Vision-Assisted mmWave Beam Management for Next-Generation Wireless Systems: Concepts, Solutions and Open Challenges

Kan Zheng, *Senior Member, IEEE*, Haojun Yang, *Member, IEEE*, Ziqiang Ying, Pengshuo Wang, and Lajos Hanzo, *Life Fellow, IEEE*

***Abstract*—Beamforming techniques have been widely used in the millimeter wave (mmWave) bands to mitigate the path loss of mmWave radio links as the narrow straight beams by directionally concentrating the signal energy. However, traditional mmWave beam management algorithms usually require excessive channel state information overhead, leading to extremely high computational and communication costs. This hinders the widespread deployment of mmWave communications. By contrast, the revolutionary vision-assisted beam management system concept employed at base stations (BSs) can select the optimal beam for the target user equipment (UE) based on its location information determined by machine learning (ML) algorithms applied to visual data, without requiring channel information. In this paper, we present a comprehensive framework for a vision-assisted mmWave beam management system, its typical deployment scenarios as well as the specifics of the framework. Then, some of the challenges faced by this system and their efficient solutions are discussed from the perspective of ML. Next, a new simulation platform is conceived to provide both visual and wireless data for model validation and performance evaluation. Our simulation results indicate that the vision-assisted beam management is indeed attractive for next-generation wireless systems.**

***Index Terms*—Millimeter wave (mmWave), Beamforming, Machine learning (ML), Next-generation wireless systems.**

I. Introduction

BEAMFORMING-aided directional transmission plays a critical role in improving the spatial spectrum efficiency. Due to the expected wide deployment of millimeter wave (mmWave) communications, beamforming techniques are receiving much attention in the context of multiple-input and multiple-output (MIMO) systems designed for the mmWave frequency bands [1]. However, given a large number of antennas, tracking the movement of multiple concurrent user equipments (UEs) dramatically increases the complexity, overhead and latency of signal processing at the base stations (BSs) using mmWave massive MIMO schemes [2]. These problems may be exacerbated for a high number of antennas even in line of sight (LoS) channels. In order to overcome these challenges, computer vision (CV)-aided machine learning (ML) algorithms may be harnessed as promising solutions for beamforming. Motivated by the spatial sparsity of mmWave wireless channels exhibiting predominant LoS characteristics, mmWave beams pointing to the target UEs can be efficiently selected and adapted according to the location information of UEs derived by ML algorithms [3].

In order to implement a vision-assisted beam management system, several technical challenges have to be overcome. The traditional vision-based object tracking algorithms, such as the sparse representation and correlation filtering, have difficulty in accurately locating high-mobility UEs in real-time [4]. Furthermore, in complex environments, tracking multiple UEs in the face of blockage and uneven light, the localization accuracy of UEs tends to degrade significantly. As a result, the traditional CV-related ML algorithms cannot satisfy the high location accuracy required by mmWave communications. Finally, gathering abundant labelled data from real-world environments including both visual data and wireless signals to train ML models is still challenging.

Nevertheless, vision-assisted beam management methods that predict the optimal mmWave beams have been investigated in the last few years. A framework used for dataset generation was also proposed for cooperatively exploiting both visual and wireless data [5]. However, these methods may still be plagued by a number of issues:

- The robustness of the existing ML models for vision-assisted beam management has to be improved [6]. When the classical image classification models designed for prediction are used for mmWave link blockage prediction and beam prediction, the target accuracy cannot be always satisfied, for example due to the over-fitting issues. In this context, a preliminary study was conducted in [7] by relying on a simple dataset, for investigating vision-assisted beam management in multi-user scenarios.
- The scalability of the methods is not guaranteed in practical scenarios. For example, the existing methods do not obey the so-called modular design principles, making

Kan Zheng is with the College of Electrical Engineering and Computer Sciences, Ningbo University, Ningbo, 315211, China (E-mail: zhengkan@nbu.edu.cn).

Haojun Yang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (E-mail: yanghaojun.yhj@gmail.com).

Ziqiang Ying and Pengshuo Wang are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, 100876, China (E-mail: yingzq0116@163.com, wshuo@bupt.edu.cn).

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (E-mail: lh@ecs.soton.ac.uk).

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62201301. Lajos Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/W016605/1 and EP/X01228X/1 as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028).

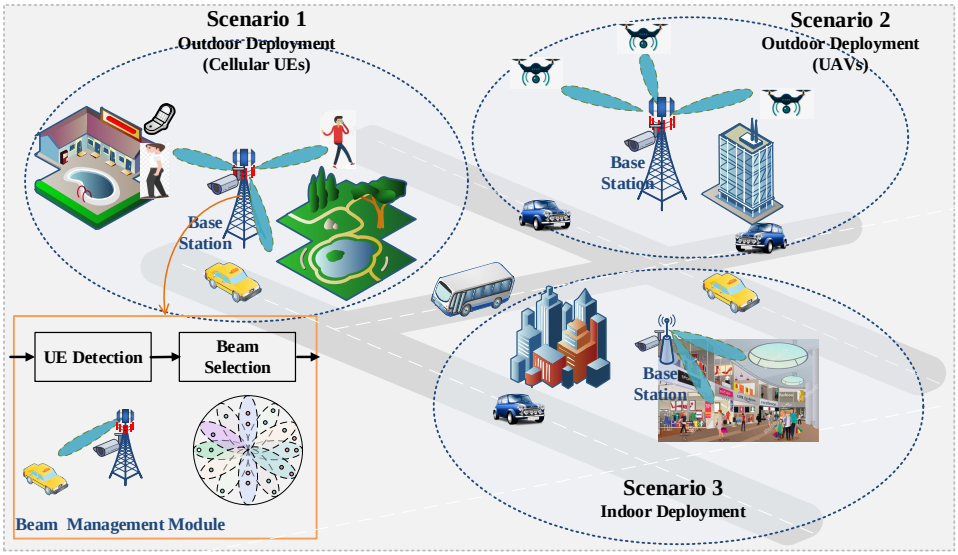


Fig. 1. Illustration of vision-assisted beam management system.

them difficult to upgrade flexibly or to modify them simply when for example a crucial parameter, like the size of the beamforming codebook, changes.

