A Hybrid Sparse-Semantic Image Classification Framework to Support Marine and Coastal Monitoring and Management

Ying Shi [a], Lan Song[b, c *], Luocheng Wu [d], Huanhuan Li [e]
[a]Department of Basic Courses, Suzhou City University, Suzhou, China
[b]Institute of Environmental Science (CML), Leiden University, The Netherlands
[c] College of Engineering, Eastern Institute of Technology, Ningbo, China
[d] Institute of Sound and Vibration Research, University of Southampton, Southampton, UK
[e] School of Engineering, University of Southampton, Southampton, UK
* Corresponding author: lansong@eitech.edu.cn

Abstract: Effective ocean and coastal management increasingly relies on timely, reliable, and interpretable information derived from large volumes of visual data collected by satellites, Unmanned Aerial Vehicles (UAVs), and shipborne sensing systems. Such data underpin a wide range of management and governance functions, including environmental monitoring, maritime surveillance, port operations, and coastal planning. However, the operational use of automated image classification in marine and coastal contexts remains constrained by complex visual conditions and the persistent scarcity of labelled data, limiting its integration into routine decision-support workflows. This study develops a hybrid sparse-semantic image classification framework designed to enhance the robustness and applicability of marine and coastal image understanding in management-oriented settings. The framework integrates structured local representations, which capture geometric and spatial characteristics of marine scenes through a sparsity-driven encoding strategy with locality, non-negativity, and semantic consistency constraints, with global semantic features learned via self-supervised Vision Transformers (ViTs). Specifically, semantic information extracted from the Classification (CLS) token of a pretrained self-Distillation with No Labels (DINO) model is fused with sparse local descriptors to form interpretable and discriminative image representations that remain effective under limited labelled data conditions. Experiments conducted on four public benchmark datasets and three maritime image datasets demonstrate consistent performance improvements over representative state-of-the-art approaches. Beyond classification accuracy, the proposed framework provides a transferable and expandable analytical tool to support marine environmental monitoring, maritime traffic surveillance, and coastal management practices by enhancing upstream image interpretation within monitoring and decision-support workflows in data-constrained marine and coastal contexts.

Keywords: Marine and coastal management; Image classification; Self-supervised learning; Maritime surveillance; Decision support

1. Introduction

Effective ocean and coastal management increasingly depends on timely, reliable, and actionable information to support planning, regulation, and decision-support workflows across complex marine and coastal systems. Advances in Earth observation technologies, Unmanned Aerial Vehicles (UAVs), autonomous platforms, and shipborne sensing systems have led to a rapid expansion in the availability of visual data covering marine environments, coastal zones, and maritime activities. These data streams play a central role in management-oriented

applications such as marine environmental monitoring (Galgani et al., 2024; Ghaban et al., 2025; Puskic et al., 2026), maritime traffic surveillance (Dong et al., 2024; Teixeira et al., 2025; Yang et al., 2024), port and terminal operations management (Abu Bakar et al., 2025; He et al., 2025; Yorulmaz and Susoy, 2025), habitat assessment and conservation planning (Agrillo et al., 2025; Wang et al., 2025; Xiao et al., 2025), and maritime safety and security governance (Adetunji et al., 2025; Wang et al., 2024). While the volume and diversity of marine and coastal imagery continue to grow, transforming these data into management-relevant knowledge remains a persistent challenge, particularly in large-scale and heterogeneous operational contexts.

Recent studies have increasingly addopted data-driven approaches to support maritime and coastal management. For instance, predictive modelling has been applied to forecast shipping-related emissions and identify key driving factors, thereby informing policy development and operational decision-making within the shipping sector (Xu et al., 2025c). In parallel, optimisation-based methods have been used to analyse trade-offs in infrastructure planning and regulatory design, including shore-to-ship electricity deployment and port competition strategies (Xu et al., 2025a). Additional research has investigated the low-carbon efficiency of ship-port systems (Zhang et al., 2025) and assessed the environmental impacts of fuel policies (J. Shi et al., 2025), highlighting the need for robust monitoring and evaluation frameworks to support evidence-based governance. Furthermore, large-scale initiatives such as the Maritime Silk Road underscore the increasing complexity of cross-regional carbon management and policy coordination (Xu et al., 2025b). Collectively, these studies reveal an increasing demand for robust, data-driven analytical tools capable of supporting decision-relevant processes in complex and data-constrained marine and coastal environments.

Image classification constitutes a foundational analytical component in many marine and coastal monitoring pipelines, enabling raw visual data to be translated into semantically meaningful categories that can inform downstream decision-support systems, regulatory oversight, and management interventions (Rangel-Buitrago et al., 2026; Vehmaa et al., 2024; Yang et al., 2021). However, compared with terrestrial environments, marine and maritime imagery is typically acquired under highly variable and adverse conditions, including complex illumination, strong reflections, low contrast, scale variability, background clutter, and partial occlusion. These characteristics substantially increase the difficulty of extracting reliable and discriminative image representations, often leading to degraded classification performance when methods are deployed in real-world marine and coastal settings.

In parallel, labelled marine imagery remains scarce, expensive to obtain, and unevenly distributed across application domains, posing a major constraint for the adoption of supervised data-driven approaches in routine management and governance processes.

From a methodological perspective, sparse coding-based image classification techniques have been widely explored due to their robustness, interpretability, and ability to capture local geometric and structural characteristics of visual scenes (Cao et al., 2024; Cheng et al., 2023; Shu, 2026). By representing visual patterns as sparse combinations of dictionary atoms, sparse coding can effectively capture local geometric and structural characteristics while suppressing

redundant information. When combined with handcrafted local descriptors such as the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), sparse coding has demonstrated resilience to scale and illumination variations, making it attractive for complex visual environments, including marine scenes. Nevertheless, such approaches mainly rely on low-level handcrafted features and have limited capability to capture high-level semantic information, which is increasingly important for distinguishing visually similar marine categories.

More recently, self-supervised deep learning models, particularly Vision Transformers (ViTs) (Dosovitskiy et al., 2021), trained using frameworks such as self-Distillation with No Labels (DINO) (Caron et al., 2021), have shown strong capability in learning transferable global semantic representations without requiring large quantities of labelled data. The Classification (CLS) token produced by a pretrained DINO model encodes holistic image-level semantics and has demonstrated strong generalisation across diverse visual recognition tasks. This property is especially appealing for marine and coastal applications, where annotated datasets are often limited. However, deep semantic features alone typically lack explicit modelling of local geometric structures and offer limited interpretability, which may reduce robustness and transparency in challenging visual conditions.

These observations highlight the need for image representation frameworks that can simultaneously preserve local geometric structure, capture global semantic meaning, and remain robust and interpretable under limited labelled data conditions. Such frameworks are particularly important for marine and coastal management, where analytical outputs must support operational decision making, risk assessment, and governance processes across diverse environmental and institutional contexts. Sparse coding and self-supervised ViT models offer complementary strengths: sparse coding provides structured and interpretable local representations, while ViTs deliver powerful global semantic understanding. Leveraging this complementarity, this study proposes a hybrid sparse-semantic image classification framework tailored to the analytical needs of marine and coastal monitoring and management.

Within the proposed framework, local SIFT descriptors are encoded using a Histogram intersection and Semantic information-based Locality Non-negative Laplacian Sparse Coding (HS-LNLSC) scheme to preserve neighbourhood similarity, non-negativity, and structural consistency. In parallel, global semantic representations are extracted from the CLS token of a pretrained DINO ViT. The two complementary feature types are fused at the feature level to construct discriminative image descriptors, which are subsequently classified using a linear multi-class Support Vector Machine (SVM). By combining structured local representations with deep semantic features, the proposed approach enhances robustness, interpretability, and classification performance in complex marine and maritime scenarios.

The HS-LNLSC-DINO framework generates robust and discriminative image-level representations by integrating structured local sparse features with self-supervised global semantic features. This combination enables the simultaneous encoding of fine-grained spatial characteristics, such as vessel contours and coastline geometries, and higher-level contextual information, including scene categories and environmental conditions. As such, the framework

provides a strong analytical basis for applications in marine environmental monitoring, vessel traffic analysis, and coastal infrastructure assessment.

It should be noted, however, that the present study evaluates performance at the image-classification level. The extent to which these representations can be translated into actionable insights for downstream applications, such as risk prioritisation or policy evaluation, has not been empirically examined and remains an important direction for future work. Accordingly, the proposed framework is positioned as an analytical module within broader monitoring and decision-support workflows, rather than as a standalone decision-making tool. Its primary contribution is to improve the reliability and interpretability of upstream image analysis, thereby enhancing the quality of information available to monitoring systems. Issues related to domain adaptation and deployment in operational settings are beyond the scope of this study.

The main contributions of this study are summarised as follows:

(1) A hybrid sparse-semantic image classification framework is developed to support robust marine and coastal image analysis under complex visual conditions and limited labelled data constraints relevant to management practice.

(2) A structured local feature encoding strategy is designed to preserve geometric and spatial characteristics of marine scenes through locality, non-negativity, and semantic consistency constraints.

(3) Self-supervised Vision Transformer features are integrated to enhance global semantic understanding and cross-domain generalisation in data-scarce marine and coastal monitoring contexts.

(4) Extensive experiments on four public benchmark datasets and three maritime image datasets demonstrate consistent performance improvements over representative state-of-the-art methods, highlighting the potential of the proposed framework as a transferable analytical tool for marine environmental monitoring, maritime surveillance, and coastal management applications.

The HS-LNLSC-DINO framework contributes a robust and interpretable approach to image-based analysis in marine and coastal contexts. By strengthening information extraction under low-sample and visually complex conditions, it enhances the quality of data inputs for monitoring and assessment processes, thereby supporting indirectly decision-relevant workflows without explicitly performing policy or operational decision-making.

The remainder of this paper is organised as follows. Section 2 reviews related literature. Section 3 details the proposed framework. Section 4 presents experimental results. Section 5 discusses management implications and practical considerations for integrating the proposed framework into marine and coastal monitoring workflows. Section 6 concludes the paper and discusses future research directions.

2. Literature review

The automated interpretation of marine and coastal imagery has become an increasingly important analytical component of ocean and coastal management systems, supporting environmental monitoring, maritime surveillance, port operations, and coastal governance.

Visual data collected from satellites, UAVs, and shipborne sensing platforms are now routinely used to inform situational awareness, risk assessment, and decision-support workflows across marine and coastal domains. Image classification plays a central role in these workflows by transforming large volumes of heterogeneous visual data into semantically meaningful categories that can be integrated into monitoring dashboards, early-warning systems, and management decision-support tools.

Existing research on image classification for complex natural environments can be broadly organised into three interrelated methodological streams that are particularly relevant to marine and coastal applications: (i) structured local representation learning based on sparse coding, (ii) semantic-aware and joint representation-classification models, and (iii) deep and self-supervised representation learning.

This section reviews representative studies within each stream and critically examines their strengths and limitations from the perspective of robustness, interpretability, and applicability to marine and coastal monitoring and management.

2.1. Sparse coding and structured local representations

Sparse coding has been widely investigated for image classification due to its ability to model local structural characteristics and produce interpretable representations (Cao et al., 2024; Cheng et al., 2023; Shu, 2026). By expressing visual features as sparse linear combinations of dictionary atoms, sparse coding highlights salient local patterns while suppressing redundancy, which is advantageous for scenes with cluttered backgrounds and variable appearance. To enhance descriptive capability, Yang et al. incorporated Spatial Pyramid Matching (SPM) into Sparse Coding (SCSPM), enabling multi-scale spatial layout modelling (Yang et al., 2009). Related spatially aware extensions have also demonstrated effectiveness in capturing spatial context for image categorisation (Lazebnik et al., 2006; Perronnin et al., 2010).

