HAR datasets, and discussing their effectiveness in capturing both spatial and temporal aspects of human actions.

5.3.2 Second Paper: Exploring Attention Techniques and Developing a GNN-ViT Hybrid Model for HAR

The second paper will focus on exploring various attention mechanisms in HAR, assessing recent transformer models, and developing a GNN-ViT hybrid model. Attention mechanisms, such as self-attention and temporal attention, have been transformative in improving model performance across various deep learning applications by allowing models to focus on the most relevant parts of the input data Vaswani et al. (2017). In HAR, attention techniques can significantly enhance a model's ability to prioritize important spatial and temporal features, leading to more accurate recognition of actions in complex video sequences. This paper will explore different attention mechanisms and assess how their integration into HAR models improves the overall performance, particularly in capturing long-range dependencies between frames.

The study will culminate in the creation of a GNN-ViT hybrid model, combining the ability of GNNs to model relationships in graph-structured data with the global attention capabilities of ViTs. This hybrid model will be designed to handle both spatial and temporal relationships in HAR more effectively by leveraging GNNs for modeling skeleton or body joint data and ViTs for extracting and attending to relevant global features in video frames. The models developed will be compared to existing state-of-the-art methods to determine their efficacy in improving HAR accuracy and generalization, particularly in complex scenarios.

5.3.3 Third Paper: Creating an Arabic Sign Language Video Dataset and Surveying Existing Sign Language Datasets

The third paper will focus on addressing a critical gap in HAR research by creating a large-scale Arabic sign language video dataset and conducting a survey of existing sign language datasets in academia. While sign language recognition has been studied extensively for languages like American Sign Language (ASL) Athitsos et al. (2008) and German Sign Language Li et al. (2020a), there is a notable lack of resources for Arabic sign language Moustafa et al. (2024). This paper aims to fill this gap by developing a diverse, annotated dataset specifically for Arabic sign language gestures, which will be invaluable for researchers developing HAR models tailored to this language. The dataset will include a wide range of signs performed by different individuals, capturing variations in gestures, facial expressions, and hand movements.

This new dataset will provide the foundation for developing models that can accurately recognize Arabic sign language in real-world applications.

In addition to creating the dataset, the paper will include a survey of existing sign language video datasets used in academic research, such as RWTH-PHOENIX-Weather (German) Forster et al. (2012) and ASLLVD (American) Neidle et al. (2012). This survey will provide a comparative analysis of these datasets in terms of size, diversity, and annotation quality, highlighting the specific challenges faced in sign language recognition. After creating the Arabic sign language dataset, the paper will train and evaluate all the models developed in this thesis, including the GNN-CNN and GNN-ViT hybrid models, on this new dataset. This will allow for an assessment of how well these models perform in recognizing sign language compared to traditional action recognition tasks, contributing to advancements in sign language recognition and making these technologies more accessible to Arabic-speaking communities.

5.4 Thesis Contribution

The current thesis makes several significant contributions to the field of HAR. Firstly, it introduces innovative architectures, specifically the CNN-ViT-based models for HAR, which combine CNNs with ViTs to enhance both spatial and temporal feature extraction. This hybrid model offers a novel approach to addressing one of the primary limitations of traditional models (i.e. balancing spatial and temporal feature extraction). CNNs are well known for their efficacy in capturing spatial features, but they struggle with long-range temporal dependencies. In contrast, ViTs excel at capturing global temporal patterns due to their self-attention mechanisms. By integrating both architectures, the thesis presents a more comprehensive and efficient model for HAR that shows superior performance in benchmark datasets, particularly in scenarios involving complex human actions.

Another critical contribution of the thesis is its advancement in transfer learning for HAR through the proposed TransNet and TransNet+ architectures. TransNet model benefits from disassembling the complex 3D-CNN architecture into 2D-CNN and 1D-CNN architectures. By leveraging pre-trained 2D CNNs, such as MobileNet and VGG16, it efficiently performs spatial feature extraction while reducing the need for large HAR-specific datasets. Simultaneously, the 1D-CNN serves as a fast and accurate temporal feature extractor. TransNet+ goes a step further by incorporating autoencoders pre-trained on related tasks, such as human semantic segmentation, to direct feature extraction more effectively. This approach significantly enhances model generalization and reduces the dependency on large datasets, a persistent limitation in HAR research. The thesis demonstrates that transfer learning, when combined with effective feature extraction strategies, can result in more robust HAR models that can be applied to various real-world environments.

The thesis also contributes to the ongoing discourse on data augmentation techniques in HAR. It provides a comprehensive overview of data augmentation techniques in computer vision, focusing on image classification and segmentation. It addresses challenges such as the large data requirements of deep neural networks and the risk of overfitting when data is scarce, emphasizing the role of data augmentation in expanding datasets to improve model generalization. However, the thesis highlights the importance of task-specific augmentation strategies, as different tasks prioritize distinct features. For instance, random rotation enhances classification by introducing shape variations while preserving texture but degrades segmentation performance due to its reliance on precise shape features. The findings underscore the necessity of tailoring augmentation techniques to the feature priorities of specific tasks, particularly for segmentation, where preserving shape integrity is paramount.

Finally, the thesis future works emphasizes the importance of developing inclusive and diverse datasets for HAR applications. In particular, the creation of an Arabic sign language video dataset addresses a crucial gap in the field, enabling more accessible research and development in sign language recognition. This contribution not only advances HAR but also promotes the development of assistive technologies for underrepresented linguistic communities. By training the proposed models on this new dataset, the thesis highlights the potential for hybrid CNN-ViT models and data augmentation strategies to improve sign language recognition systems, making them more adaptable and accurate across various sign languages.

5.5 Concluding Remark

I am pleased to have completed my research on this topic, as it has greatly improved my understanding of HAR mechanisms and deep learning methods. I hope this thesis will contribute to advancing knowledge in this area. However, there are still important questions about the relationship between HAR and deep learning that need further investigation. Throughout this research, more complex questions have emerged than I initially considered, and I look forward to future work in this exciting field.

Chapter 6

Appendix

6.1 Statistical Significance Analysis Methods

This appendix presents two statistical significance analysis methods: the paired-samples t-test and the one-sample t-test (Montgomery and Runger, 2020; Devore, 2000) in Algorithm 2 and Algorithm 3, respectively. The symbols and variables used in Algorithms 2 and 3 are summarized in Table 6.1.