- The implementation issues pertaining to the complexity of computation and overhead costs have to be addressed before the wide deployment of vision-assisted beam management becomes a reality. For example, an exploratory strategy was proposed for reducing the overhead associated with beam selection, where information from localization and vision sensors is integrated [8].

Indeed, there is a paucity of literature addressing the associated challenges of vision-assisted mmWave beam management techniques. The scope of this article is thus to study the interplay of ML-based computer vision and beam management in mmWave systems. Specifically, the main contributions of this article can be summarized as follows.

1) We first present a comprehensive framework for a vision-assisted mmWave beam management system, including the typical deployment scenarios as well as a pair of major concerns, namely the user equipment detection and beam selection.
2) Then, three main technical challenges and their efficient solutions are discussed from the perspective of ML. In particular, we study the salient issues of lightweight compression, the deleterious effects of inadequately labeled data, as well as the associated robustness aspects.
3) Next, we portray the development of our own simulation platform to provide both visual and wireless data for model validation and performance evaluation. This unified platform is universally applicable in terms of producing data for those scenarios where the wireless characteristics vary tremendously. Our simulation results also show that vision-assisted beam management is indeed attractive for next-generation wireless systems.
4) Lastly, the related open topics are discussed from a practical perspective in order to guide future research.

The article is organized as follows. A detailed description of vision-assisted mmWave beam management systems is provided in Section II. Next, we discuss some ML-related challenges and solutions conceived for mmWave beam management in Section III. We then present our performance results and discuss some potential open topics. Finally, our conclusions are given in Section VI.

II. Holistic Framework of Vision-Assisted Beam Management System

A. Typical Deployment Scenarios of Vision-Assisted Beam Management Systems

The next-generation wireless systems are expected to operate in multiple bands, including the sub-6 GHz and mmWave bands. In general, the signal propagation of the sub-6 GHz bands is more resilient to blockages, thereby the sub-6 GHz bands are used for the services that require low or medium data rates. By contrast, as a benefit of their abundance of spectral resources, the mmWave bands are expected to support multi-Gigabit services. In order to take full advantage of their benefits, a dual-band system in which the BS and UEs use both the sub-6 GHz and mmWave transceivers is considered in this article. The vision-assisted beam management may be enabled only when the LoS condition is met, which may also be combined with sub-6 GHz systems [3]. The rich bandwidth potential of mmWave communications can be used both for the backhauls and for the user access links under a variety of potential deployment scenarios.Thus, we mainly focus attention on those scenarios, where the vision-assisted beam management can be harmoniously integrated. Some of them are illustrated in Fig. 1, and are discussed in more detail below.

1) Scenario 1 – Outdoor Deployment (Cellular UEs): When the wireless channels under outdoor environments are spatially sparse, i.e., dominated by LoS propagation, vision-assisted beam management can be indeed conveniently

adopted at the BSs. Then, all the cellular UEs can be served by BSs on the mmWave band.

2) Scenario 2 – Outdoor Deployment (UAVs): At the time of writing, mmWave communications are widely used for unmanned aerial vehicle (UAV) communications. The camera deployed at BSs can also readily capture the video of the UAV flying by without obstruction. Therefore, it is eminently suitable for vision-assisted beam management in this scenario.

3) Scenario 3 – Indoor Deployment: In order to significantly increase the system capacity in high-density indoor environments, cameras installed at the BSs are capable of capturing images of nearby UEs. However, there may be lots of objects, which increases the recognition complexity of the CV algorithms.

B. Description of Vision-Assisted Beam Management System

As shown in Fig. 1, a BS equipped with a high-definition camera first captures the video scenes. Then, the ML-based vision model embedded in the BS is activated for localizing and tracking the target UEs. By collaboratively utilizing the image/video information, the beam management module finally selects the optimal beams for the target UEs among a pre-defined beam pattern codebook.

The vision-assisted beam management is generally divided into two steps, namely the UE detection, and the ensuing beam selection for the target UEs. More explicitly, the former determines whether any target UEs exist in the view of the camera, while the latter is responsible for providing both the location information and the optimal beams for the target UEs.

1) UE Detection: In the traditional ML-based object detection models, each frame of the video stream is processed to generate the object locations as the output, which is usually time-consuming. However, the target UEs are not captured by the camera all the time, and they are not always in their active communication status. To this end, it is necessary to detect the existence of active UEs before beam selection. In the proposed framework, a ML-based binary classification model can be used for determining whether active UEs exist at the current moment. For the sake of illustration, the active state of the k-th UE is defined as "1", while the inactive state as "0" . Then, based on the captured image, the active/passive state of the k-th UE can be predicted by

$$S_k = F_{\mathcal{P}_1}\left(\mathsf{Image}, k\right), \tag{1}$$

where $F_{\mathcal{P}_1}(\cdot)$ is the ML-based binary classification model that has to be investigated, and $\mathcal{P}_1$ is the parameter set of the model. Additionally, to strike an attractive tradeoff between the complexity and accuracy, the UE-related information including the sub-6GHz channel state information and network signaling might be taken into account.