However, conventional sparse coding often suffers from encoding instability, whereby visually similar local features may be assigned to different codewords, leading to inconsistent representations. To address this issue, Gao et al. introduced Laplacian Sparse Coding (LSC), which incorporates Laplacian regularisation to encourage similar local descriptors to have similar codes (Gao et al., 2010). Wang et al. proposed Locality-constrained Linear Coding (LLC), enforcing locality constraints so that nearby descriptors contribute more strongly to reconstruction (Wang et al., 2010). Min et al. further integrated Laplacian regularisation into LLC to strengthen neighbourhood uniformity (Min et al., 2016). Subsequent studies explored adaptive neighbourhood selection and manifold-preserving constraints to further improve locality-aware coding (Lin et al., 2011; Mairal et al., 2008).

Non-negativity constraints have also been introduced to enhance stability and interpretability. Lee and Seung proposed Matrix Factorization (NMF) to learn part-based representations (Lee and Seung, 1999), followed by extensions such as Robust NMF (RNMF) (Zhang et al., 2011). Hoyer integrated NMF with sparse coding principles to derive non-negative sparse coding (Hoyer, 2002), while Cai et al. introduced graph-constrained NMF to preserve intrinsic data structures (Cai et al., 2011). Han et al. proposed Non-negative Laplacian Sparse Coding (Lap-

NMF-SPM), integrating Laplacian regularisation, NMF, and sparse coding to maintain dependencies among local features (Han et al., 2015).

Most of these methods rely on Euclidean distance to measure feature similarity, which is not always suitable for histogram-based descriptors. To overcome this limitation, Wu et al. proposed the Histogram Intersection Kernel (HIK) as a more appropriate similarity measure (Wu and Rehg, 2009), which has been shown to be effective for distribution-based features (Vedaldi and Zisserman, 2012). Chen et al. incorporated the intersection of histograms into LLC for improved environmental categorisation (Chen et al., 2018), while Wan et al. developed Elastic-net and Histogram intersection-based Non-negative Local Sparse Coding (EH-NLSC), jointly modelling sparsity, locality, and non-negativity (Wan et al., 2019).

These structured local representation methods are attractive for marine imagery because they preserve fine-grained geometric patterns that are important for recognising vessels, coastlines, and sea-surface structures. Nevertheless, they mainly rely on handcrafted descriptors and have limited capability to capture high-level semantic information.

2.2. Semantic-aware and joint representation-classification models

Traditional image classification pipelines usually treat feature encoding and classifier learning as separate stages, which may result in semantic information loss. To address this limitation, semantic representation learning has been introduced into image representation frameworks (Zhang et al., 2013). From a generative perspective, Rasiwasia et al. formulated compact semantic spaces based on Gaussian mixture modelling for Scene categorization and retrieval (Rasiwasia and Vasconcelos, 2008a, 2008b). Discriminative approaches incorporate semantic cues more explicitly. Zhang et al. proposed a joint image representation and classification framework by integrating sparse coding with a Random Semantic Space (RSS) (Zhang et al., 2015). Shen et al. combined Global visual features and associated label semantics within a unified segmentation and classification framework (Shen and Zeng, 2019). Additional studies have explored semantic-aware sparse and joint representation-classification models to improve discriminative power (Ruan et al., 2025; Xie et al., 2025).

Moreover, semantic-aware and joint representation-classification models have attracted increasing attention in the maritime and marine engineering domain, where visual data are typically affected by complex backgrounds, dynamic illumination conditions, and a scarcity of annotated samples. By jointly learning visual representations and high-level semantic information, these approaches mitigate semantic ambiguity and enhance feature discriminability, thereby improving robustness in challenging marine environments. Consequently, semantic-aware modelling strategies have demonstrated promising performance in a variety of maritime vision tasks, including maritime object recognition, ship detection, and coastal scene understanding. In particular, incorporating contextual and semantic information into feature learning has been shown to significantly improve the interpretation of marine and coastal imagery under adverse sea-surface and atmospheric conditions (Kanjir et al., 2018; Yang et al., 2025).

Although these approaches partially bridge the gap between low-level features and semantic categories, they remain largely dependent on handcrafted descriptors and local modelling strategies, which limit their ability to capture complex, high-level semantics commonly encountered in marine and coastal imagery. To address these limitations, the proposed HS-LNLSC formulation explicitly integrates locality, non-negativity, and semantic consistency constraints within a unified optimisation framework. These constraints are designed to preserve neighbourhood structure, ensure physically meaningful non-negative representations, and enforce consistency across semantically related local patches. Specifically, the locality constraint encourages similar local descriptors to be represented using nearby dictionary atoms, thereby preserving intrinsic geometric relationships. The non-negativity constraint enhances interpretability by ensuring additive and physically meaningful feature contributions. In addition, the semantic consistency constraint promotes coherence among local features belonging to similar semantic regions, improving discriminative capability under complex visual conditions.

Compared with conventional sparse coding variants, the proposed HS-LNLSC method provides more stable and interpretable local feature representations, particularly in low-sample and visually complex maritime scenarios. By jointly preserving geometric structure and incorporating semantic consistency, the method improves discriminability in challenging conditions such as low contrast, background clutter, and small target objects, which are commonly encountered in marine and coastal imagery.

2.3. Deep and self-supervised representation learning for marine imagery

Deep learning has substantially advanced image representation learning, particularly through Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Smirnov et al., 2014), recurrent architectures such as LSTM (Graves, 2012), and Transformer-based models (Vaswani et al., 2017). ViTs model images as sequences of patches and have demonstrated strong performance in image classification (Dosovitskiy et al., 2021), with subsequent works improving data efficiency and architectural design (Liu et al., 2021; Touvron et al., 2021).

Self-supervised learning approaches, such as DINO, further enhance ViT by enabling robust representation learning without large labelled datasets (Caron et al., 2021). The CLS token produced by pretrained DINO models encodes global semantic information and shows strong transferability, making such models attractive for marine and coastal applications where labelled data are scarce.

However, ViT-based representations primarily focus on global semantic modelling and typically lack explicit mechanisms for preserving local geometric structures. Consequently, they are often treated as black-box features with limited interpretability and may exhibit reduced robustness in visually complex environments.

Recent marine remote sensing studies highlight these challenges. Bakirci et al. conducted an extensive review of deep learning-based approaches for satellite-driven ship detection and ocean monitoring, noting that background clutter and scale variation remain major sources of

error (Bakirci, 2025). Wang et al. proposed Ship-YOLO for ship detection in remote sensing imagery, demonstrating improvements for small targets and complex backgrounds, yet still relying on deep features alone(W. Shi et al., 2025). Beyond ship detection, deep learning-based visual analysis methods have been increasingly applied to a wide range of maritime tasks, including mooring line monitoring and oil spill detection. For instance, Khatri et al. proposed an optimised Faster R-CNN framework for mooring line detection in marine images. By refining the feature extraction strategy of ResNet-50 and reallocating high-level features to the region proposal and classification stages, the method improves localisation accuracy and robustness in complex marine scenes (Khatri et al., 2026). To address the inefficiency and limited accuracy of traditional manual inspection and radar-based approaches for oil spill detection, He et al. proposed a YOLOv12-based framework for accurate oil spill detection and boundary delineation in UAV imagery. By constructing a diverse remote sensing dataset and incorporating high-resolution inputs, pretrained initialisation, and cosine annealing optimisation, the proposed method significantly enhances detection accuracy and robustness in complex marine environments, supporting timely ecological risk assessment and response (He et al., 2026). These studies indicate that while deep learning offers strong semantic modelling capability, complementary local structural modelling remains important for robust marine image understanding.

According to the above discussion, the referenced studies primarily focus on four commonly adopted approaches, denoted as SCSPM, LSC, ViT, and DINO. For a clearer comparison, Table 1 summarises the advantages, disadvantages, and application scenarios of these four methods.

Based on Table 1, existing approaches can be broadly categorised into traditional sparse coding methods and deep learning-based models. Sparse coding techniques, such as SCSPM and LSC, emphasise local geometric structure through sparsity constraints, offering strong noise robustness and interpretability, but often suffer from high computational cost and sensitivity to parameter tuning, which limits scalability. In contrast, deep learning-based models, including ViT and DINO, excel at learning global semantic representations and long-range dependencies, particularly under self-supervised settings, albeit at the expense of substantial data and computational requirements. The complementary characteristics highlighted in Table 1 suggest that integrating sparse coding with deep semantic learning has the potential to yield more robust and efficient image representation frameworks.

Table 1. Comparative analysis of four methods.

Methods	Advantages	Disadvantages	Applications
SCSPM	(1) Efficient and compact feature encoding; (2) Feature selection and dimensionality reduction; (3) Robustness to noise and redundant data; (4) Nonlinear mapping.	(1) Higher computational complexity; (2) Sensitivity to parameters; (3) Risk of overfitting.	(1) Image processing: Image compression, denoising, and feature extraction; (2) Signal processing; (3) Machine learning; (4) Neuroscience.
LSC	(1) Preservation of local structural information; (2) Enforcement of data non-negativity;	(1) Higher computational complexity; (2) Dependency on parameters.	(1) Image compression, denoising, and feature representation; (2) Signal processing;

	(3) Addressing the instability of encoding.		(3) Bioinformatics.
ViT	(1) Strong capability to capture long-range dependencies in images; (2) Flexibility to handle images of varying sizes with patch embeddings; (3) High performance on large-scale image classification tasks.	(1) High dependence on large-scale training data for optimal performance; (2) High computational and memory costs compared to CNNs; (3) Less effective on small datasets without pretraining.	(1) Image classification and recognition; (2) Object detection and segmentation; (3) Medical image analysis and remote sensing.
DINO	(1) Self-supervised learning: does not require labelled data; (2) Generation of high-quality visual representations suitable for downstream tasks; (3) Robust to data augmentations and variations.	(1) Training can be unstable without careful hyperparameter tuning; (2) Sensitive to batch size and learning rate schedules; (3) High computational resource requirement for large-scale training.	(1) Unsupervised feature learning for image classification; (2) Object detection and segmentation pretraining; (3) Image retrieval and clustering.

2.4. Research gaps

Despite substantial progress, several critical gaps remain in the context of marine and coastal image classification:

(1) Limited incorporation of robust similarity measures and semantic cues within structured local encoding.

Most sparse coding-based methods focus on locality, non-negativity, and manifold consistency but rely on Euclidean distance and perform encoding independently of semantic information. A unified structured local encoding framework that simultaneously integrates robust similarity measurement and semantic guidance remains underexplored.

(2) Insufficient integration of structured local representations with deep semantic features.

Existing approaches typically emphasise either handcrafted local modelling or deep semantic learning. Few studies provide principled integration of these two complementary paradigms within a unified framework.

(3) Lack of feature-level fusion strategies that preserve both interpretability and semantic discrimination.

Many hybrid methods adopt heuristic fusion schemes that do not explicitly balance fine-grained local details and global semantic context, leading to suboptimal robustness and transparency.

(4) Limited validation in complex marine and maritime scenarios.

Most methods are evaluated primarily on generic benchmark datasets, with insufficient large-scale validation on maritime imagery characterised by severe illumination changes, background clutter, and limited labelled data.

In summary, these gaps motivate the development of a unified hybrid sparse-semantic image representation framework that combines structured local modelling and self-supervised deep semantic features to support robust marine and coastal image classification. The proposed HS-LNLSC-DINO framework is designed to address these challenges in a complementary and application-oriented manner.

3. Methodology

This section introduces the proposed hybrid sparse-semantic image classification framework, referred to as HS-LNLSC-DINO. The framework combines structured sparse representations with global semantic features extracted from DINO, with the goal of leveraging both local geometric structures and high-level semantic information for robust image classification. An overview of the framework is provided first, followed by detailed descriptions of each component.