Algorithm 2 processes the data in Table 4.7 from five experimental runs for each model across two contexts out of the total three contexts. It pairs the results from the first run of each context, followed by pairing the results from the second run of each context, continuing in this manner until all five runs have been paired. The algorithm then computes the two-tailed p-value, denoted as p_p, for the paired-samples t-test. Algorithm 3 utilizes the performance results of the state-of-the-art models along with the performance result of the proposed model from Table 4.8. The algorithm then computes the two-tailed p-value, denoted as p_o, for the one-sample t-test.

Symbol	Definition
c_{1i}	Performance of a model in the first context (i.e., 12 frames) of the *i*-th run in the paired-samples test.
c_{2i}	Performance of a model in the second context (i.e., 18 frames) of the *i*-th run in the paired-samples test.
c_{3i}	Performance of a model in the third context (i.e., 24 frames) of the *i*-th run in the paired-samples test.
n_p	Number of paired observations (i.e., the number of runs) in the paired-samples t-test.
d_i	Differences between paired observations ($c_{1i} - c_{2i}$) in the paired-samples t-test.
$\bar{d}$	Mean of the differences between paired observations in the paired-samples t-test.
s_{dp}	Standard deviation of the differences in the paired-samples t-test.
t_{sp}	t-statistic value for the paired-samples t-test.
d_{fp}	Degrees of freedom for the paired-samples t-test, calculated as $n_p - 1$.
p_p	Two-tailed p-value for the paired-samples t-test.
n_c	Number of comparisons for the paired-samples t-test (i.e., 12 vs. 18, 12 vs. 24, and 18 vs. 24 frames).
m_i	Performance of the state-of-the-art *i*-th model used in the one-sample t-test.
n_o	Population size, i.e., the number of state-of-the-art models.
μ_o	Mean performance of the state-of-the-art models.
s_{do}	Standard deviation of the performance of the state-of-the-art models.
t_{so}	t-statistic value for the one-sample t-test.
d_{fo}	Degrees of freedom for the one-sample t-test, calculated as $n_o - 1$.
p_o	Two-tailed p-value for the one-sample t-test.
m_p	Observed performance of the proposed model in the one-sample t-test.
α	Significance level for hypothesis testing, typically set at 0.05.
α_a	Adjusted significance level using the Bonferroni correction.

TABLE 6.1: List of symbols and variables used in the paired-samples t-test (i.e., Algorithm 2) and one-sample t-test (i.e., Algorithm 3).

Algorithm 2 Paired-Samples t-Test Algorithm

```
Input: The model performance on two different frame contexts: $(c_{1i}, c_{2i})$, where $i = 1, 2, \ldots, n_p$ with $n_p = 5$ (experimental runs); the number of comparisons $n_c = 3$; and the significance level $\alpha = 0.05$.
Output: Two-tailed p-value $p_p$.
Calculate the differences between the paired observations:
    $$d_i = c_{1i} - c_{2i}, \quad i = 1, 2, \ldots, n_p.$$
Compute the mean of the differences:
    $$\bar{d} = \frac{\sum_{i=1}^{n_p} d_i}{n_p}.$$
Compute the standard deviation of the differences:
    $$s_{dp} = \sqrt{\frac{\sum_{i=1}^{n_p} (d_i - \bar{d})^2}{n_p - 1}}.$$
Calculate the t-statistic value:
    $$t_{sp} = \frac{\bar{d}}{s_{dp} / \sqrt{n_p}}.$$
Determine the degrees of freedom:
    $$d_{fp} = n_p - 1.$$
Calculate the two-tailed p-value:
    $$p_p = 2 \times f_{cp}(|t_{sp}|, d_{fp}),$$
    where the $f_{cp}$ function uses a statistical t-distribution table/software (e.g., the SciPy library and Python programming language) to find the critical t-value corresponding to the calculated t-statistic ($t_{sp}$) and degrees of freedom ($d_{fp}$).
Apply the Bonferroni correction:
    $$\alpha_a = \alpha / n_c,$$
    where $n_c$ is the number of comparisons and $n_c = 3$ for the case in Table 4.7.
if $p_p < \alpha_a$ then
    Reject the null hypothesis ($H_{p0}$), i.e., there is a significant difference.
else
    Fail to reject the null hypothesis, i.e., there is not enough evidence to suggest a significant difference.
end if
```

Algorithm 3 One-Sample T-test Algorithm

```
Input: Performance results of state-of-the-art models: $m_1, m_2, \ldots, m_{n_o}$; the performance result of the proposed model: $m_p$; and the significance level $\alpha = 0.05$.
Output: Two-tailed p-value $p_o$.
Compute the population mean:
    $\mu_0 = \frac{\sum_{i=1}^{n_o} m_i}{n_o}.$
Compute the population standard deviation:
    $s_{do} = \sqrt{\frac{\sum_{i=1}^{n_o} (m_i - \mu_0)^2}{n_o - 1}}.$
Calculate the t-statistic:
    $t_{so} = \frac{m_p - \mu_0}{s_{do} / \sqrt{n_o}}.$
Determine the degrees of freedom:
    $d_{fo} = n_o - 1.$
Calculate the two-tailed p-value:
    $p_o = 2 \times f_{co}(|t_{so}|, d_{fo}),$
where the $f_{co}$ function uses a statistical t-distribution table/software (e.g., the SciPy library and Python programming language) to find the critical t-value corresponding to the calculated t-statistic ($t_{so}$) and degrees of freedom ($d_{fo}$).
if $p_o < \alpha$ then
    Reject the null hypothesis ($H_{o0}$), i.e., there is a significant difference.
else
    Fail to reject the null hypothesis, i.e., there is not enough evidence to suggest a significant difference.
end if
```

References

tocchapterReferences

Amany Abdelbaky and Saleh Aly. Human action recognition based on simple deep convolution network pcanet. In *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, pages 257–262. IEEE, 2020.

Muhammad Abdullah, Mobeen Ahmad, and Dongil Han. Facial expression recognition in videos: An cnn-lstm based model for video classification. In *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–3. IEEE, 2020.

Amina Adadi and Mohammed Berrada. Explainable ai for healthcare: from black box to interpretable models. In *Embedded systems and artificial intelligence: proceedings of ESAI 2019, Fez, Morocco*, pages 327–337. Springer, 2020.

Basant Adel, Asmaa Badran, Nada E Elshami, Ahmad Salah, Ahmed Fathalla, and Mahmoud Bekhit. A survey on deep learning architectures in human activities recognition application in sports science, healthcare, and security. In *The International Conference on Innovations in Computing Research*, pages 121–134. Springer, 2022.