2) Beam Selection for Target UEs: The goal of beam selection is to find the optimal beam from the codebook for maximizing the signal-to-noise ratio (SNR). The traditional beam management schemes generally require the channel state information (CSI) to be obtained by channel estimation, which requires substantial overhead. Instead, a vision-assisted method requiring no CSI knowledge is conceived for solving the beam selection problem. Firstly, the position of target UEs can be determined by a ML-based object detection model from the images captured by the camera. Then, the angles of the target UEs are estimated by exploiting the location information. Finally, the optimal beam index is selected by maximizing the SNR, albeit other metrics may also be used. The complete procedure is as follows.

*a) **Object Detection**:* Given the presence of some target UEs, each frame of the video stream can be processed to locate the target UEs by ML-based object detectors. In general, the family of ML-based object detectors may be divided into two types, namely the single-stage detectors, such as the so-called You Only Look Once type models [9], and the two-stage detectors, such as region based convolutional neural network (R-CNN) related models. The two-stage detector first adopts a 'region proposal network' for generating the region of interest (RoI), and then utilizes classification models for determining the category of region. In contrast to the two-stage detector, the single-stage one directly predicts the category of each feature map without first generating RoI. Hence, the two-stage detector typically attains higher detection accuracy, while the single-stage detector has higher detection speed. In our proposed framework, either of them may be chosen flexibly according to the specific requirements of different application scenarios.

*b) **Angle Prediction**:* For the beam management, the angle information of the UEs' physical location within the geographical coverage of the BS is required for selecting the optimal beam in terms of the real physical *world coordinate*. However, the outputs of ML-based object detection models are the UEs' location in the image captured by the camera, i.e., the location in the *pixel coordinate*. Thus, it is paramount to establish the mapping relationship between these two locations in the cases of vision-assisted beam management applications.

*c) **Beam Selection**:* Given the predicted angle, the beam selection can provide the index of the optimal beam. Let $\mathbf{w}_k$ denote the beamforming vector of the k-th UE. Then the optimal beam can be predicted as follows:

$$\mathbf{w}_k = G_{\mathcal{P}_2}\left(\mathsf{Angle}, \mathsf{Codebook}, k\right), \tag{2}$$

where $G_{\mathcal{P}_2}(\cdot)$ and $\mathcal{P}_2$ are the prediction model and parameter set, respectively. For instance, upon considering the simple case of a uniform linear array in the 2D space, the codebook is composed of Q beams having an identical angular separation of π/Q. Therefore, the task of the beam selection is simplified to estimating the range that the predicted angle falls into.

III. Challenges for Vision-Assisted Beam Management System

A. Lightweight Compression for Prediction Model

The limited computing and storage capabilities of embedded systems make the real-time implementation of the ML-based models in mmWave communication systems challenging. Again, the YOLO object detector is applicable to localize the target UE. Even though YOLO is faster than other detectors, it still contains too many convolutional layers. For example, the backbone network in Version 3 of YOLO

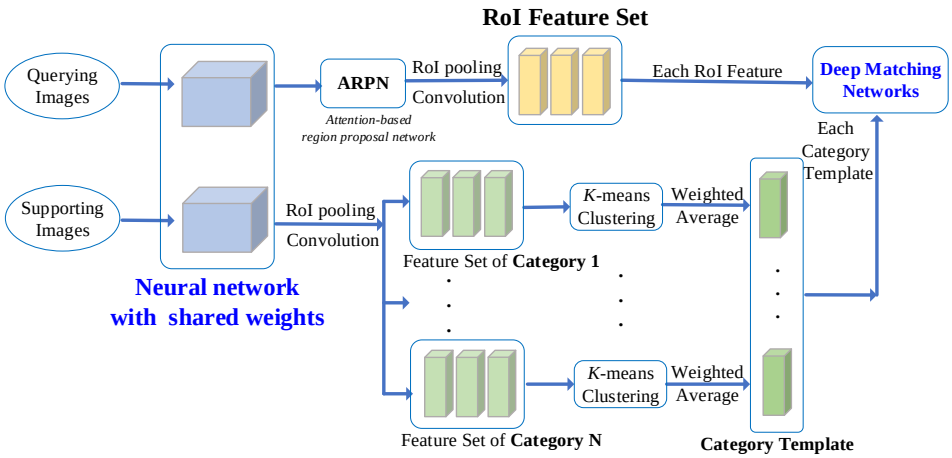


Fig. 2. Illustration of N-way K-shot ML models based on metric learning for object detection.

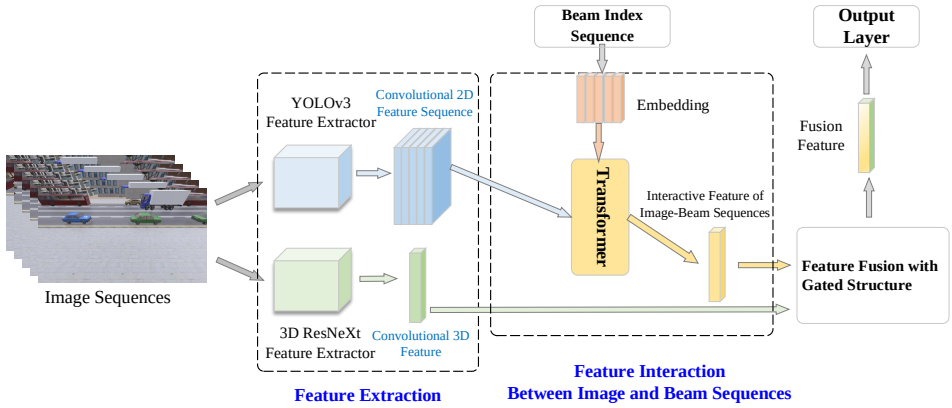


Fig. 3. Illustration of a vision-assisted beam management system based on image sequences, where the structures of the YOLOv3 and Transformer are the classical ones, and the 3D ResNeXt-101 is shown in [10].