3.1. The proposed framework

The proposed framework consists of five main stages, as illustrated in Fig. 1. First, local SIFT features are computed from each image to capture invariant local patterns. Second, these descriptors are encoded using Histogram Intersection-based Local Non-negative LSC (HI-LNLSC) to generate discriminative image representations. Third, semantic information is incorporated into the HI-LNLSC framework to jointly model image representation and classification, resulting in the HS-LNLSC feature representation. Through these three steps, a structured sparse representation is obtained, preserving local geometric relationships. Fourth, global semantic features are extracted from the CLS token of a pretrained DINO model, providing image-level semantic representations learned in a self-supervised manner. Simultaneously, the sparse and semantic features are fused at the feature level to form the final feature representation. Finally, classification is performed using a linear multi-class SVM.

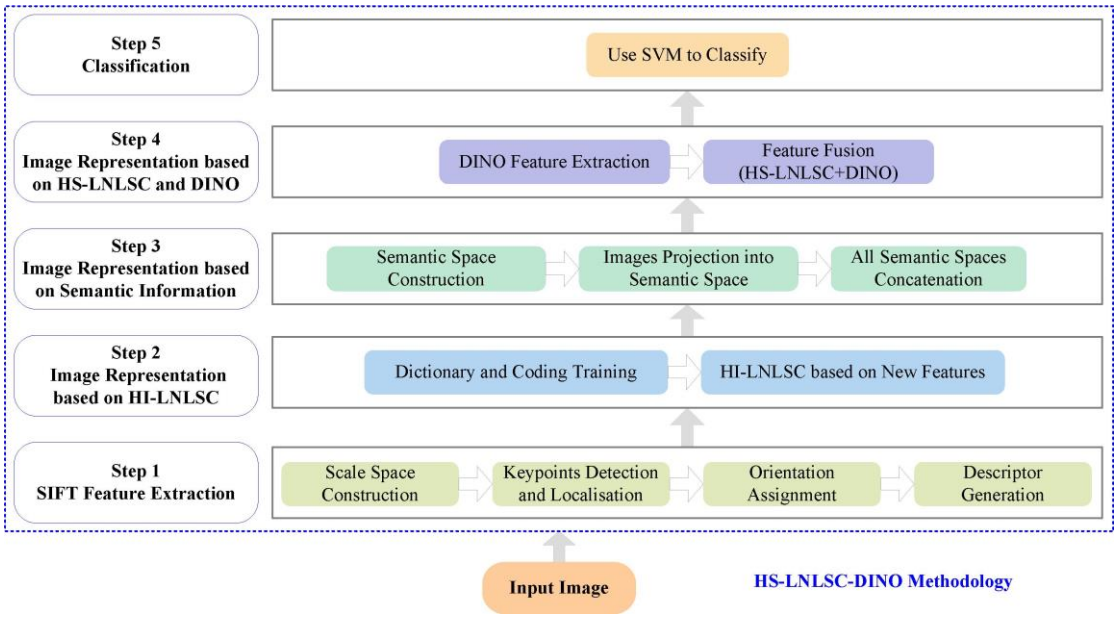


Fig. 1. An overview of the HS-LNLSC-DINO framework.

3.2. HS-LNLSC

This subsection introduces the LNLSC approach, which extends the LSC model by jointly enforcing locality and non-negativity constraints. To further improve locality modelling, the distance metric of the optimisation framework is enhanced by integrating histogram intersection. Building on this formulation, the HS-LNLSC method is developed through combining image representation and classification to leverage semantic information.

In the first stage, the HI-LNLSC method is used to encode local image features, with Max Pooling (MP) applied to generate the initial image representations. In the second stage, a randomly sampled subset of these image representations is used to build the semantic space. A trained classifier then projects all training images into this space, producing the image feature representations that incorporate semantic information.

By jointly incorporating locality and non-negativity constraints, histogram intersection-based similarity, and semantic information, this framework substantially improves the effectiveness of image representation and the overall classification performance of the HS-LNLSC method.

3.2.1. HI-LNLSC-based image representation

This section explains how the HI-LNLSC method generates the original image representations. It achieves this by embedding histogram intersection into the LNLSC optimisation framework, which redefines the similarity measure between feature vectors and dictionary atoms, thereby providing a more effective way to quantify their similarity.

(1) Dictionary and corresponding coding training

As a result of the prohibitive complexity of constructing local adaptors (Wang et al., 2010) and the Laplacian matrix for the entire set of local features, a subset of template features is randomly sampled and employed to train the dictionary and coding. The original formulation of the HI-LNLSC approach is described below. Let X represent the non-negative input feature matrix, while D and S denote the non-negative dictionary and the associated non-negative sparse representation, respectively, with the definitions of $X=[x_1,x_2,...,x_N]\in R^{M\times N}$, $D=[d_1,d_2,...,d_Q]\in R^{M\times Q}$, $V=[v_1,v_2,...,v_N]\in R^{Q\times N}$. The optimisation problem is formulated by integrating locality and non-negativity constraints into the LSC framework as follows:

$$\begin{cases} \min\limits_{D,S} \sum\limits_{i=1}^{N} (\left\| x_i - Ds_i \right\|_2^2 + \lambda \left\| b_i \odot v_i \right\|_2^2) + \beta tr(VLV^T) \\ s.t. \qquad sparseness(d_j) = S_d \\ \qquad\quad \left\| d_j \right\|_2^2 \le 1, D \ge 0, V \ge 0, \forall j \end{cases} \tag{1}$$

where λ, β and S_d are predefined constant parameters. The sparseness measure is S_d defined by jointly incorporating the l_2 -norm and l_1 -norm, as expressed below:

$$sparseness(d_j) = \frac{\sqrt{M} - \left\| d_j \right\|_1 / \left\| d_j \right\|_2}{\sqrt{M} - 1} \tag{2}$$

where M represents the dimension of d_j, namely, $d_j \in R^{M\times 1}$.

In the present work, the conventional Euclidean distance used in the LLC framework to evaluate the similarity between feature vectors and dictionary atoms is replaced by a histogram intersection-based similarity measure. $b_i \in R^Q$ represents a local weighting operator, which is given by:

$$\begin{aligned} b_i &= \sigma / dist(x_i, D) \\ dist(x_i, D) &= [dist(x_i, d_1), ..., dist(x_i, d_M)]^T \end{aligned} \tag{3}$$

Here, σ is a parameter that regulates the strength of weight regularisation, and $dist(x_i, d_j)$ denotes the distance between x_i and d_j, evaluated based on histogram intersection. The computation can be expressed as:

$$dist(x_i, d_j) = \sum_{k=1}^{Q} \min(x_{ik}, d_{jk}) \tag{4}$$

Here, Q denotes the histogram dimension, which corresponds to the number of dictionary atoms. Additionally, x_{ik} and d_{jk} denote the k-th components of the feature vector x_i and the dictionary atom d_j, respectively.

To solve Eq. (1), an alternating optimisation strategy is adopted, in which D and V are iteratively updated by fixing one variable while optimising the other. First, with X and D held fixed, V is optimised. By applying the Method of Lagrange Multiplier (MLM) along with the Karush-Kuhn-Tucker (KKT) theorem, the corresponding update strategy for V is derived as follows:

$$\begin{aligned} &v_{ij} = v_{ij} \frac{(D^T X + \beta V W)_{ij}}{(D^T D V + \beta V D + \lambda diag(g_i)\bar{V})_{ij}}, \\ &\forall\, i = 1, 2, ..., M,\ \ j = 1, 2, ..., N \end{aligned} \tag{5}$$

where $diag(g_i)$ denotes a diagonal matrix whose diagonal entries are given by g_i, $g_i = [b_{i1}^2, b_{i2}^2, ..., b_{iN}^2]$ is an N-dimensional vector represented as a row vector, and $b_{ij}^2 = (\sigma / dist(x_j, d_i))^2$ and $\bar{V} = [diag(v_1), diag(v_2), ..., diag(v_N)]^T$ represent vectors of ones and zeros, respectively.

Next, with X and V fixed, the optimisation is performed with respect to D. For this problem, the diagonal matrix Λ is derived by solving the corresponding Lagrange dual formulation using the conjugate gradient method. The result Λ is then substituted into the following expression to compute D:

$$D = (XV^T)(VV^T + \Lambda)^{-1} \tag{6}$$

(2) HI-LNLSC based on new features

A set of template features, X, is randomly selected in Section 3.2.1 (1) to train the matrices D and V. For an unseen local feature matrix Y, the learned matrices D and V are directly used for HI-LNLSC encoding, leading to the following optimisation formulation:

$$\begin{cases} \min\limits_{V} \|Y - DS\|_F^2 + \lambda \|b \odot S\|_F^2 + \dfrac{\beta}{2} \sum\limits_{ji} \|s_j - v_i\|_2^2 w_{ji} \\ s.t.\ s_{ij} \geq 0, \forall i, j \end{cases} \tag{7}$$

Here, v_i denotes the i-th element of the column vector in V, and S denotes the sparse representation of the updated feature matrix Y. The similarity matrix W is constructed using a K-nearest neighbour strategy based on histogram intersection. If a template feature x_i is among the K-nearest neighbours of a new feature y_j, then w_{ji} is set to 1; otherwise, it is set to 0. This similarity metric is consistent with Eq. (4).

The update rule for S can be derived by applying the MLM and the KKT conditions, as outlined below.

$$s_{ij} = s_{ij} \frac{(D^T X + \beta V W^T)_{ij}}{(D^T D S + \frac{1}{2}\beta S A + \lambda diag(g_i)\bar{S})_{ij}} \tag{8}$$

where A is a diagonal matrix with $a_{jj} = \sum_i w_{ji}$ as its diagonal elements, $diag(g_i)$ follows the definition provided in Eq. (5), and $\bar{S} = [diag(s_1), diag(s_2), \cdots, diag(s_N)]^T$ denotes a vector of ones. After training D and V on the template features, Eq. (8) is applied to encode new features using HI-LNLSC.

To achieve feature fusion, the MP method is utilised in this paper. The detailed procedure is outlined below.

$$f_l = \max\{|s_{1l}|, |s_{2l}|, \cdots, |s_{Nl}|\}, l = 1, 2, \cdots, Q \tag{9}$$

Here, the l-th component of the sparse representation s_N is denoted by s_{Nl}, and f_l corresponds to the l-th component of the vector f. Consequently, each spatial pyramid region is represented by a Q-dimensional descriptor f, as defined by $f = [f_1, f_2, \cdots, f_Q]^T$.

After computing the region-level representations, the SPM method is applied to derive the final image descriptor.

3.2.2. Semantic information-based image representation

To address the limitations of the HI-LNLSC method, specifically its relatively independent processing, as well as the semantic gap between visual features and human interpretation, the HS-LNLSC approach is introduced. This approach aims to effectively combine both visual cues and semantic information, thereby modelling interactions between semantic entities and their contextual surroundings. First, the HI-LNLSC approach is used to perform encoding of local image features, resulting in an initial image representation. Next, a semantic space is created to refine and derive the resulting visual representation. Ultimately, a SVM is utilised to categorise the images based on these semantic-aware representations.

(1) Semantic space construction.

The semantic space is formulated as a unified representation domain constructed from various visual representations, providing a basis for semantic modelling of images. Each semantic space is constructed via classifier training on a randomly sampled image subset.