Preeti Agarwal and Mansaf Alam. A lightweight deep learning model for human activity recognition on edge devices. *Procedia Computer Science*, 167:2364–2373, 2020.

Hamid Ahmadabadi, Omid Nejati Manzari, and Ahmad Ayatollahi. Distilling knowledge from cnn-transformer models for enhanced human action recognition. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 180–184. IEEE, 2023.

K Akila. Recognition of inter-class variation of human actions in sports video. *Journal of Intelligent & Fuzzy Systems*, 43(4):5251–5262, 2022.

Khaled Alomar and Xiaohao Cai. Transnet: A transfer learning-based network for human action recognition. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1825–1832. IEEE, 2023.

Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging*, 9 (2):46, 2023.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Talha Anwar and Seemab Zakir. Effect of image augmentation on ecg image classification using deep learning. In *2021 International Conference on Artificial Intelligence (ICAI)*, pages 182–186. IEEE, 2021.

Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, page 142, 2015.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

J Arunnehru, G Chamundeeswari, and S Prasanna Bharathi. Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos. *Procedia computer science*, 133:471–477, 2018.

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Ruzena Bajcsy, Franc Solina, and Alok Gupta. Segmentation versus object representation—are they separable? In *Analysis and interpretation of range images*, pages 207–223. Springer, 1990.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos.

Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017a.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017b.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12:43–77, 1994.

SH Shabbeer Basha, Viswanath Pulabaigari, and Snehasis Mukherjee. An information-rich sampling technique over spatio-temporal cnn for classification of human actions in videos. *Multimedia Tools and Applications*, 81(28):40431–40449, 2022.

Florian Baumann. Action recognition with hog-of features. In *German Conference on Pattern Recognition*, pages 243–248. Springer, 2013.

Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41):30509–30555, 2020.

Mohammad Mahdi Bejani and Mehdi Ghatee. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, 54(8): 6391–6438, 2021.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Gibran Benitez-Garcia, Lidia Prudente-Tixteco, Luis Carlos Castro-Madrid, Rocio Toscano-Medina, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Luis

Javier Garcia Villalba. Improving real-time hand gesture recognition with semantic segmentation. *Sensors*, 21(2):356, 2021.

Eze Benson, Michael P Pound, Andrew P French, Aaron S Jackson, and Tony P Pridmore. Deep hourglass for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 419–428. Springer, 2018.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

Jordan J Bird, Chloe M Barnes, Luis J Manso, Anikó Ekárt, and Diego R Faria. Fruit quality and defect image classification with conditional gan data augmentation. *Scientia Horticulturae*, 293:110684, 2022.

Aaron F Bobick and James W Davis. Action recognition using temporal templates. In *Motion-Based Recognition*, pages 125–146. Springer, 1997.

Sornkitja Boonprong, Chunxiang Cao, Wei Chen, Xiliang Ni, Min Xu, and Bipin Kumar Acharya. The classification of noise-afflicted remotely sensed data using three machine-learning techniques: effect of different levels and types of noise on accuracy. *ISPRS International Journal of Geo-Information*, 7(7):274, 2018.

Alan C Bovik. *The essential guide to video processing*. Academic Press, 2009.

Ajay Kumar Boyat and Brijendra Kumar Joshi. A review paper: noise models in digital image processing. *arXiv preprint arXiv:1505.03489*, 2015.

Gianni Brauwers and Flavius Frasincar. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3279–3298, 2021.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jason Brownlee. *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery, 2018.

Neil G Burnet, Jessica E Scaife, Marina Romanchikova, Simon J Thomas, Amy M Bates, Emma Wong, David J Noble, Leila Ea Shelley, Simon J Bond, Julia R Forman, et al. Applying physical science techniques and cern technology to an unsolved problem in radiation treatment for cancer: the multidisciplinary 'voxtox'research programme. *CERN ideaSquare journal of experimental innovation*, 1(1):3, 2017.

Xiaohao Cai, Raymond Chan, Mila Nikolova, and Tieyong Zeng. A three-stage approach for segmenting degraded color images: smoothing, lifting and thresholding (SLaT). *J. Sci. Comput.*, 72(3):1313–1332, 2017. ISSN 0885-7474. . URL https://doi.org/10.1007/s10915-017-0402-2.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-8(6):679–698, 1986.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Ángela Casado-García, César Domínguez, Manuel García-Domínguez, Jónathan Heras, Adrián Inés, Eloy Mata, and Vico Pascual. Clodsa: a tool for augmentation in classification, localization, detection, semantic segmentation and instance segmentation tasks. *BMC bioinformatics*, 20(1):1–14, 2019.

Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

Bhaskar Chakraborty, Michael B Holte, Thomas B Moeslund, Jordi Gonzalez, and F Xavier Roca. A selective spatio-temporal interest point detector for human action recognition in complex scenes. In *2011 International Conference on Computer Vision*, pages 1776–1783. IEEE, 2011.

Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

Bo Chen, Fangzhou Meng, Hongying Tang, and Guanjun Tong. Two-level attention module based on spurious-3d residual networks for human action recognition. *Sensors*, 23(3):1707, 2023.

Chen Chen, Chen Qin, Cheng Ouyang, Zeju Li, Shuo Wang, Huaqi Qiu, Liang Chen, Giacomo Tarroni, Wenjia Bai, and Daniel Rueckert. Enhancing mr image segmentation with realistic adversarial data augmentation. *Medical Image Analysis*, 82:102597, 2022.

Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*, 2021.

Hong Chen and Sung-Tae Jung. Enhancement of tongue segmentation by using data augmentation. *The Journal of Korea Institute of Information, Electronics, and Communication Technology*, 13(5):313–322, 2020.

Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1910–1921, 2022.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

Shoushui Chen, Xin Yang, and Guo Cao. Impulse noise suppression with an augmentation of ordered difference noise detector and an adaptive variational method. *Pattern Recognition Letters*, 30(4):460–467, 2009.

Xingyuan Chen, Peishi Jiang, Justine EC Missik, Zhongming Gao, Brittany Verbeke, and Heping Liu. Opening the black box of lstm models using xai. In *AGU Fall Meeting Abstracts*, volume 2020, pages H191–06, 2020.

Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5): 545–563, 2021.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Marco Domenico Cirillo, David Abramian, and Anders Eklund. What is the best data augmentation for 3d brain tumor segmentation? In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 36–40. IEEE, 2021.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20: 273–297, 1995.