(called 'YOLOv3' [9]) comprises 53 convolutional layers, and the channels in each of these convolutional layers are typically quite large, namely up to 1024 channels. Hence, the model size and computation complexity of YOLO become the barriers to its time-critical applications such as mmWave beam management. Therefore, it is essential to compress the model volume for increasing its prediction speed, while guaranteeing its accuracy.

One of the most common methods of model compression is network pruning [11]. In this method, sophisticated rules can be applied to neural networks so that the relatively insignificant weights or branches are removed, thereby reducing the number of model parameters and increasing the inference speed. According to the granularity of pruning objects, the typical network pruning schemes can be divided into weight pruning and structured pruning techniques. The former compresses those relatively insignificant weights in the networks. This technique has a high degree of flexibility, but modest inference speed acceleration. For the latter one, the coarser-grained convolution kernels, channels, and layers might be removed, resulting in both a higher compression ratio and faster inference speed.

Since YOLO contains a large number of convolutional layers and hundreds or thousands of channels, a structured pruning method is preferred for obtaining a satisfactory compression effect. In particular, YOLO can be pruned at both *the channel and layer levels*. Channel pruning provides the compression of the model width, while the layer pruning reduces the depth of the models. With the help of network pruning, the volume of YOLO model can be substantially reduced, hence its prediction speed is significantly improved.

B. Efficiency Improvement of Prediction Model Having Inadequately Labeled Data

Due to a large number of parameters in the ML-based models, a dataset having a huge number of labeled data is required for training the models. However, there might be insufficient labeled data to fully fit the ML models in practical vision-assisted mmWave communication systems. First of all, gathering visual data (such as RGB images) and wireless data (such as channel responses) requires completely different equipment and devices. Furthermore, realistic physical test scenarios have to be constructed, relying on practical equipment placing and data synchronization. Additionally, a long test period is needed in order to collect enough data. As a result, the data collection process itself is complex and time-consuming. Finally, the

visual datasets collected have to label the bounding boxes for all UEs in the images. Thus, for practical applications, another challenge is how to achieve excellent prediction accuracy in the beam management module, when the dataset is small or moderate.

The output layer, which is used to map the feature vector to the required classification space, is typically a fully connected layer or a 1×1-convolutional layer in both the single-stage and two-stage object detectors. In general, the output layer parameters of object detectors are randomly initialized and iteratively updated thereafter, when a new dataset is adopted in the training. However, there are also a number of other parameters in the models that have to be fine-tuned. As a result, having inadequate data may cause over-fitting during the learning process, gravely affecting the localization of objects. Localization performance degradation may lead to the spurious angle prediction for beam selection algorithms.

Therefore, in order to improve the modelling process in the face of inadequately labeled data, we propose to use a ML scheme relying on the ***'metric learning***[1]***'*** technique of [12] for accurately localizing the UEs. Fig. 2 presents a N-way K-shot ML scheme conceived for object detection based on metric learning. In this scheme, a RoI set is generated for the querying images based on region proposal networks. For the supporting images, the features of all categories are generated according to the labeled frames. Our proposed scheme calculates the similarity between each predicted RoI feature and the corresponding feature template. Theoretically, the higher the similarity score, the higher the prediction accuracy becomes for the bounding box associated with the RoI.

C. Robustness and Applicability for Prediction Model

Only adopting low-complexity single frame image based methods cannot cope well with multi-user scenarios, especially in cluttered environments. In the case of completely invisible UEs, the BSs cannot identify them, because they may be totally obscured when simply analyzing a single image frame at a time. Explicitly, single image contains only the location information and environmental information about the UEs seen at the time, but cannot provide extra information concerning the movement of the target UEs or the changing camera-view of the surrounding environment. Hence the single-frame processing loses sight of the spatial and temporal correlation of moving objects. Therefore, how to exploit the image sequences in the video data to improve the performance of the object detectors and beam management remains a challenge.

To enhance the robustness and applicability, the vision-assisted beam management may process the image sequences for a total of N consecutive video frames, i.e., not only the current frame but also those from $N-1$ previous frames. Compared to the schemes based on the current individual video frame, the improved schemes using a sequence of video frames can capture both the spatial coherence of each video frame and its temporal inter-frame correlation.

Fig. 3 shows the overall framework of a vision-assisted beam management system based on image sequences [13]. The framework primarily consists of three main steps. In the first step, we extract specific features of the image sequences. In the lower branch of Fig. 3, the 3D convolution is applied for extracting the features containing both spatial and temporal contents. In the upper branch of Fig. 3, the 2D convolution is used for processing each image separately. Then, the interactions among the image-beam sequence features take place. The Transformer scheme of Fig. 3 having several encoder layers is used for interactive sequence modeling of the features in the beam index sequence and the image sequence features obtained by 2D convolution. The final step is to design a suitable output layer according to the specific beam management tasks so as to select the optimal beam.