Let X denote the initial image feature representations of H training samples obtained utilising the HI-LNLSC approach, where the samples belong to E classes with corresponding labels denoted by z. A subset of P($P \le H$) images is randomly selected from these representations to build the semantic space, and the above procedure is iteratively performed R times. The resulting sets from each iteration are denoted as $\{(u^{1,1}, z^{1,1}), ..., (u^{P,1}, z^{P,1})\}, ..., \{(u^{1,R}, z^{1,R}), ..., (u^{P,R}, z^{P,R})\}$. For the r-th random selection $\{(u^{1,r}, z^{1,r}), ..., (u^{P,r}, z^{P,r})\}$, a SVM is employed to construct the corresponding semantic space, as follows:

$$f_e^r(z^{p,r}) = \bar{z}_e^{p,r} = w_e^r u^{p,r} + b_e^r, \quad p = 1,2,...,P, \quad e = 1,2,...,E \tag{10}$$

An optimisation problem based on the hinge loss function is subsequently formulated as follows:

$$\min_{w_e^r} \left\| w_e^r \right\|^2 + \alpha \sum_{p=1}^{P} p(\bar{z}_e^{p,r}, z^{p,r}) \tag{11}$$

Solving Eq. (11) yields the corresponding values of w_e^r and b_e^r. Each component of the resulting semantic space is associated with an individual classifier, which is trained using a randomly chosen subset of samples. Given E image classes, the resulting semantic space consists of E dimensions.

(2) Image projection into semantic space.

After training the SVM, all training images are projected into the constructed semantic space as follows:

$$u_{r,k}^{ss} = (f_1^r(u^h); f_2^r(u^h); ...; f_E^r(u^h)), r = 1,2,...,R; h = 1,2,...,H \tag{12}$$

Here, the superscript '*ss*' is used to indicate the 'semantic space'.

(3) All semantic spaces concatenation.

Once all spaces are constructed, all training images are mapped into each of the R learned spaces. The final image representation is then obtained by concatenating the corresponding feature vectors across these spaces, as follows:

$$u_k^{ss} = (u_{1,h}^{ss}; u_{2,h}^{ss}; ...; u_{R,h}^{ss}), h = 1,2,...,H \tag{13}$$

Thus, the locally structured image representation is obtained via Eq. (13).

3.3. HS-LNLSC-DINO

While HS-LNLSC effectively captures local geometric structures and preserves discriminative local relationships through locality, non-negativity, and histogram intersection constraints, it primarily relies on handcrafted local descriptors and lacks explicit modelling of high-level semantic information. In contrast, DINO excels at learning rich global semantic representations but does not explicitly encode local geometric structures, often treating these features as black-box representations. Motivated by the complementary strengths of these two approaches, this section introduces a hybrid sparse-semantic image representation framework, termed HS-LNLSC-DINO, which integrates structured sparse coding with self-supervised semantic features.

3.3.1. DINO feature extraction

In parallel with local sparse encoding, global semantic features are extracted using DINO. Given the same input image, it is divided into fixed-size, non-overlapping patches, which are then processed by the Transformer encoder. Let $f_{CLS} \in R^{Dino}$ denote the output CLS token of DINO. This token encodes holistic, image-level semantic information learned through self-distillation, without the need for labelled supervision, and has demonstrated strong transferability across various visual tasks.

Unlike local sparse features, f_{CLS} captures long-range contextual dependencies and high-level semantics, making it particularly well-suited for complex marine and maritime scenes, where global context plays a crucial role.

3.3.2. Sparse-semantic feature fusion

To leverage the complementary strengths of structured sparse representations and deep semantic features, the HS-LNLSC and DINO features are fused at the feature level. The local structural features of the image, obtained through Eq. (13), are denoted by $f_{\text{HS-LNLSC}}^{\text{T}}$, and the final image representation is constructed as follows:

$$f = \left[f_{\text{HS-LNLSC}}^{\text{T}}, \alpha f_{\text{CLS}}^{\text{T}} \right]^{\text{T}} \tag{14}$$

where α is a scaling factor that balances the contributions of the local sparse features and global semantic features.

This fusion strategy preserves fine-grained local geometric details through HS-LNLSC, while simultaneously incorporating global semantic understanding via the DINO CLS token. As a result, the fused representation achieves enhanced robustness, improved discriminative capability, and better interpretability compared to purely sparse or purely deep representations.

Finally, the fused feature vector f is fed into a multi-class linear SVM for classification. The use of a linear classifier further emphasises the discriminative power of the proposed representation without introducing additional model complexity.

Rather than relying on simple feature concatenation, the proposed framework adopts a principled feature-level integration of structured local descriptors derived from HS-LNLSC and global semantic representations obtained from DINO CLS tokens. This hybrid design effectively combines fine-grained geometric information with high-level contextual semantics, thereby enhancing discriminative capability and robustness under data-limited conditions typical of maritime image analysis. By leveraging the complementary strengths of local and global representations, the framework achieves more reliable and interpretable classification performance than either component used in isolation.

3.4. Algorithmic description

Once the formulations of matrices *D* and *V* are established, the dictionary and corresponding sparse representations of the template features are obtained via a two-stage procedure. First, Algorithm 2 iteratively updates the diagonal matrix *D* associated with the Lagrange multipliers. Then, Algorithm 3 identifies the optimal approximation that enforces the desired sparsity, with *D* updated accordingly. The specific algorithmic procedure is provided in Appendix A and B.

Algorithm 1 is then employed to learn both the dictionary for HS-LNLSC-DINO and its associated sparse codes. This algorithm iteratively updates both *D* and *V* until a predefined stopping criterion is reached. Notably, Algorithm 1 consolidates the operations of Algorithms 2 and 3 in the Appendix.

In conclusion, the overall procedure is summarised below.

Algorithm 1 (HS-LNLSC-DINO)

Input: Non-negative feature matrix X; Initial dictionary D; Initial sparse coding V; Laplacian matrix L; Parameters λ, β, σ, S_d; Training samples$H$; Training times R.

Output: Updated dictionary D; Sparse coding V; Category annotations.

Step 1. Initialise the variables X and V, namely, $X = X / \max(X(:))$, $V = V / \|V\|_1$.

Step 2. Convergence Check: While convergence criteria are not met, repeat the following steps:

Step 3. Update sparse coding V with Eq. (5).

Step 4. Normalise the dictionary D and sparse coding V, namely, $d_{ij} = d_{ij} / \sqrt{\sum_i v_{ij}}$, $v_{ij} = v_{ij} / \sqrt{\sum_i v_{ij}}$.

Step 5. Update Lagrange dual matrix Λ according to **Algorithm 2**.

Step 6. Each column vector of matrix D is projected individually according to **Algorithm 3** to acquire the updated values. Let D^* represent the optimal dictionary and V represent the corresponding sparse coding.

Step 7. Check the convergence condition for Step 8 in **Algorithm 2.**

Step 8. If the convergence condition is satisfied, proceed to the next step.

Step 9. If not, return to Step 3 and repeat the process.

Step 10. Post-Processing: After updating D and V, the sparse codes S for the new features are computed using Eq. (8).

Step 11. Apply SPM via Eq. (9) to implement MP for the encoded features and acquire the initial image representation.

Step 12. Semantic Space Construction: Eq. (10) is employed to build the semantic space, after which w_e^r and b_e^r are obtained using Eq. (11).

Step 13. All images in the training set are embedded within the semantic space in accordance with Eq. (12).

Step 14. Integrate all semantic spaces to form the final image representation using Eq. (13).

Step 15. Fuse the HS-LNLSC and DINO features to obtain HS-LNLSC-DINO features according to Eq. (14).

Step 16. Implement a SVM to categorise the image based on its semantic representation.

4. Experiments

The section evaluates the proposed hybrid sparse-semantic framework in terms of accuracy, robustness, and practical suitability for marine and coastal monitoring tasks. The experiments are organised as follows. Section 4.1 introduces the datasets and implementation settings, with particular attention to the characteristics that reflect operational marine imaging conditions (e.g., clutter, scale variation, and limited labelled data). Section 4.2 presents ablation studies to quantify the contribution of each component and to examine design choices relevant to deployment (feature selection, Transformer layer depth, and training data availability). Section 4.3 benchmarks the proposed approach against representative methods to assess its competitiveness under standard evaluation protocols. Section 4.4 investigates generalisation and stability to evaluate reliability under repeated sampling and across diverse datasets. Section 4.5 presents a discussion of the computational complexity of the proposed algorithm.

4.1. Datasets and experimental settings

4.1.1. Datasets description

To evaluate performance across both general-purpose and domain-specific conditions, experiments are conducted on four widely used benchmark datasets (Corel-10, Scene-15, Caltech-101, and Caltech-256) and three maritime image datasets representing realistic marine visual complexity. Table 2 summarises the benchmark datasets, and Fig. 2 provides example images from Corel-10. These benchmark datasets are included to ensure comparability with established baselines and to verify that the proposed framework is not domain-specific.

For marine and coastal relevance, three maritime datasets are considered: Open Seaship, Seaship-trimming, and SMD-trimming. The Open Seaship dataset comprises 31,455 images across seven common ship categories (passenger ships, mixed types, container ships, bulk carriers, ore carriers, general cargo ships, and fishing vessels), as summarised in Table 3. In addition, the Seaship dataset represents the complete collection of maritime images, whereas Seaship-trimming refers to a pre-processed subset specifically employed for ablation studies or small-sample experiments. This distinction has been applied consistently throughout the manuscript to ensure clarity and avoid potential confusion. The Singapore Maritime Dataset (SMD) includes imagery captured in diverse operational settings (on-board, near-infrared, and on-shore) with substantial variability in background, illumination, and target scale, as summarised in Table 4. Since the original SMD was designed for detection rather than classification, experiments employ the SMD-trimming dataset to enable controlled evaluation within an image classification pipeline. Example images from Seaship are shown in Fig. 3.

Overall, the selected maritime datasets reflect key challenges encountered in practice, including background interference, small targets, viewpoint changes, and limited labelled samples, conditions that directly affect the reliability of monitoring outputs in marine and coastal management settings.

Table 2. Four benchmark image datasets.

Datasets	Classes	Images Per Class	Total Images
Corel-10	10	100	1000
Scene-15	15	200~400	4485
Caltech-101	101	31~800	9144
Caltech-256	256	≥80	29,780

Fig. 2. Sample images from the Corel-10 dataset.