Ionut Cosmin Duta, Bogdan Ionescu, Kiyoharu Aizawa, and Nicu Sebe. Spatio-temporal vector of locally max pooled features for action recognition in videos. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3097–3106, 2017.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

Somayeh Danafar and Niloofar Gheissari. Action recognition for surveillance applications using optic flow and svm. In *Computer Vision–ACCV 2007: 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18-22, 2007, Proceedings, Part II 8*, pages 457–466. Springer, 2007.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. *arXiv preprint arXiv:2209.02535*, 2022.

Debapratim Das Dawn and Soharab Hossain Shaikh. A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *The Visual Computer*, 32:289–306, 2016.

Sharana Dharshikgan Suresh Dass, Hrishav Bakul Barua, Ganesh Krishnasamy, Raveendran Paramesran, and Raphaël C-W Phan. Actnetformer: Transformer-resnet hybrid method for semi-supervised action recognition in videos. In *International Conference on Pattern Recognition*, pages 343–359. Springer, 2025.

Geoff Delaney, Susannah Jacob, Carolyn Featherstone, and Michael Barton. The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 104(6):1129–1137, 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2):20539517211035955, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

Jay L Devore. Probability and statistics. *Pacific Grove: Brooks/Cole*, 2000.

Chhavi Dhiman and Dinesh Kumar Vishwakarma. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on Image Processing*, 29:3835–3844, 2020.

Sotirios Diamantas and Kostas Alexis. Optical flow based background subtraction with a moving camera: Application to autonomous driving. In *International Symposium on Visual Computing*, pages 398–409. Springer, 2020.

Ali Diba, Ali Mohammad Pazandeh, and Luc Van Gool. Efficient two-stream motion and appearance 3d cnns for video classification. *arXiv preprint arXiv:1608.08851*, 2016.

Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*, 2017.

Giovanni Diraco, Gabriele Rescio, Andrea Caroppo, Andrea Manni, and Alessandro Leone. Human action recognition in smart living services and applications: context awareness, data availability, personalization, and privacy. *Sensors*, 23(13):6040, 2023.

Gilad Divon and Ayellet Tal. Viewpoint estimation—insights & model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018.

Y. Djenouri and A.N. Belbachir. A hybrid visual transformer for efficient deep human activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

PN Druzhkov and VD Kustikova. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis*, 26(1):9–15, 2016.

Nidhi Dua, Shiva Nand Singh, and Vijay Bhaskar Semwal. Multi-input cnn-gru based human activity recognition using wearable sensors. *Computing*, 103(7):1461–1478, 2021.

Debidatta Dwibedi, Pierre Sermanet, and Jonathan Tompson. Temporal reasoning in videos using convolutional gated recurrent units. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1111–1116, 2018.

Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001.

Hany El-Ghaish, Mohamed E Hussien, Amin Shoukry, and Rikio Onai. Human action recognition based on integrating body pose, part shape, and motion. *IEEE Access*, 6: 49040–49055, 2018.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016.

RW Farebrother. Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–250, 1976.

Ahmeed Suliman Farhan, Muhammad Khalid, and Umar Manzoor. Combined oriented data augmentation method for brain mri images. *IEEE Access*, 2025.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, 2012.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Chi Geng and JianXin Song. Human action recognition based on convolutional neural networks with a convolutional auto-encoder. In *2015 5th International Conference on Computer Sciences and Automation Engineering (ICCSAE 2015)*, pages 933–938. Atlantis Press, 2016.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019.

Ilias Giannakos, Eirini Mathe, Evaggelos Spyrou, and Phivos Mylonas. A study on the effect of occlusion in human activity recognition. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 473–482, 2021.

Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30, 2017.

Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Mahesh Gour, Sweta Jain, and T Sunil Kumar. Residual learning based cnn for breast cancer histopathological image classification. *International Journal of Imaging Systems and Technology*, 30(3):621–635, 2020.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

Alexander Grushin, Derek D Monner, James A Reggia, and Ajay Mishra. Robust human action recognition via long short-term memory. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013.

Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition using sparse representation on covariance manifolds of optical flow. In *2010 7th IEEE international conference on advanced video and signal based surveillance*, pages 188–195. IEEE, 2010.

Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10905–10914, 2019.

Booz Allen Hamilton. Find the nuclei in divergent images to advance medical discovery. `https://www.kaggle.com/c/data-science-bowl-2018/data`, 2018.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. Db-lstm: Densely-connected bi-directional lstm for human action recognition. *Neurocomputing*, 444:319–331, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Wenfeng He, Chulong Zhang, Jingjing Dai, Lin Liu, Tangsheng Wang, Xuan Liu, Yuming Jiang, Na Li, Jing Xiong, Lei Wang, et al. A statistical deformation model-based data augmentation method for volumetric medical image segmentation. *Medical Image Analysis*, 91:102984, 2024.

Kartik Hegde, Rohit Agrawal, Yulun Yao, and Christopher W Fletcher. Morph: Flexible acceleration for 3d cnn-based video understanding. In *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 933–946. IEEE, 2018.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

Koki Hirota and Takashi Komuro. Grasping action recognition in vr environment using object shape and position information. In *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–2. IEEE, 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Alexander Hoelzemann and Kristof Van Laerhoven. Digging deeper: Towards a better understanding of transfer learning for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 50–54, 2020.

Berthold KP Horn. Bg schunck determining optical flow. *Artificial intelligence*, 17(1-3): 185–203, 1981.

Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

Rui Hou, Chen Chen, Rahul Sukthankar, and Mubarak Shah. An efficient 3d cnn for action/object segmentation in video. *arXiv preprint arXiv:1907.08895*, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Yu Hu, Yongkang Wong, Wentao Wei, Yu Du, Mohan Kankanhalli, and Weidong Geng. A novel attention-based hybrid cnn-rnn architecture for semg-based gesture recognition. *PloS one*, 13(10):e0206049, 2018.

Zheng-ping Hu, Rui-xue Zhang, Yue Qiu, Meng-yao Zhao, and Zhe Sun. 3d convolutional networks with multi-layer-pooling selection fusion for video classification. *Multimedia Tools and Applications*, 80(24):33179–33192, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA annual symposium proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.

Catherine Huyghe, Nacim Ihaddadene, Thomas Haessle, and Chabane Djeraba. Human action recognition based on body segmentation models. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE, 2021.