IV. Simulation Methodology and Evaluation

At the current state-of-the-art, it is quite difficult to collect and label both the visual and wireless data in real-time. Hence, we resort to simulations for generating labeled data for training and testing. Fig. 4 shows our simulation platform conceived for vision-assisted mmWave beam management. As this stage, only an outdoor scenario is used for validating the models in the platform. An open source framework is proposed to speed up the implementation of other scenarios. As a result, it is advantageous to revise the details of the scenario when generating wireless data, such as the number and orientation of rays, channels, user positions, etc. Furthermore, diverse UEs are involved, as well as other entities, such as trees, bushes, sidewalks, benches and buildings. Specifically, our own-developed and open-source platform is based on MATLAB software and only requires a text file for defining a scenario. During the phase of initialization, a series of visual and wireless sequences are created for modelling real-world physical environments. To create visual and wireless datasets, all sequences are respectively processed by the animation modeling software[2] and the ray-tracing software in the second step. Finally, the datasets can be used for evaluating and validating the performance of ML-based models for beam management.

A. Initialization

In the initialization phase, the types and attributes of entities are described in intricate detail. Using unified definitions is an efficient and compatible way of ensuring the appropriate relationship between the visual and wireless data generation. In particular, the scenario definition includes the system parameters, antenna arrays, BSs, reflectors, and mobile users. Each of them contains the following information, i.e.,

- **System parameters:** This includes parameters used to describe how the platform works. For example, the total

[1]In general, many approaches in ML require a measure of distance among data points. Typically, with the aid of priori domain knowledge, some standard distance metrics are adopted, such as Euclidean, Cosine, etc. Nevertheless, it is difficult to design metrics that are well-suited to the particular data and task of interest. Therefore, ***'metric learning'*** technique is investigated to automatically construct task-specific distance metrics from weakly supervised data, which is more beneficial for the case of inadequately labeled data.

[2]Normally, the animation modeling software is designed for creating complex 3D objects, rendering them to images, and making animation from frames.

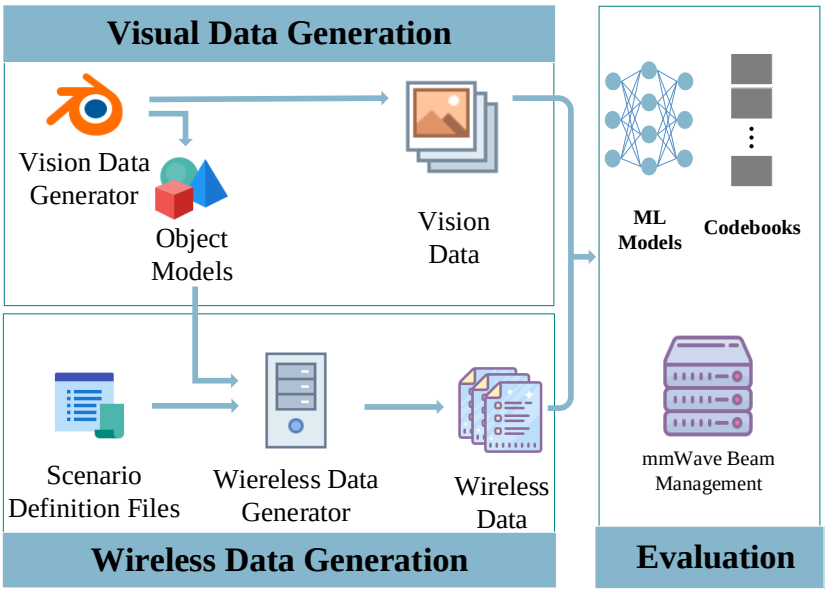


Fig. 4. Simulation platform for vision-assisted beam management system.

number of frames for simulations, the average number of video frames calculated per second in simulations, the maximum number of mmWave reflections calculated in the ray tracing process, and the size of beam codebook, etc.

- **Antenna arrays:** The key parameters of the antennas such as the size of antenna arrays and the antenna spacing are defined.
- **Base stations:** The location of BSs, the configuration of the camera deployed at BSs, the antenna arrays used by BSs and diverse other parameters are described in detail.
- **Reflectors:** The position, shape and material of reflectors are given.
- **User equipments:** Similar to the reflectors, we have to define the parameters of UEs such as the location and appearance. Additionally, the UE antenna arrays, the UE motion trajectories and other parameters are specified as well.

It is noted that both the visual and wireless simulations require the aforementioned information. Based on the information defined in a given scenario, the visual and wireless simulation processes have to be synchronized so that the data generated tally correctly.

B. Data Generation

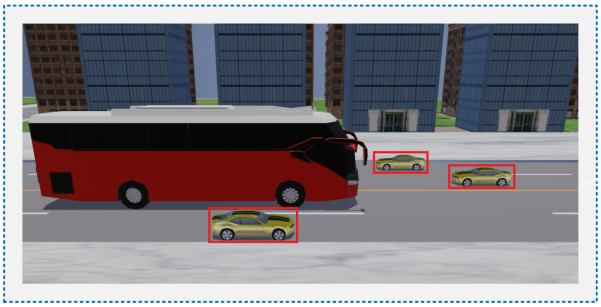
The data generation includes both visual and wireless data. Although different simulation environments and processes are used for generating these two types of data, there still exists a corresponding relationship between them for ensuring that the data produced conforms to the scenario definition. Here, we first introduce the process of generating both types of data, and then we describe how to synchronize and merge these data.