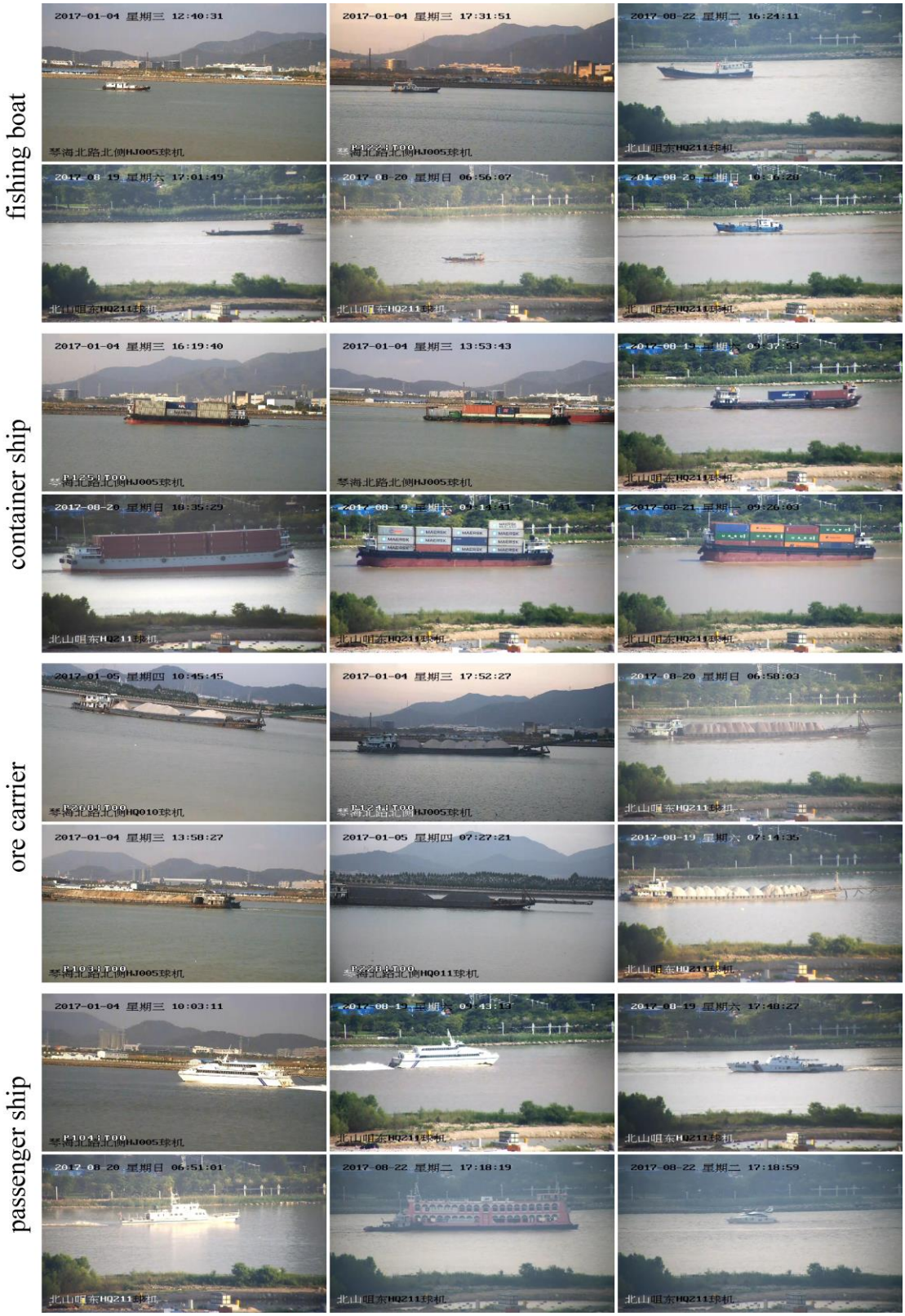


Fig. 3. Some pictures of the Seaship.

Table 3. The Seaship dataset.

Ship Class	Images	Percentage
Passenger ship	3171	0.1008
Mixed type	3440	0.1094
Container ship	3657	0.1163
Bulk cargo carrier	5067	0.1610

Ore carrier	5126	0.1630
General cargo ship	5342	0.1698
Fishing boat	5652	0.1797
Total	31455	1

Table 4. The SMD dataset.

Subdataset	Videos (Annotated)	Labelled Frames	Number of Labels
VIS on-board	11 (4)	2400	3173
NIR	30 (23)	11,286	83,174
VIS on-shore	40 (36)	17,967	154,495
Total	81 (63)	31,653	240,842

4.1.2. Experimental settings

The experimental pipeline follows a hybrid representation strategy combining structured local encoding and self-supervised global semantic representations.

Structured local feature extraction and coding. Dense SIFT descriptors are extracted using a sliding window (step size: 8 pixels) with a fixed window size. Each local descriptor has dimensionality 128. A dictionary of size 1024 is learned for encoding. To clarify the selection of key parameters in the HS-LNLSC-DINO framework, we provide the following rationale. The locality regularisation parameter (λ), Laplacian smoothness (β), and histogram intersection weight (σ)were determined through a combination of empirical stability analysis across multiple datasets and guidance from previous studies (Han et al., 2015; Wang et al., 2010). Preliminary experiments demonstrated that the model's performance is relatively insensitive to small variations around these values, indicating that the chosen default settings achieve consistent and robust performance without the need for dataset-specific tuning. The fusion parameter α , which governs the relative contribution of local and global features, was set to 0.4 to ensure a stable and balanced integration of HS-LNLSC and DINO representations. Similarly, the sparsity parameter S_d was set to 0.4, providing sufficiently discriminative sparse coding while preventing excessively sparse representations that could compromise model performance.

Semantic space construction. A semantic space is constructed using a fixed proportion of images (P=0.3H) and repeated sampling (R=30) to reduce sensitivity to data selection and to improve robustness.

DINO feature extraction. Images are resized to a fixed resolution prior to feeding into a pretrained DINO ViT with patch size 16. The network weights are frozen during feature extraction to reduce overfitting risk, which is particularly relevant for operational marine datasets where labelled data may be limited. The CLS token is used as the global semantic representation. Key configurations are summarised in Table 5.

Classification and evaluation. For all experiments, classification is performed using a multi-class linear SVM to ensure consistent evaluation and to isolate the effect of representation learning from classifier complexity. Unless otherwise specified, accuracy is reported as mean ± standard deviation across repeated runs.

To evaluate model performance under realistic operational constraints, training protocols with varying numbers of samples per class (50, 100, 150, 200, and 250) are adopted, simulating low-data conditions common in maritime monitoring scenarios. The rationale for these choices is explicitly described to clarify that the protocols are designed to assess robustness under limited labelled data rather than arbitrary selection.

Table 5. Parameter settings used in each stage of the experiment.

Image Datasets	**Corel-10**	**Scene-15**	**Caltech-101**	**Caltech-256**	**Seaship**	**Seaship-trimming**	**SMD-trimming**
The SIFT feature extraction stage							
Approaches	SIFT	SIFT	SIFT	SIFT	SIFT	SIFT	SIFT
Step size	8	8	8	8	8	8	8
Window size	16×16	16×16	16×16	16×16	16×16	16×16	16×16
The HS-LNLSC stage							
Dictionary size	1024	1024	1024	1024	1024	1024	1024
λ	0.4	0.4	0.4	0.4	0.4	0.4	0.4
β	0.2	0.2	0.2	0.2	0.2	0.2	0.2
σ	100	100	100	100	100	100	100
Sparseness S_d	0.4	0.4	0.4	0.4	0.4	0.4	0.4
The construction of the semantic space stage							
The number of images utilised to form the semantic space (P)	0.3H	0.3H	0.3H	0.3H	0.3H	0.3H	0.3H
The times of repetition selected (R)	30	30	30	30	30	30	30
The DINO feature extraction stage							
The size of the input image	320×320	320×320	320×320	320×320	320×320	320×320	320×320
Patch size	16×16	16×16	16×16	16×16	16×16	16×16	16×16
α	0.4	0.4	0.4	0.4	0.4	0.4	0.4

4.2. Ablation studies

To identify the components that most critically influence performance under realistic marine imaging conditions, a comprehensive set of ablation studies is conducted from three complementary perspectives: (i) feature source and fusion strategy, (ii) the choice of Transformer layers for semantic representation learning, and (iii) sensitivity to the size of the training dataset. Collectively, these analyses provide actionable insights for practical deployment, particularly with respect to computational efficiency, annotation burden, and robustness in data-scarce scenarios.

4.2.1. Selection of feature combinations

This experiment investigates the relative contributions of different feature sources and their combinations to classification performance, with the objective of identifying the most effective

representation strategy within the proposed framework. Three categories of features are considered: (1) structured sparse representations generated by HS-LNLSC, (2) patch-level features derived from intermediate DINO representations (denoted as DINO-Patch), and (3) global semantic features extracted from the CLS token of DINO (denoted as DINO-CLS). Based on these feature types, five representative configurations are evaluated: HS-LNLSC, DINO-Patch, DINO-CLS, HS-LNLSC+DINO-Patch, and HS-LNLSC+DINO-CLS.

Experiments are conducted on two representative datasets, Seaship and Scene-15, under identical experimental settings. For each class, 50 images are randomly selected for training, with the remaining samples reserved for testing. To further investigate the influence of semantic feature depth, DINO representations are extracted from multiple Transformer layers (L2, L4, L6, L8, L10, and L12). A multi-class linear SVM classifier is employed for all configurations to ensure a fair and consistent comparison.

The quantitative results are reported in Tables 6 and 7, with corresponding performance trends illustrated in Fig. 4. Several important observations emerge. First, HS-LNLSC alone yields stable and competitive performance across all Transformer layers on both datasets, indicating strong robustness and minimal sensitivity to semantic depth. This behaviour is expected, as HS-LNLSC is built upon handcrafted local descriptors and structured sparse coding, which are independent of Transformer-based semantic representations.

Second, DINO-CLS consistently outperforms DINO-Patch on both datasets, and its performance improves markedly as deeper Transformer layers are used. This finding suggests that the CLS token progressively encodes more discriminative global semantic information at deeper layers, whereas patch-level features mainly emphasise local visual patterns that are less effective when used in isolation for classification.

Third, the combination of HS-LNLSC and DINO-CLS achieves the best performance across nearly all layers and datasets. HS-LNLSC+DINO-CLS consistently surpasses both individual components, demonstrating strong complementarity between structured local sparse representations and global semantic features. By contrast, integrating HS-LNLSC with DINO-Patch yields only marginal gains or even performance degradation, implying that patch-level semantic features provide limited additional information beyond that already captured by HS-LNLSC.

Overall, these results indicate that HS-LNLSC+DINO-CLS constitutes the most effective feature combination. HS-LNLSC contributes stable and interpretable local geometric modelling, while DINO-CLS supplies high-level semantic discrimination. Their integration produces a unified representation that is both robust and semantically expressive. Consequently, HS-LNLSC+DINO-CLS is selected as the optimal feature configuration and adopted in all subsequent experiments.

Table 6. Classification accuracy on the Seaship dataset with different feature combinations (%).

Methods \ Layers	L2	L4	L6	L8	L10	L12
HS-LNLSC	78.73	78.73	78.73	78.73	78.73	78.73
DINO-Patch	55.94	57.18	56.54	54.86	56.05	53.08
DINO-CLS	41.13	55.97	77.53	84.64	89.48	90.88

HS-LNLSC+DINO-Patch	56.57	57.98	57.28	55.39	55.51	55.15
HS-LNLSC+DINO-CLS	**79.52**	**80.93**	**81.12**	**87.14**	**90.92**	**92.76**

Table 7. Classification accuracy on the Scene-15 dataset with different feature combinations (%).

Methods \ Layers	L2	L4	L6	L8	L10	L12
HS-LNLSC	87.53	87.53	87.53	87.53	87.53	87.53
DINO-Patch	68.64	67.90	67.35	66.92	67.98	68.21
DINO-CLS	60.74	82.68	90.45	94.39	96.28	96.27
HS-LNLSC+DINO-Patch	67.95	68.95	69.21	67.45	69.25	68.15
HS-LNLSC+DINO-CLS	**88.46**	**90.01**	**93.55**	**95.19**	**96.59**	**97.51**

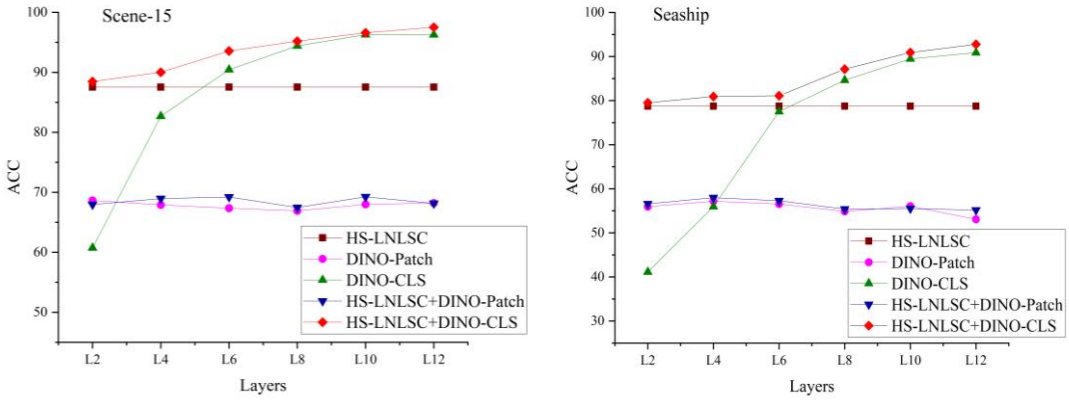


Fig. 4. Classification accuracy for five different methods in two image datasets.