Noor A Ibraheem, Mokhtar M Hasan, Rafiqul Z Khan, and Pramod K Mishra. Understanding color models: a review. *ARPN Journal of science and technology*, 2(3): 265–275, 2012.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Sarah Ali Abdelaziz Ismael, Ammar Mohammed, and Hesham Hefny. An enhanced deep learning approach for brain cancer mri images classification using residual networks. *Artificial intelligence in medicine*, 102:101779, 2020.

Neziha Jaouedi, Noureddine Boujnah, and Med Salim Bouhlel. A new hybrid deep learning model for human action recognition. *Journal of King Saud University-Computer and Information Sciences*, 32(4):447–453, 2020.

I Jegham et al. Multi-view vision transformer for driver action recognition. *SpringerLink*, 2022.

Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

MI Jordan. Serial order: a parallel distributed processing approach. technical report, june 1985-march 1986. Technical report, California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, 1986.

Soumya Joshi, Dhirendra Kumar Verma, Gaurav Saxena, and Amit Paraye. dropouts in training a convolutional neural network model for image classification. In *International Conference on Advances in Computing and Data Sciences*, pages 282–293. Springer, 2019.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

M Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020.

M Kalfaoglu et al. Human action recognition with transformers. *SpringerLink*, 2022.

Ibrahem Kandel and Mauro Castelli. Improving convolutional neural networks performance for image classification using test time augmentation: a case study using mura dataset. *Health information science and systems*, 9(1):1–22, 2021.

Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization. *arXiv preprint arXiv:1707.07103*, 2017.

Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2): 358–367, 1988.

Sachin Kansal, Sagar Jha, and Prathamesh Samal. Dl-dare: Deep learning-based different activity recognition for the human–robot interaction environment. *Neural Computing and Applications*, 35(16):12029–12037, 2023.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, pages 1–27, 2021.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Vladimir Khryashchev and Roman Larionov. Wildfire segmentation on satellite images using deep learning. In *2020 Moscow Workshop on Electronic and Networking Technologies (MWENT)*, pages 1–5. IEEE, 2020.

Chulyeon Kim, Jiyoung Lee, Taekjin Han, and Young-Min Kim. A hybrid framework combining background subtraction and deep neural networks for rapid person detection. *Journal of Big Data*, 5(1):1–24, 2018a.

Pil-Soo Kim, Dong-Gyu Lee, and Seong-Whan Lee. Discriminative context learning with gated recurrent unit for group activity recognition. *Pattern Recognition*, 76: 149–161, 2018b.

Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European conference on computer vision*, pages 201–214. Springer, 2012.

Marco Klaiber, Daniel Sauter, Hermann Baumgartl, and Ricardo Buettner. A systematic literature review on transfer learning for 3d-cnns. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–10. IEEE, 2021.

Yongqiang Kong, Yunhong Wang, and Annan Li. Spatiotemporal saliency representation learning for video action recognition. *IEEE Transactions on Multimedia*, 24:1515–1528, 2021.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Raivo Koot, Markus Hennerbichler, and Haiping Lu. Evaluating transformers for lightweight action recognition. *arXiv preprint arXiv:2111.09641*, 2021.

Okan Kopuklu, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

Pranjal Kumar, Siddhartha Chauhan, and Lalit Kumar Awasthi. Human activity recognition (har) using deep learning: Review, methodologies, progress and future research directions. *Archives of Computational Methods in Engineering*, 31(1):179–219, 2024.

Rahul Kumar and Shailender Kumar. A survey on intelligent human action recognition techniques. *Multimedia Tools and Applications*, 83(17):52653–52709, 2024.

Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 345–362. Springer, 2020.

Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. Handling annotation uncertainty in human activity recognition. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 109–117, 2019.

Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64: 107–123, 2005.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Rethinking data augmentation: Self-supervision and self-distillation. 2019.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3054–3063, 2021.

Mei Chee Leong, Dilip K Prasad, Yong Tsui Lee, and Feng Lin. Semi-cnn architecture for effective spatio-temporal learning in action recognition. *Applied Sciences*, 10(2): 557, 2020.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020a.

Hua Li and Jun Wang. A neural network model for optical flow computation. In *Neural Networks and Pattern Recognition*, pages 57–76. Elsevier, 1998.

Jun Li, Xianglong Liu, Mingyuan Zhang, and Deqing Wang. Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recognition*, 98:107037, 2020b.

Lianwei Li, Shiyin Qin, Ning Yang, Li Hong, Yang Dai, and Zhiqiang Wang. Lvnet: A lightweight volumetric convolutional neural network for real-time and high-performance recognition of 3d objects. *Multimedia Tools and Applications*, 83 (21):61047–61063, 2024.

Wei Li, Chen Chen, Mengmeng Zhang, Hengchao Li, and Qian Du. Data augmentation for hyperspectral image classification with deep cnn. *IEEE Geoscience and Remote Sensing Letters*, 16(4):593–597, 2018a.

Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2020c.

Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 203–220. Springer, 2016.

Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018b.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

Lydia Lindner, Dominik Narnhofer, Maximilian Weber, Christina Gsaxner, Malgorzata Kolodziej, and Jan Egger. Using synthetic training data for deep learning-based gbm segmentation. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6724–6729. IEEE, 2019.

Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *arXiv Preprint, CoRR, abs/1506.00019*, 2015.

An-An Liu, Ning Xu, Yu-Ting Su, Hong Lin, Tong Hao, and Zhao-Xuan Yang. Single/multi-view human action recognition via regularized multi-task learning. *Neurocomputing*, 151:544–553, 2015.

GuoJun Liu, XiangLong Tang, JianHua Huang, JiaFeng Liu, and Da Sun. Hierarchical model-based human motion tracking via unscented kalman filter. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Honghui Liu, Changjian Wang, and Yuxing Peng. Data augmentation with illumination correction in sematic segmentation. In *Journal of Physics: Conference Series*, volume 2025, page 012009. IOP Publishing, 2021.

Xiao Liu, De-yu Qi, and Hai-bin Xiao. Construction and evaluation of the human behavior recognition model in kinematics under deep learning. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–9, 2020.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

Zhiqiang Liu, Paul Chow, Jinwei Xu, Jingfei Jiang, Yong Dou, and Jie Zhou. A uniform architecture design for accelerating 2d and 3d cnns on fpgas. *Electronics*, 8(1):65, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Wei-Lwun Lu and James J Little. Simultaneous tracking and action recognition using the pca-hog descriptor. In *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pages 6–6. IEEE, 2006.