1) Visual Data Generation: The visual data is generated by some special animation modeling software, such as Blender [14], which facilitates the construction of 3D object models. Hence, the first step in generating the visual data is to create a 3D model of the reflectors and users by the animation modeling software. Then, we have to assign textures or materials to the objects in the scenario, in order to make them more realistic. Note that the material mentioned here differs from that in the scenario definition. The former only determines the visual effect of the generated image, while the latter determines the propagation of electromagnetic waves. Next, the cameras have to be deployed correctly at the BSs. The second step is to define the movement animation of users. In general, the animation consists of a sequence of images. There are some frames referred as the key frames, and the position and shape of the 3D model in the other frames can be determined by interpolation between a pair of consecutive key frames. In the scenario defined, the objects are regarded as rigid bodies, and each frame only contains the position change users. Finally, the animation generated is the last step exported from the animation modeling software.

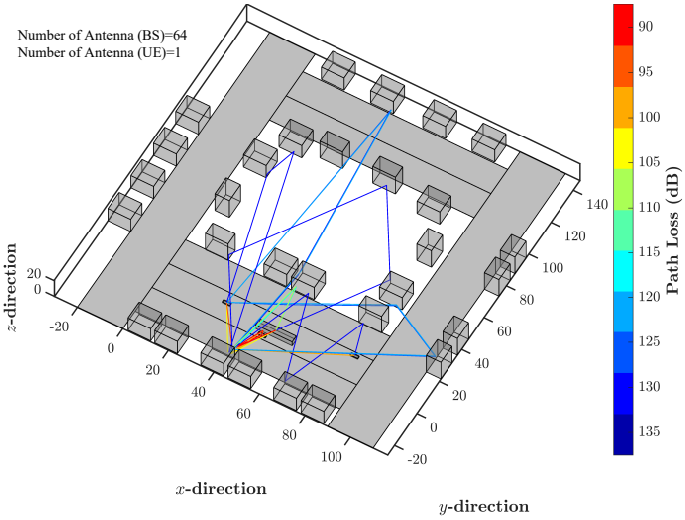
2) Wireless Data Generation: The wireless data is generated by the software that supports ray-tracing technology [15], e.g., MATLAB. Due to the challenge of generating complex 3D objects in MATLAB, the 3D model of the reflectors and users must be obtained by loading external data. Then, the transmitter and receiver are correctly positioned, i.e., co-located with either the BSs or the other UEs that might move at a given speed and in a certain direction. Next, we calculate the propagation-related information, such as the signal power, delay, angle of departure and angle of arrival, using ray-tracing technology. Likewise, according to the geometric channel model, the wireless channels in the current scenario are constructed using the above propagation information. Furthermore, the codebook indices corresponding to the optimal beam are calculated, which is crucial for the wireless data.

3) Data Synchronization: As seen from the above data generation phase, both the object model and UE motion

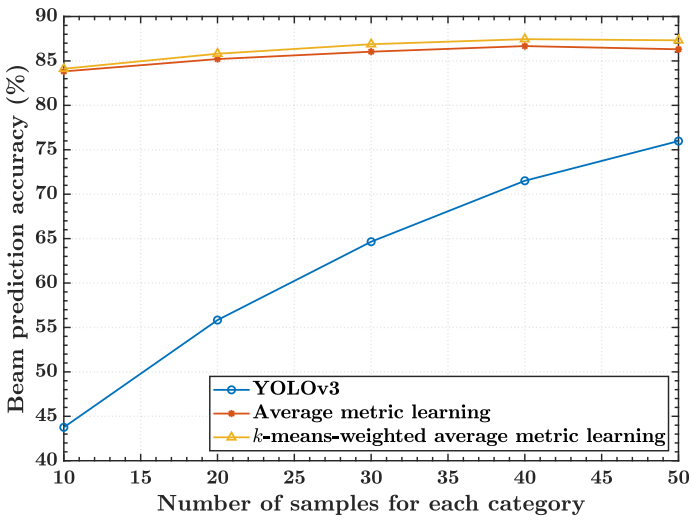
(a) Illustration of an example of mmWave communication scenario.



(b) An example for single frame image with bounding box labelled by object detection.



(c) An example for ray tracing at one frame.



(d) Performances of beam prediction.

Fig. 5. Performance evaluation results.

should be consistent across the pair of generation processes. Specifically, the following methods are used in our platform.

- Consistency of object model: With the aid of the animation modeling software, we create 3D models of the objects and export them in STL format[3]. Then, we import the required 3D model into MATLAB software with ray-tracing to make sure that the object models are consistent.
- Consistency of UE movement: The concept of frames is introduced into the ray-tracing process. Hence, the UEs remain in the same positions for these two generation processes, which ensures that the UE movements are consistent.

C. Evaluation and Validation

Fig. 5(a) shows an example of the mmWave communication scenario in an urban environment, where a BS communicates with three vehicular UEs in cars. Moreover, two types of buildings built from different materials are involved in this scenario. In general, the reflection properties of objects are strongly influenced by their materials. To accurately model real-world environments, we set the materials of the dark-coloured and light-colored buildings to concrete and brick, while the materials of all vehicles are assumed to be metal. We then characterise the performance of the platform based on this pre-defined scenario.

1) Dataset Validation from Visual and Wireless Aspects: In the proposed framework of vision-assisted beam management, object detection plays a crucial role in both the existence detection and beam selection tasks. Thus, Fig. 5(b) presents the validation results for datasets from the visual perspective. Explicitly, it shows a single image frame labelled by the bounding boxes of three moving UEs in cars. The accuracy of the labeling results demonstrates that the visual datasets generated accurately characterise the movement of objects in each frame, and that the proposed vision-assisted beam management framework can also effectively track objects in real time.