4.2.2. Influence of Transformer layer selection

Following the identification of HS-LNLSC+DINO-CLS as the most effective feature combination in Section 4.2.1, the impact of Transformer layer selection on overall classification performance is further examined. Because different layers of a ViT encode semantic information at progressively higher levels of abstraction, determining the layer that provides the most discriminative and complementary semantic representation when fused with HS-LNLSC is of practical importance.

In this experiment, CLS-token features are extracted from multiple Transformer layers of the DINO model, namely L2, L4, L6, L8, L10, and L12, while the HS-LNLSC component is kept unchanged. To further investigate the robustness of the HS-LNLSC-DINO framework under more constrained training conditions, we reduce the number of training samples per class to 30, while maintaining the same protocol as in Section 4.2.1. This allows assessment of relative performance across ablation variants without affecting comparability.

The quantitative results are reported in Table 8, and the corresponding trends are illustrated in Fig. 5. A clear performance increase is observed as the Transformer depth grows, indicating that deeper layers encode increasingly discriminative semantic information. In contrast, shallow layers (L2 and L4) mainly capture low-level visual cues such as edges and textures, which offer limited additional complementarity to HS-LNLSC and therefore lead to smaller performance gains.

Mid-to-deep layers (L6 and L8) yield substantial improvements, suggesting a favourable trade-off between semantic abstraction and generalisation capability. Among these, L8 consistently demonstrates strong performance across both datasets, indicating that features at

this depth effectively complement the structured sparse representations produced by HS-LNLSC.

The highest classification accuracy is consistently obtained using features from L12. This outcome indicates that semantic representations extracted from the deepest Transformer layer are the most beneficial when integrated with HS-LNLSC. Consequently, Layer 12 (L12) is selected as the optimal semantic feature extraction layer and is adopted in all subsequent experiments, including analyses of training-sample sensitivity and generalisation performance.

Table 8. Classification accuracy for different Transformer layers on the two image datasets (%).

Datasets \ Layers	L2	L4	L6	L8	L10	L12
Seaship	79.52	80.93	81.12	87.14	90.92	**92.76**
Scene-15	88.46	90.01	93.55	95.19	96.59	**97.51**

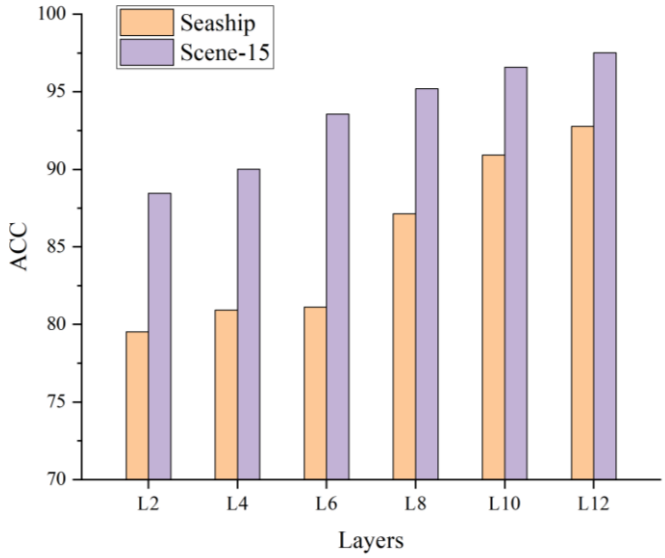


Fig. 5. Classification accuracy for different Transformer layers on the two image datasets.

4.2.3. Impact of the number of training samples

This subsection evaluates the robustness of the proposed HS-LNLSC-DINO framework under varying training sample sizes, a critical consideration for real-world marine and maritime applications where annotated data are often scarce. The objective is to characterise the evolution of classification performance as the amount of training data increases and to assess the data efficiency of the proposed approach.

Experiments are conducted on three datasets with distinct characteristics: Seaship, SMD-trimming, and Scene-15. For the Seaship and SMD-trimming datasets, the number of training samples per class is set to 50, 100, 150, 200, and 250, whereas for the Scene-15 dataset, 50, 100, 150, and 200 samples per class are considered. In all cases, the remaining samples are reserved for testing. The optimal configuration identified in the preceding subsections, namely, the HS-LNLSC+DINO-CLS feature fusion strategy with semantic features extracted from Layer 12 (L12), is adopted throughout.

The quantitative results are reported in Table 9 and visualised in Fig. 6. As expected, classification accuracy increases monotonically with the number of training samples, reflecting typical learning behaviour. More importantly, strong performance is observed even in low-data regimes. With only 50 training samples per class, the HS-LNLSC-DINO framework achieves competitive accuracy across all three datasets, demonstrating high data efficiency.

This favourable behaviour can be attributed to the complementary nature of the hybrid representation. The structured sparse coding component captures stable and discriminative local geometric patterns that generalise well with limited supervision, while the self-supervised DINO features contribute transferable semantic knowledge learned from large-scale unlabeled data. As additional training samples become available, the classifier is able to further exploit these rich representations, leading to consistent performance gains. Collectively, these results indicate that the proposed framework not only achieves high accuracy with sufficient training data but also demonstrates strong robustness and generalisation in low-sample scenarios, underscoring its suitability for practical marine and maritime image classification tasks.

Table 9. Classification accuracy for different Training images across the three image datasets (%).

Datasets	**Training Images**				
	50	**100**	**150**	**200**	**250**
Seaship	92.76	96.38	97.85	98.12	98.75
SMD-trimming	94.90	97.35	98.65	99.13	99.68
Scene-15	97.51	98.50	99.26	99.61	/

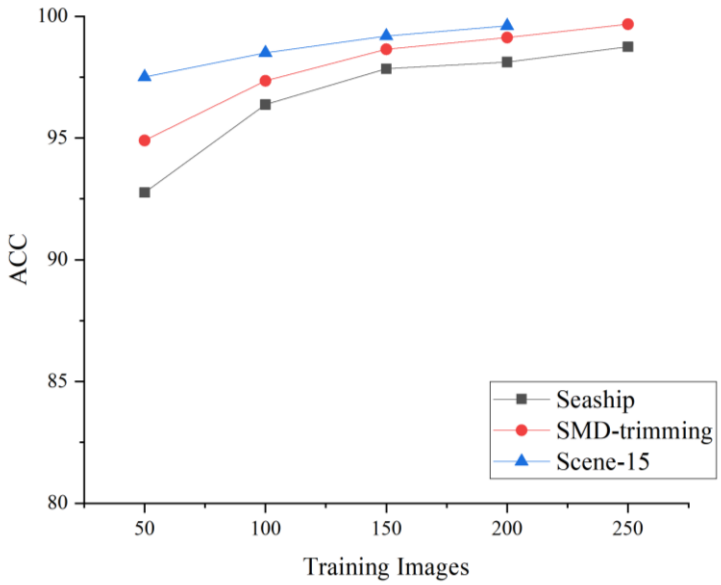


Fig. 6. Classification accuracy for different Training images on the three image datasets.

4.3. Comparison with state-of-the-art methods

To comprehensively assess the effectiveness of the proposed HS-LNLSC-DINO framework, extensive comparisons are conducted with a broad spectrum of representative state-of-the-art image classification methods. The selected methods encompass diverse image representation paradigms, such as SCSPM, LLC, LSC, Lap-NMF-SPM, EH-NLSC, RSS, and HS-NLLSC (Y. Shi et al., 2025) approaches.

Regarding the four benchmark datasets, different training-testing splits are adopted according to commonly used evaluation protocols. In particular, with respect to the Corel-10 and Scene-15 datasets, 50 and 100 images from each class are randomly chosen to train the model, while the leftover images serve as the testing set. For the Caltech-101 dataset, experiments are conducted using 15 and 30 training images per category, while all remaining images are reserved for testing. In the case of Caltech-256, 15, 30, 45, and 60 images per class are randomly selected for training, and the remaining images are used for testing. For the three maritime datasets, the number of training samples per category is varied to assess performance under different data availability conditions. Specifically, 50, 100, 150, 200, and 250 images from each class are randomly chosen for training, while the remaining images are allocated for testing.

Table 10 presents the comparative results between the proposed HS-LNLSC-DINO method and several representative cutting-edge approaches over four benchmark datasets. The evaluated methods include classical sparse coding models (SCSPM, LLC, LSC), structured sparse representation techniques (Lap-NMF-SPM, EH-NLSC), semantic-based models (RSS), HS-NLLSC, and deep feature-based baselines. Across all datasets and experimental settings, HS-LNLSC-DINO consistently achieves the highest classification accuracy, demonstrating clear and robust performance advantages.

In addition, statistical significance analyses were performed to evaluate the reliability of HS-LNLSC-DINO's performance gains relative to the second-best baseline across all datasets and training sample sizes. As shown in Table 10, the reported P-values indicate that most improvements are statistically significant ($P < 0.05$), suggesting that the observed advantages are unlikely to arise from random variation. These findings further support the robustness and generalizability of the proposed framework across diverse datasets and varying sample conditions.

The superior performance of HS-LNLSC-DINO can be attributed to its integrated sparse-semantic representation mechanism. The HS-LNLSC component incorporates locality, non-negativity constraints, Laplacian regularisation, and histogram intersection similarity, facilitating the stable and consistent encoding of local SIFT features while preserving spatial geometric relationships. This design effectively addresses the coding instability and feature cancellation issues commonly encountered in traditional sparse coding frameworks. In parallel, the incorporation of global semantic features extracted from DINO introduces high-level contextual information that is challenging to capture using handcrafted descriptors alone.

By contrast, traditional sparse coding approaches, such as SCSPM, LSC, and Lap-NMF-SPM, primarily rely on Euclidean-distance-based encoding and operate independently of semantic information, which limits their discriminative capability in complex scenes. While EH-NLSC enhances similarity measurement through histogram intersection, the absence of Laplacian regularisation reduces its ability to effectively model spatial structure. Semantic-based methods like RSS incorporate label information at the representation level but still rely on unstable sparse coding during the encoding process, resulting in limited performance improvements.

Through the joint exploitation of structured local representations and globally discriminative semantic features, HS-LNLSC-DINO effectively bridges the gap between handcrafted sparse modelling and deep representation learning. This complementary integration results in more robust, informative, and semantically consistent image representations, leading to significant performance improvements over existing methods.

Table 10. Classification accuracy (mean ± standard deviation) across four benchmark image datasets (%); P-value versus second-best.