Weirui Lu, Xiaofen Xing, Bolun Cai, and Xiangmin Xu. Listwise view ranking for image cropping. *IEEE Access*, 7:91904–91911, 2019.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Mark Lyksborg, Oula Puonti, Mikael Agn, and Rasmus Larsen. An ensemble of 2d convolutional neural networks for tumor segmentation. In *Scandinavian conference on image analysis*, pages 201–211. Springer, 2015.

Rui Ma, Pin Tao, and Huiyun Tang. Optimizing data augmentation for semantic segmentation on small-scale dataset. In *Proceedings of the 2nd international conference on control and computer vision*, pages 77–81, 2019.

Upal Mahbub, Hafiz Imtiaz, and Md Atiqur Rahman Ahad. An optical flow based approach for action recognition. In *14th international conference on computer and information technology (ICCIT 2011)*, pages 646–651. IEEE, 2011.

Najeeb ur Rehman Malik, Syed Abdul Rahman Abu-Bakar, Usman Ullah Sheikh, Asma Channa, and Nirvana Popescu. Cascading pose features with cnn-lstm for multiview human action recognition. *Signals*, 4(1):40–55, 2023.

Dimitrios Mallios and Xiaohao Cai. Deep rectum segmentation for image guided radiation therapy with synthetic data. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 975–979. IEEE, 2021.

Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. Three-stream 3d/1d cnn for fine-grained action classification and segmentation in table tennis. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 35–41, 2021.

Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European conference on computer vision*, pages 579–596. Springer, 2016.

Francisco Massa, Renaud Marlet, and Mathieu Aubry. Crafting a multi-task cnn for viewpoint estimation. *arXiv preprint arXiv:1609.03894*, 2016.

Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Wei Sun, Frederick Tung, and Leonid Sigal. Interpretable spatio-temporal attention for video action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

BH Menze and AB Jakab. S., kalpathy-cramer j, farahani k, kirby j, burren y, porz n, slotboom j, wiest r, leemput kv. the multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans Med Imaging*, 34:1993–2024, 2015.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

Andres Milioto and Cyrill Stachniss. Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using cnns. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7094–7100. IEEE, 2019.

S Milyaev and I Laptev. Towards reliable object detection in noisy images. *Pattern Recognition and Image Analysis*, 27(4):713–722, 2017.

Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Al-Fuqaha Mohammadi. Mohammadi m., al-fuqaha a., sorour s., guizani m. *Deep learning for IoT big data and streaming analytics: A survey, IEEE Communications Surveys & Tutorials*, 20(4):2923–2960, 2018.

Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John wiley & sons, 2020.

Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors*, 23(4):2182, 2023.

Mehdi Habibzadeh Motlagh, Mahboobeh Jannesari, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha. Breast cancer histopathological image classification: A deep learning approach. *BioRxiv*, page 242818, 2018.

AMA Moustafa, MS Mohd Rahim, Mahmoud M Khattab, Akram M Zeki, Safaa S Matter, Amr Mohmed Soliman, and Abdelmoty M Ahmed. Arabic sign language recognition systems: A systematic review. *Indian Journal of Computer Science and Engineering*, 15:1–18, 2024.

Khan Muhammad, Amin Ullah, Ali Shariq Imran, Muhammad Sajjad, Mustafa Servet Kiran, Giovanna Sannino, Victor Hugo C de Albuquerque, et al. Human action recognition using attention based lstm network with dilated cnn features. *Future Generation Computer Systems*, 125:820–830, 2021.

Ronald Mutegeki and Dong Seog Han. A cnn-lstm approach to human activity recognition. In *2020 international conference on artificial intelligence in information and communication (ICAIIC)*, pages 362–366. IEEE, 2020.

Loris Nanni, Michelangelo Paci, Sheryl Brahnam, and Alessandra Lumini. Comparison of different image data augmentation approaches. *Journal of Imaging*, 7 (12):254, 2021.

Sharan Narang, Erich Elsen, Gregory Diamos, and Shubho Sengupta. Exploring sparsity in recurrent neural networks. *arXiv preprint arXiv:1704.05119*, 2017.

Humza Naveed, Saeed Anwar, Munawar Hayat, Kashif Javed, and Ajmal Mian. Survey: Image mixing and deleting for data augmentation. *Engineering Applications of Artificial Intelligence*, 131:107791, 2024.

Tiago S Nazaré, Gabriel B Costa, Welinton A Contato, and Moacir Ponti. Deep convolutional neural networks and noisy images. In *Iberoamerican Congress on Pattern Recognition*, pages 416–424. Springer, 2017.

Carol Neidle, Ashwin Thangali, and Stan Sclaroff. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC*, 2012.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

Xin Ning, Fangzhe Nan, Shaohui Xu, Lina Yu, and Liping Zhang. Multi-view frontal face image generation: a survey. *Concurrency and Computation: Practice and Experience*, page e6147, 2020.

Eshed Ohn-Bar and Mohan Trivedi. Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 465–470, 2013.

Kazuki Omi, Jun Kimata, and Toru Tamaki. Model-agnostic multi-domain learning with domain-specific adapters for action recognition. *IEICE TRANSACTIONS on Information and Systems*, 105(12):2119–2126, 2022.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Michalis Papakostas, Theodoros Giannakopoulos, Fillia Makedon, and Vangelis Karkaletsis. Short-term recognition of human activities using convolutional neural networks. In *2016 12th international conference on signal-image technology & internet-based systems (SITIS)*, pages 302–307. IEEE, 2016.

Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54(3):2259–2322, 2021.

R Pascanu. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2013.

S Nissi Paul and Y Jayanta Singh. Survey on video analysis of human walking motion. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 7(3): 99–122, 2014.

Pornntiwa Pawara, Emmanuel Okafor, Lambert Schomaker, and Marco Wiering. Data augmentation for plant classification. In *International conference on advanced concepts for intelligent vision systems*, pages 615–626. Springer, 2017.

Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):773–786, 2018.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Tri-Cong Pham, Chi-Mai Luong, Muriel Visani, and Van-Dung Hoang. Deep cnn and data augmentation for skin lesion classification. In *Asian Conference on Intelligent Information and Database Systems*, pages 573–582. Springer, 2018.

Enrico Picco, Piotr Antonik, and Serge Massar. High speed human action recognition using a photonic reservoir computer. *Neural Networks*, 2023.