On the other hand, Fig. 5(c) illustrates the signal power of the randomly generated rays between the BS and vehicular UEs, which comes from the wireless datasets. The larger the distance between rays, the lower the received power, which confirms the trends of the generated wireless datasets. Additionally, due to the mobility of various objects, wireless

[3]The STL format is a universal format for displaying 3D models, which is widely supported by related animation modeling software.

TABLE I
SIMULATION RESULTS FOR ML MODELS WITH LIGHTWEIGHT COMPRESSION

Model Type	Classification Performance		Compression Performance		
	mAP[1] Score	Beam Prediction Accuracy	The Number of Parameters	Model Size	FPS[2]
No-Pruning Model	85.12	90.34%	61.52 M	236.52 MB	39
Channel Pruning Only Model	84.68	90.15%	12.64 M	48.32 MB	77
Channel and Layer Pruning Model	84.42	90.06%	11.17 M	42.71 MB	91

[1] Mean average precision
[2] Frames per second

datasets can occasionally contain zero data. For example, whenever the car farthest from the BS runs into the shadow of a bus, no rays are detected for this frame, resulting in beam tracking outage.

2) Results for Inadequately Labeled Data: Fig. 5(d) shows the beam prediction accuracy of the improved models using metric learning in the case of inadequately labeled data. For each category, the metric learning normally requires only one feature template. Nevertheless, there may be K samples in each category, thereby having K feature vectors. Therefore, the K feature vectors should be combined to produce a representative category template. Three schemes are considered for comparison, i.e., Version 3 of YOLO, the average metric learning, and the k-means-weighted average metric learning. Specifically, the average metric learning combines all feature vectors using the arithmetic mean method over K samples. To overcome the homogenization of arithmetic mean, the k-means-weighted average metric learning is also studied, in which K samples are first classified by the k-means method, and the cluster features obtained are then combined by the weighted averaging method.

It is clearly shown that both metric learning schemes perform better than YOLOv3, achieving an accuracy of about 84.12% with only 10-shot learning. Furthermore, the k-means-weighted average metric learning slightly outperforms the average metric learning. In conclusion, the metric learning schemes are more efficient than YOLOv3 when data are inadequately labelled.

3) Results for Lightweight Compression: The performances of the improved ML-based models having different lightweight compression are also evaluated. Prior to discussing the simulation results, we briefly highlight the performance metric, i.e., mean average precision (mAP) score. As a derivative of the average precision (AP), mAP is the average of AP score, while the AP score generally is obtained by calculating the area under the precision-recall (PR) curve. To summarize, the AP score is calculated for each category, then averaged to determine the final mAP score.

Table I presents the simulation results for the cases of no-pruning, channel pruning only, as well as channel-pruning and layer-pruning. As illustrated in Table I, the classification performance of both pruning models degrades compared to the no-pruning model. However, the accuracy erosion is modest for two pruning models. For instance, with regard to the channel and layer pruning model, the mAP performance and beam prediction accuracy only deteriorates by about 0.8% and 0.3%, respectively. By contrast, the pruning operation results in a significant model size reduction and an acceleration of the inference speed. The number of parameters and the model size are reduced by about 82%, which is more beneficial for the practical deployment of latency-sensitive applications.

V. OPEN DISCUSSION

A. Combination with Hierarchical Beam Search

Usually, the explicit training required for finding the best beam directions in the angular domain is indispensable. In contrast to the classical exhaustive search based training, hierarchical training has been proposed as a promising technique of reducing both the complexity and the overhead. However, a trade-off must be struck between the phase shift resolution of training and the complexity imposed. For example, when a low phase shift is chosen for the first stage of training, the beam direction can be selected more accurately. However, this imposes higher feedback delay and higher overhead, or vice versa. To strike a compelling trade-off, a vision-assisted beam management scheme can be used as the first stage of training, because it does not rely on UE feedback for beam selection. Subsequently, the accuracy of beam search can be further improved through a fine-tuning of CSI-based beam management along with a lower phase shift in the following training stage.

B. Uplink and Downlink Beam Matching

As a result of the propagation differences between the uplink and downlink, especially for frequency division duplex (FDD) systems, the downlink beam selection based on the uplink channel estimation operation usually requires calibration to improve accuracy. On the other hand, the location of the user can be accurately determined by vision-assisted beam management regardless of the frequency band. Therefore, how to use this information to support beam matching on both the uplink and downlink becomes a very interesting topic.

C. Dual-Band Communications with Sub-6 GHz

Recently, the dual-band communication mode including mmWave and sub-6 GHz communications is becoming increasingly popular. Therefore, another open challenge is how to exploit the extra information at sub-6 GHz so as to enhance the mmWave beamforming performance. Intuitively, the

proposed vision-assisted mmWave communications depends on having LoS propagation for its accurate operation, and it is vulnerable to blockage. For instance, when multiple UEs are captured by the camera without any additional details, vision-assisted beamforming may falter. On the other hand, sub-6 GHz communications generally works well for both non-line of sight (NLoS) and LoS channels, and it is capable of providing the related control information, including CSI and other user-specific information. This information can assist in the detection of active UEs and multi-user discrimination when using vision-assisted mmWave communications. Additionally, for further reducing the complexity of exploiting sub-6 GHz communications, the above-mentioned hierarchical beam search technique can be used for sub-6 GHz to provide prompt user-specific information.