Methods	Corel-10	Scene-15	Caltech-101 (15)	Caltech-101 (30)	Caltech-256 (15)	Caltech-256 (30)	Caltech-256 (45)	Caltech-256 (60)
SCSPM	86.76± 1.18	81.12 ± 0.45	66.87 ± 0.45	72.10 ± 1.14	27.53 ± 0.42	33.86 ± 0.55	37.35 ± 1.64	40.08 ± 0.79
LLC	87.83 ± 1.03	81.53 ± 0.87	68.57 ± 0.88	72.54 ± 0.71	31.27 ± 0.85	34.17 ± 0.33	35.93 ± 0.51	37.58 ± 0.49
LSC	88.43± 0.75	89.65 ± 0.41	70.32 ± 1.35	74.86 ± 0.53	29.88 ± 0.15	35.67 ± 0.33	38.37 ± 0.46	40.35 ± 0.24
Lap-NMF-SPM	91.24 ± 0.95	90.46 ± 0.87	74.35 ± 0.94	76.81 ± 0.49	35.24 ± 0.83	37.46 ± 0.32	39.87 ± 0.75	41.35 ± 0.72
EH-NLSC	93.64 ± 0.78	91.82 ± 0.67	73.34 ± 0.62	78.89 ± 0.39	35.89 ± 0.56	38.87 ± 0.59	41.65 ± 0.53	43.61 ± 0.47
RSS	95.72 ± 0.78	92.45 ± 0.93	77.63 ± 0.89	82.91 ± 0.22	40.16 ± 0.53	44.96 ± 0.85	48.25 ± 0.47	51.32 ± 0.41
HS-NLLSC	98.86 ± 0.23	97.56 ± 0.64	81.54 ± 0.48	87.73 ± 0.67	46.34 ± 0.45	49.86 ± 0.65	53.60 ± 0.78	57.68 ± 0.12
HS-LNLSC-DINO	**99.84 ± 0.17**	**98.50 ± 0.12**	**97.63 ± 0.11**	**99.21 ± 0.16**	**87.24 ± 0.19**	**92.07 ± 0.16**	**94.86 ± 0.26**	**96.86 ± 0.11**
P-value (vs second-best)	5.95×10^{-3}	1.21	1.37×10^{-10}	2.95×10^{-8}	6.66×10^{-14}	7.11×10^{-13}	4.44×10^{-12}	0.00
Significant (P < 5%)	*	*	*	*	*	*	*	*

For clearer visualisation, Fig. 7 presents the mean classification accuracy and corresponding standard deviation across seven key methods over four benchmark datasets. HS-LNLSC-DINO consistently achieves the best performance. Compared with HS-NLLSC and other competing approaches, accuracy gains range from approximately 1% to 16% across the datasets.

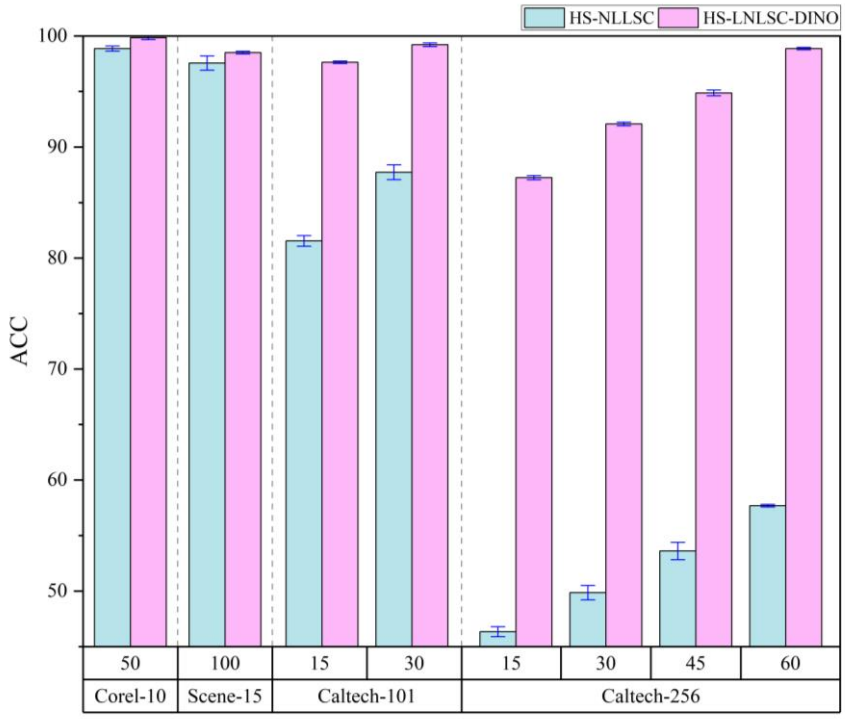


Fig. 7. Classification accuracy (mean ± standard deviation) across four benchmark image datasets.

With particular emphasis on ship classification, performance is evaluated on the Seaship, Seaship-trimming, and SMD-trimming datasets. The average classification accuracy across these three maritime datasets is reported in Table 11. For improved visual interpretation and ease of comparison, the corresponding numerical results are further depicted in Fig. 8.

Table 11. Classification accuracy (mean ± standard deviation) across the three maritime datasets (%).

Training Images		**50**	**100**	**150**	**200**	**250**
HS-NLLSC	Seaship	78.62±0.72	89.28±0.26	92.99±0.13	94.62±0.23	94.61±0.43
	Seaship-trimming	87.59±0.67	92.16±0.62	95.33±0.23	96.64±0.16	97.63±0.23
	SMD-trimming	93.22±0.36	96.41±0.41	97.73±0.29	98.97±0.15	99.28±0.08
HS-LNLSC-DINO	Seaship	**92.76±0.21**	**96.38±0.18**	**97.85±0.31**	**98.12±0.24**	**98.75±0.11**
	Seaship-trimming	**95.78±0.48**	**98.04±0.27**	**98.62±0.22**	**99.23±0.16**	**99.31±0.11**
	SMD-trimming	**94.90±0.42**	**97.35±0.28**	**98.65±0.17**	**99.13±0.17**	**99.68±0.07**

As presented in Table 11 and illustrated in Fig. 8, the proposed HS-LNLSC-DINO approach demonstrates consistently strong classification performance across all three maritime datasets. Compared to HS-NLLSC, the HS-LNLSC-DINO method achieves an accuracy improvement of approximately 1% to 15% as the number of training samples increases.

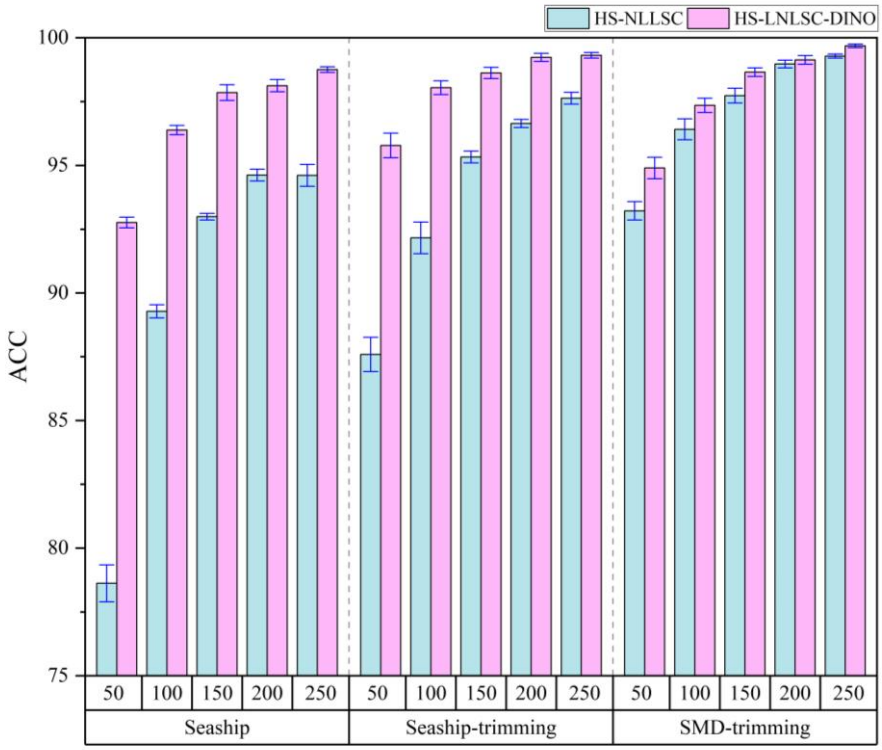


Fig. 8. Classification accuracy (mean ± standard deviation) across the three maritime datasets.

4.4. Generalisation and stability analysis

To further evaluate the robustness and generalisation capability of the proposed HS-LNLSC-DINO method, additional experiments are conducted from two perspectives: cross-dataset generalisation performance and random repeatability analysis.

4.4.1. Generalisation evaluation across datasets

Based on the results of the ablation studies in Section 4.2, the optimal configuration, specifically, the fusion of HS-LNLSC with DINO CLS features extracted from the L12 Transformer layer, is selected for all subsequent experiments. This configuration is consistently applied across all seven datasets, including four benchmark datasets and three maritime image datasets. The classification accuracies achieved on each dataset are presented in Table 12.

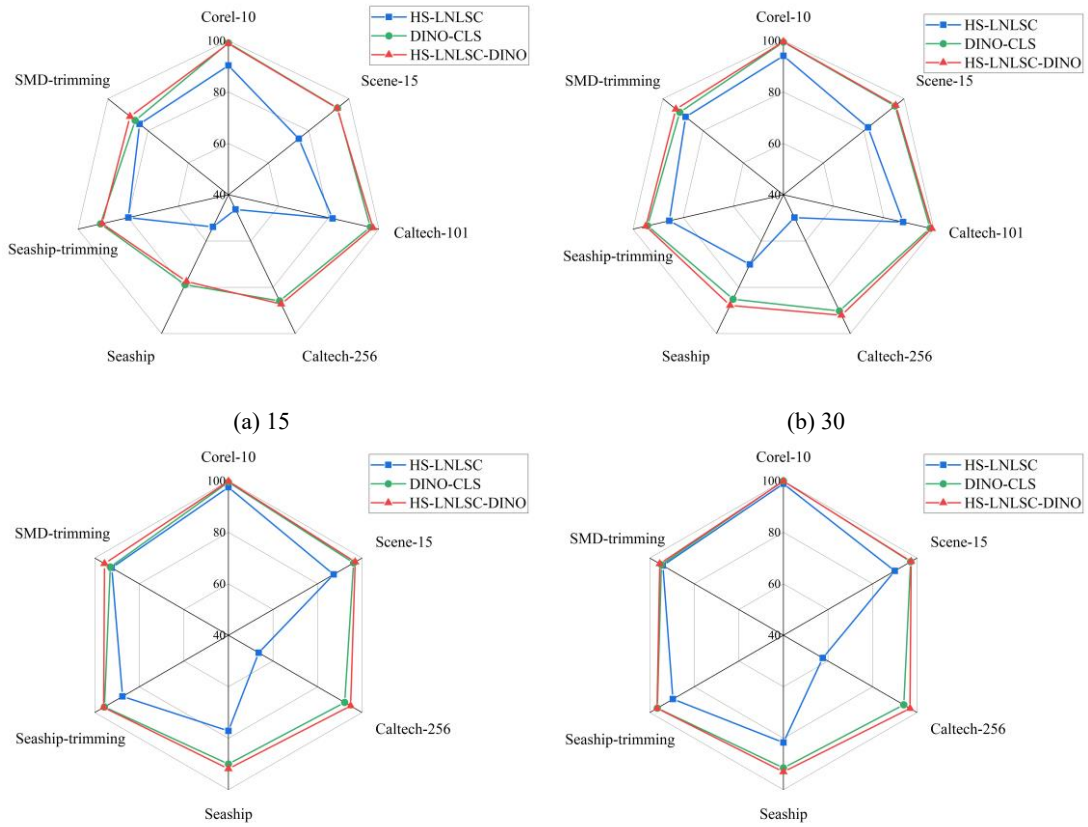
As presented in Table 12, the proposed HS-LNLSC-DINO framework consistently achieves strong performance across all datasets. Particularly notable advantages are observed on the maritime datasets, which are characterised by complex backgrounds, illumination variations, and significant scale changes, highlighting the robustness of the proposed framework in challenging visual environments. At the same time, competitive performance is maintained on standard datasets, indicating that the proposed hybrid representation is not limited to a specific domain. These results validate the effectiveness and broad applicability of integrating structured sparse representations with self-supervised semantic features.