Danil V Prokhorov, LA Feldkarnp, and I Yu Tyukin. Adaptive behavior with fixed weights in rnn: an overview. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 3, pages 2018–2022. IEEE, 2002.

Yongliang Qiao, Daobilige Su, He Kong, Salah Sukkarieh, Sabrina Lomax, and Cameron Clark. Data augmentation for deep learning based cattle segmentation in precision livestock farming. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pages 979–984. IEEE, 2020.

Tiexin Qin, Ziyuan Wang, Kelei He, Yinghuan Shi, Yang Gao, and Dinggang Shen. Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1419–1423. IEEE, 2020.

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Martin Rajnoha, Radim Burget, and Lukas Povoda. Image background noise impact on convolutional neural network training. In *2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 1–4. IEEE, 2018.

Manoj Ramanathan, Wei-Yun Yau, and Eam Khwang Teoh. Human action recognition with video data: research and evaluation challenges. *IEEE Transactions on Human-Machine Systems*, 44(5):650–663, 2014.

P Ramya and Rajendran Rajeswari. Human action recognition using distance transform and entropy based features. *Multimedia Tools and Applications*, 80: 8147–8173, 2021.

Keerthana Rangasamy, Muhammad Amir As'ari, Nur Azmina Rahmad, Nurul Fathiah Ghazali, and Saharudin Ismail. Deep learning in sport video analysis: a review. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18(4):1926–1933, 2020.

Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M Patel, Carlos D Castillo, and Rama Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, 2018.

J Ranstam and JA Cook. Lasso regression. *Journal of British Surgery*, 105(10):1348–1348, 2018.

Aditya Ravishankar, S Anusha, HK Akshatha, Anjali Raj, S Jahnavi, and J Madhura. A survey on noise reduction techniques in medical images. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 1, pages 385–389. IEEE, 2017.

Abhisek Ray, Maheshkumar H Kolekar, Raman Balasubramanian, and Adel Hafiane. Transfer learning enhanced vision-based human activity recognition: a decade-long analysis. *International Journal of Information Management Data Insights*, 3(1):100142, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Suraj Prakash Sahoo, Samit Ari, Kamalakanta Mahapatra, and Saraju P Mohanty. Har-depth: a novel framework for human action recognition using sequential learning and depth estimated history images. *IEEE transactions on emerging topics in computational intelligence*, 5(5):813–825, 2020.

Rashmi Saini and Vinod Maan. Human activity and gesture recognition: A review. In *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, pages 1–2. IEEE, 2020.

Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (CSUR)*, 54(10s):1–25, 2022.

Allah Bux Sargano, Plamen Angelov, and Zulfiqar Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *applied sciences*, 7(1):110, 2017.

Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Mahdieh Shabanian, Markus Wenzel, and John P DeVincenzo. Infant brain age classification: 2d cnn outperforms 3d cnn in small dataset. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 626–633. SPIE, 2022.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021.

Ahmed Sharshar, Ahmed H Abo Eitta, Ahmed Fayez, Mohamed A Khamis, Ahmed B Zaky, and Walid Gomaa. Camera coach: activity recognition and assessment using thermal and rgb videos. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Learning long-term dependencies for action recognition with a biologically-inspired deep network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 716–725, 2017.

Jia Shijie, Wang Ping, Jia Peiyi, and Hu Siping. Research on data augmentation for image classification based on convolution neural networks. In *2017 Chinese automation congress (CAC)*, pages 4165–4170. IEEE, 2017.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Na Shu, Q Tang, and Haihua Liu. A bio-inspired approach modeling spiking neural networks of visual cortex for human action recognition. In *2014 international joint conference on neural networks (IJCNN)*, pages 3450–3457. IEEE, 2014.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014a.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014b.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

Pawan Kumar Singh, Soumalya Kundu, Titir Adhikary, Ram Sarkar, and Debotosh Bhattacharjee. Progress of human action recognition research in the last ten years: a comprehensive survey. *Archives of Computational Methods in Engineering*, pages 1–41, 2021.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015.

Anastasios Stamoulakatos, Javier Cardona, Craig Michie, Ivan Andonovic, Pavlos Lazaridis, Xavier Bellekens, Robert Atkinson, Md Moinul Hossain, and Christos

Tachtatzis. A comparison of the performance of 2d and 3d convolutional neural networks for subsea survey video classification. In *OCEANS 2021: San Diego–Porto*, pages 1–10. IEEE, 2021.

Daobilige Su, He Kong, Yongliang Qiao, and Salah Sukkarieh. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics. *Computers and Electronics in Agriculture*, 190:106418, 2021.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841, 2019.

D Sun, F Dornaika, and N Barrena. Hsmix: Hard and soft mixing data augmentation for medical image segmentation. *Information Fusion*, 115:102741, 2025.

Lin Sun, Kui Jia, Kevin Chen, Dit-Yan Yeung, Bertram E Shi, and Silvio Savarese. Lattice long short-term memory for human action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2147–2156, 2017.

Mengshu Sun, Pu Zhao, Mehmet Gungor, Massoud Pedram, Miriam Leeser, and Xue Lin. 3d cnn acceleration on fpga using hardware-aware pruning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

supervise.ly. Supervisely person dataset. https://supervise.ly, 2018.

G.A.S. Surek, L.O. Seman, S.F. Stefenon, V.C. Mariani, and L.d.S. Coelho. Video-based human activity recognition using deep learning approaches. *Sensors*, 23(14):6384, 2023.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Risako Tanigawa, Yasunori Ishii, Kazuki Kozuka, and Takayoshi Yamashita. Invisible-to-visible: Privacy-aware human instance segmentation using airborne ultrasound via collaborative learning variational autoencoder. *arXiv preprint arXiv:2204.07280*, 2022.

Luke Taylor and Geoff Nitschke. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE, 2018.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5552–5561, 2019.

Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11):1473–1488, 2008.

Anwaar Ulhaq, Naveed Akhtar, Ganna Pogrebna, and Ajmal Mian. Vision transformers for action recognition: A survey. *arXiv preprint arXiv:2209.05700*, 2022.

Irem Ulku and Erdem Akagündüz. A survey on deep learning-based architectures for semantic segmentation on 2d images. *Applied Artificial Intelligence*, pages 1–45, 2022.

Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE access*, 6:1155–1166, 2017.

Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1510–1517, 2017.