D. Multi-Cell Beam Management

The coverage distance of mmWave communications is typically small, and UEs often appear at the cell edge. The beam selection of cell-edge UEs can be handled more accurately by adopting vision-assisted beam management. Specifically, the videos obtained by the cameras of multiple adjacent BSs can be processed jointly. Due to the fact that the same UE is captured in multiple images at the same time, its position can be more accurately determined using ML algorithms as well as the beam direction. By aligning the beams of two adjacent BSs for the target UE, a more reliable communication connection can be achieved. The issues associated with channel feedback overhead can be avoided by such a vision-aided multi-cell beam management.

VI. Conclusions

Vision-assisted beam management is paving the way for improved mmWave communications by relying on machine learning models of analyzing visual data. This enables us to tackle several important challenges of ML-based model implementation for mmWave beamforming. In particular, sophisticated network pruning has been used to compress the models for reducing the complexity. Additionally, a model based on metric learning has been shown to be an effective option for dealing with the problem of inadequately labeled data in practical applications. A ML model based on image sequences has also been conceived for multi-user scenarios and to mitigate the blockage problems. Then, an animation modeling software and ray-tracing software were used for successfully building a new simulation platform to generate various labeled visual and wireless data for performance evaluation and model validation. Our simulation results show that ML-based models work well with vision-assisted mmWave beam management schemes. Furthermore, some open challenges are presented for the guide of future research works. Additionally, dual-function radar communication (DFRC) systems may be capable of simultaneously performing wireless communications and remote sensing, when they become available but have a huge complexity. The alluring topic of combining these technologies is also interesting for future research. Suffice to say that the vision-based system investigated in this treatise only requires a low-cost camera and object-recognition software.

References

[1] E. Bjornson, L. Van der Perre, S. Buzzi, and E. G. Larsson, "Massive MIMO in sub-6 GHz and mmWave: Physical, practical, and use-case differences," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 100–108, Apr. 2019.

[2] X. Liu, J. Yu, H. Qi, J. Yang, W. Rong, X. Zhang, and Y. Gao, "Learning to predict the mobility of users in mobile mmWave networks," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 124–131, Feb. 2020.

[3] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10 193–10 208, Oct. 2021.

[4] P. Druzhkov and V. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 9–15, Jul. 2016.

[5] M. Alrabeiah, J. Booth, A. Hredzak, and A. Alkhateeb, "ViWi vision-aided mmWave beam tracking: Dataset, task, and baseline solutions," *arXiv:2002.02445v3*, pp. 1–7, Feb. 2020.

[6] Y. Tian and C. Wang, "Vision-aided beam tracking: Explore the proper use of camera images with deep learning," in *Proc. IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, Norman, OK, USA, Sep. 2021, pp. 1–5.

[7] H. Ahn, I. Orikumhi, J. Kang, H. Park, H. Jwa, J. Na, and S. Kim, "Machine learning-based vision-aided beam selection for mmWave multiuser MISO system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1263–1267, Jun. 2022.

[8] G. Reus-Muns, B. Salehi, D. Roy, T. Jian, Z. Wang, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on visual and location data for V2I mmWave beamforming," in *Proc. International Conference on Mobility, Sensing and Networking (MSN)*, Exeter, United Kingdom, Dec. 2021, pp. 559–566.

[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767v1*, pp. 1–6, Apr. 2018.

[10] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6546–6555.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015, pp. 448–456.

[12] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4012–4021.

[13] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4:36, pp. 1–47, Aug. 2020.

[14] "Blender," Blender Foundation. [Online]. Available: https://www.blender.org

[15] Q. Li, H. Shirani-Mehr, T. Balercia, A. Papathanassiou, G. Wu, S. Sun, M. K. Samimi, and T. S. Rappaport, "Validation of a geometry-based statistical mmWave channel model using ray-tracing simulation," in *Proc. IEEE 81st Vehicular Technology Conference (VTC Spring)*, Glasgow, UK, May 2015, pp. 1–5.

Kan Zheng (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), China, in 1996, 2000, and 2005, respectively. He is currently a full professor with Ningbo University, Ningbo, Zhejiang, China. He has rich experience in research and standardization of new emerging technologies. He has authored over 200 journal articles and conference papers in the field of wireless communications, vehicular networks, IoT, security, and so on. He holds editorial board positions with several journals and also served in the organizing/TPC committees for conferences.

Haojun Yang (Member, IEEE) received the B.S. degree in communication engineering and the Ph.D. degree in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014 and 2020, respectively. He is currently a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. His research interests include ultra-reliable and low-latency communications, radio resource management and vehicular networks.

Haojun Yang Ziqiang Ying received the B.S. degree in electronic information science and technology and the M.S. degree in information and communication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2018 and 2021, respectively. His research interests include deep learning, millimeter wave communication system, vehicular networks and machine vision.

Pengshuo Wang received his B.E. degree in information engineering in 2020 and is now working towards his M.E. degree in information and communication engineering at Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests include machine learning for wireless communication.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.org/wiki/Lajos_Hanzo) (FIEEE'04) received his Master degree and Doctorate in 1976 and 1983, respectively from the Technical University (TU) of Budapest. He was also awarded the Doctor of Sciences (DSc) degree by the University of Southampton (2004) and Honorary Doctorates by the TU of Budapest (2009) and by the University of Edinburgh (2015). He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He has published 2000+ contributions at IEEE Xplore, 19 Wiley-IEEE Press books and has helped the fast-track career of 123 PhD students. Over 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is also a Fellow of the Royal Academy of Engineering (FREng), of the IET and of EURASIP. He is the recipient of the 2022 Eric Sumner Field Award.