Vasiliki I Vasileiou, Nikolaos Kardaris, and Petros Maragos. Exploring temporal context and human movement dynamics for online action detection in videos. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1431–1435. IEEE, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 4041–4049, 2015.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

Heng Wang and A Kl. aser, c. schmid, and c.-l. liu,"action recognition by dense trajectories,". In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, pages 3169–3176, 2011.

Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

Jun Wang, Limin Xia, and Xin Wen. Cmf-transformer: cross-modal fusion transformer for human action recognition. *Machine Vision and Applications*, 35(5):114, 2024.

Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015a.

Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015b.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018a.

Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021a.

Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through hr optical flow estimation. In *Asian Conference on Computer Vision*, pages 514–529. Springer, 2018b.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021b.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018c.

Yu Wang, Quanjun Song, Tingting Ma, Yong Chen, Hao Li, and Rongkai Liu. Transformation classification of human squat/sit-to-stand based on multichannel information fusion. *International Journal of Advanced Robotic Systems*, 19(4): 17298806221103708, 2022.

Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017.

Zhuo Wang, Zhezhou Yu, Yao Wang, Huimao Zhang, Yishan Luo, Lin Shi, Yan Wang, and Chunjie Guo. 3d compressed convolutional neural network differentiates neuromyelitis optical spectrum disorders from multiple sclerosis using automated white matter hyperintensities segmentations. *Frontiers in Physiology*, 11:612928, 2020.

Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*, pages 650–663. Springer, 2008.

Baoxin Wu, Chunfeng Yuan, and Weiming Hu. Human action recognition based on context-dependent graph kernels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2609–2616, 2014.

Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876*, 7(8), 2015a.

Shandong Wu, Omar Oreifej, and Mubarak Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *2011 International conference on computer vision*, pages 1419–1426. IEEE, 2011.

Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 461–470, 2015b.

Kun Xia, Jianguang Huang, and Hanyu Wang. Lstm-cnn architecture for human activity recognition. *IEEE Access*, 8:56855–56866, 2020.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826, 2023.

Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.

Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Guoan Yang, Yong Yang, Zhengzhi Lu, Junjie Yang, Deyang Liu, Chuanbo Zhou, and Zien Fan. Sta-tsn: Spatial-temporal attention temporal segment network for action recognition in video. *PloS one*, 17(3):e0265115, 2022a.

Hao Yang, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J Maybank. Asymmetric 3d convolutional neural networks for action recognition. *Pattern recognition*, 85:1–12, 2019.

Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022b.

Ruoyu Yang, Shubhendu Kumar Singh, Mostafa Tavakkoli, Nikta Amiri, Yongchao Yang, M Amin Karami, and Rahul Rai. Cnn-lstm deep learning architecture for computer vision-based modal frequency detection. *Mechanical Systems and signal processing*, 144:106885, 2020.

Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022c.

Guangle Yao, Tao Lei, and Jiandan Zhong. A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, 118:14–22, 2019.

Xi Ye and Guillaume-Alexandre Bilodeau. A unified model for continuous conditional video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3603–3612, 2023.

Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

Bruce XB Yu, Yan Liu, and Keith CC Chan. Effective human activity recognition based on small datasets. *arXiv preprint arXiv:2004.13977*, 2020.

W Yu et al. Swin-fusion: Swin-transformer with feature fusion for human action recognition. *Neural Processing Letters*, 2023.

Novanto Yudistira and Takio Kurita. Gated spatio and temporal convolutional neural network for activity recognition: towards gated multimodal deep learning. *EURASIP Journal on Image and Video Processing*, 2017(1):1–12, 2017.

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005, 2019.

Jiawei Zhang, Yanchun Zhang, and Xiaowei Xu. Objectaug: object-level data augmentation for semantic image segmentation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021a.

Lijing Zhang and Yingli Liang. Motion human detection based on background subtraction. In *2010 Second International workshop on Education Technology and computer science*, volume 1, pages 284–287. IEEE, 2010.

Ning Zhang, Zeyuan Hu, Sukhwan Lee, and Eungjoo Lee. Human action recognition based on global silhouette and local optical flow. In *International Symposium on Mechanical Engineering and Material Science (ISMEMS 2017)*, pages 1–5. Atlantis Press, 2017a.

Wei Zhang, Yuchun Fang, and Zhengyan Ma. The effect of task similarity on deep transfer learning. In *International Conference on Neural Information Processing*, pages 256–265. Springer, 2017b.

Xing-Yuan Zhang, Ya-Ping Huang, Yang Mi, Yan-Ting Pei, Qi Zou, and Song Wang. Video sketch: A middle-level representation for action recognition. *Applied Intelligence*, 51(4):2589–2608, 2021b.

Xinru Zhang, Chenghao Liu, Ni Ou, Xiangzhu Zeng, Zhizheng Zhuo, Yunyun Duan, Xiaoliang Xiong, Yizhou Yu, Zhiwen Liu, Yaou Liu, et al. Carvemix: a simple data augmentation method for brain lesion segmentation. *NeuroImage*, 271:120041, 2023.

XYZ Zhang et al. A two-stream hybrid cnn-transformer network for skeleton-based human interaction recognition. *arXiv preprint arXiv:2105.02087*, 2021c.

Yi Zhang. Mest: An action recognition network with motion encoder and spatio-temporal module. *Sensors*, 22(17):6595, 2022.

Yun Zhang, Ling Wang, Xinqiao Wang, Chengyun Zhang, Jiamin Ge, Jing Tang, An Su, and Hongliang Duan. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Organic Chemistry Frontiers*, 8 (7):1415–1423, 2021d.

Xu Zheng, Tejo Chalasani, Koustav Ghosal, Sebastian Lutz, and Aljosa Smolic. Stada: Style transfer as data augmentation. *arXiv preprint arXiv:1909.01056*, 2019.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017a.

Xinyi Zhou, Wei Gong, WenLong Fu, and Fengtong Du. Application of deep learning in object detection. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 631–634. IEEE, 2017b.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Linchao Zhu, Du Tran, Laura Sevilla-Lara, Yi Yang, Matt Feiszli, and Heng Wang. Faster recurrent networks for efficient video classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13098–13105, 2020a.

Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 349–360. Springer, 2018.

Y Zhu, X Li, C Liu, M Zolfaghari, Y Xiong, C Wu, Z Zhang, J Tighe, R Manmatha, and M Li. A comprehensive study of deep video action recognition. arxiv 2020. *arXiv preprint arXiv:2012.06567*.

Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020b.

Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.

Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018.

Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *European Conference on Computer Vision*, pages 566–583. Springer, 2020.