To provide a more intuitive illustration of the results, Fig. 9 presents the classification accuracy of three methods, HS-LNLSC, DINO-CLS, and HS-LNLSC-DINO, across seven datasets. As shown in Fig. 9, the proposed HS-LNLSC-DINO method outperforms the other

two methods on all seven datasets, further demonstrating the robustness and generalisation capability of the proposed approach.

Table 12. Classification accuracy on the seven image datasets (%).

Datasets	Training Images			
	15	**30**	**45**	**60**
Corel-10	98.85	99.60	99.82	99.85
Scene-15	94.07	95.89	96.95	97.36
Caltech-101	97.63	99.21	/	/
Caltech-256	87.24	92.07	94.86	96.86
Seaship	77.50	87.82	91.87	93.04
Seaship-trimming	90.38	94.88	96.04	96.73
SMD-trimming	88.92	93.51	95.64	95.53



(a) 15 (b) 30

(c) 45 (d) 60

Fig. 9. Classification accuracy of the three methods on the seven image datasets.

4.4.2. Random repeatability and stability analysis

Beyond generalisation assessment, the stability of the proposed method is further evaluated through random repeatability experiments. Specifically, for the seven image datasets, 30 images from each category are randomly selected as training samples, with the remaining images used for testing. The average classification accuracy and the corresponding standard deviation are then calculated to assess the robustness of the method against data sampling variability. The experimental results from five random runs are shown in Table 13 and depicted in Fig. 10. Additionally, Fig. 11 displays the classification accuracy, along with the mean and standard deviation across the seven image datasets.

As shown in Table 13 and Figures 10 and 11, the proposed HS-LNLSC-DINO framework achieves not only high average classification accuracy but also relatively low standard deviation across repeated runs. This result demonstrates strong stability and limited sensitivity to random fluctuations in training sample selection. Taken together with the generalisation results, these findings confirm that HS-LNLSC-DINO exhibits robust and reliable performance. Through the effective integration of local geometric modelling and global semantic representation, the framework delivers consistent results across diverse datasets and experimental conditions, further supporting its suitability for practical image classification applications, particularly in marine and maritime environments.

While the Seaship dataset exhibits a slightly higher variance in classification accuracy (SD = 1.39%), this is largely attributable to the inherent characteristics of the dataset. Specifically, the images are relatively large, whereas the target objects occupy only a small portion of each scene, and the majority of pixels correspond to background or non-target regions. Consequently, forced resizing and feature extraction can lead to partial loss of target-specific features, which increases variability across experimental runs.

These observations are discussed here to highlight dataset-specific sensitivity, while simultaneously confirming the overall robustness and stability of the proposed HS-LNLSC-DINO framework across diverse maritime datasets.

Table 13. Stability analysis over five random runs on the seven image datasets (%).

Datasets	**Rounds**					**Mean Accuracy**	**Standard Deviation**
	1	**2**	**3**	**4**	**5**		
Corel-10	99.71	99.57	99.86	99.29	99.57	99.60	0.21
Scene-15	96.08	96.15	95.61	96.06	95.57	95.89	0.28
Caltech-101	99.01	99.15	99.17	99.36	99.39	99.21	0.16
Caltech-256	92.13	92.01	91.98	92.31	91.91	92.07	0.16
Seaship	89.45	86.51	86.57	89.12	87.49	87.82	1.39
Seaship-trimming	95.07	94.23	94.60	95.61	94.91	94.88	0.52
SMD-trimming	94.39	92.66	93.77	93.52	93.21	93.51	0.64

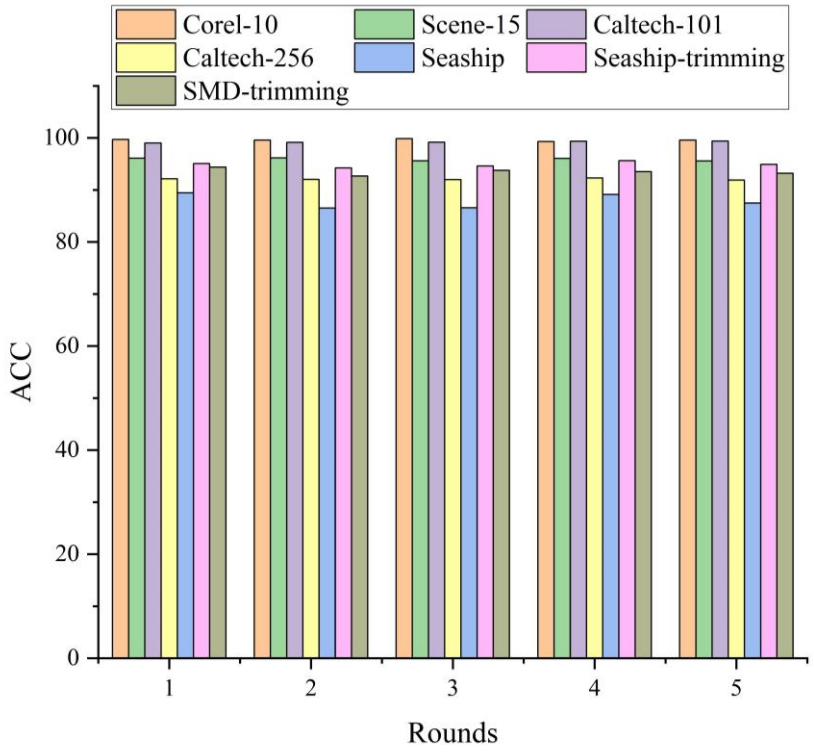


Fig. 10. Stability analysis over five random runs on the seven maritime image datasets.

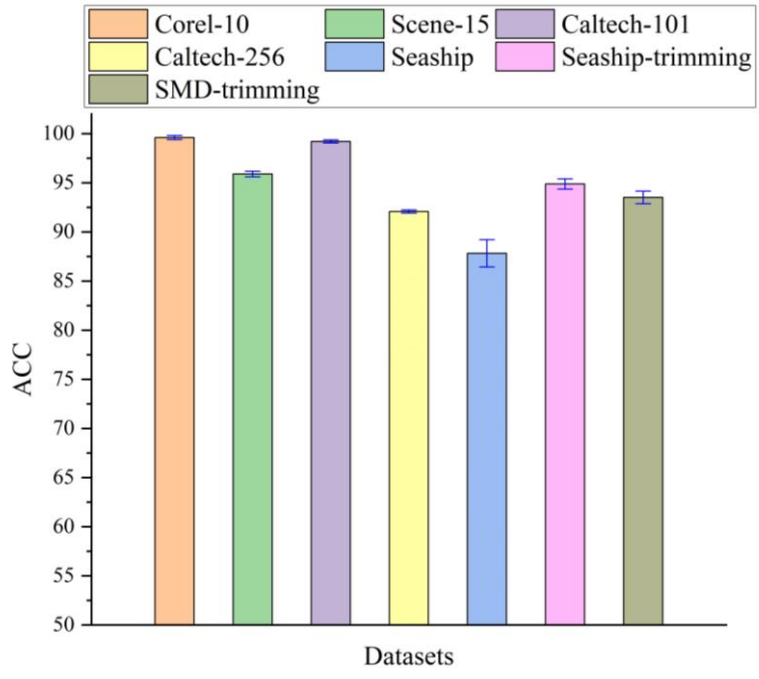


Fig. 11. Classification accuracy (mean ± standard deviation) on the seven image datasets.

4.5. Algorithm Complexity Analysis

Let n denote the number of local features per image, m the number of template features, and k the dictionary size. The computational complexity of the coding stage in LSC can be expressed as $O(n\times(m+k))$. After incorporating histogram intersection and locality constraints,

the complexity of the HI-LNLSC coding stage becomes $O(k^2)$. In the MP stage, given l pyramid levels and b histogram bins, the computational complexity is $O(n+l\times b)$. Accordingly, the total computational complexity of the HI-LNLSC stage can be summarised as $O(n+(m\times k)+k^2+l\times b)$. For the semantic information stage, let nR denote the number of cross-validation folds and C the number of categories in the dataset for SVM classification. The computational complexity of this stage can be expressed as $O(nR\times C)$. Accordingly, the overall computational complexity of the HS-LNLSC algorithm is given by $O(n+(m\times k)+k^2+l\times b+nR\times C)$.

For the DINO feature extraction stage, let p denote the number of image patches, d the token dimension, and t the number of Transformer layers. Since the pretrained DINO model is used in a frozen manner and only forward propagation is required, the computational complexity of this stage is given by $O(tp^2d)$.

The feature fusion stage consists of a simple weighted concatenation of HS-LNLSC and DINO features, whose computational cost is negligible compared with the feature extraction stages. So the complexity is $O(d_1+d_2)$. Finally, given H training samples, the linear SVM classification stage has a complexity of $O(H(d_1+d_2))$, where d_1 and d_2 denote the feature dimensions of HS-LNLSC and DINO representations, respectively. In summary, the overall computational complexity of the proposed HS-LNLSC-DINO framework can be expressed as $O(n+(m\times k)+k^2+l\times b+nR\times C+tp^2d+(d_1+d_2)+H(d_1+d_2))$.

Table 14 presents the complexity analysis of three different methods, including HS-LNLSC, DINO, and HS-LNLSC-DINO. As shown in Table 14, the proposed HS-LNLSC-DINO framework integrates both local sparse representations and global semantic features, and its overall complexity reflects the combination of these components. So the complexity is higher.

Although the hybrid framework incurs higher computational complexity compared with individual components, the additional cost is mainly concentrated in the feature extraction stage. The use of a frozen DINO backbone and a linear classifier ensures that the classification stage remains computationally efficient. Therefore, the overall computational burden remains manageable while enabling improved representation capability and classification performance in complex marine and coastal scenarios.

Table 14. The complexity analysis of three different methods.

Methods	**Complexity**
HS-LNLSC	$O(n+(m\times k)+k^2+l\times b+nR\times C)$
DINO	$O(tp^2d)$
HS-LNLSC-DINO	$O(n+(m\times k)+k^2+l\times b+nR\times C+tp^2d+(d_1+d_2)+H(d_1+d_2))$

5. Management and policy implications, and practical considerations

The findings of this study extend beyond methodological performance and offer important insights for the design, implementation, and governance of marine and coastal monitoring systems. By improving the reliability, interpretability, and scalability of image-based classification under visually complex and data-limited conditions, the proposed hybrid sparse-semantic framework provides practical support for evidence-based ocean and coastal management across diverse institutional and operational contexts.

5.1. Supporting evidence-based monitoring under data and capacity constraints

HS-LNLSC-DINO exhibits robust performance in low-sample and visually complex maritime scenarios, delivering stable and interpretable data outputs that enhance the reliability and consistency of routine marine monitoring operations. From both methodological and management perspectives, HS-LNLSC-DINO demonstrates robust performance in low-sample and visually complex maritime scenarios, maintaining stable and interpretable local feature representations. These enhancements not only strengthen the reliability of upstream image-based analyses but also improve the quality of data inputs for routine monitoring systems, particularly under data-constrained conditions where labelled samples are limited, thereby supporting more consistent and trustworthy operational monitoring. A persistent challenge in marine and coastal management is the uneven availability of labelled data across regions, activities, and monitoring objectives, often compounded by limited financial and technical resources. Many management authorities must prioritise surveillance and environmental monitoring under strict budgetary and operational constraints. The demonstrated effectiveness of the proposed framework when trained on small datasets suggests that automated image classification can be more feasibly integrated into routine monitoring programmes without requiring extensive and continuous annotation efforts. This capability enables management agencies to extract actionable information from existing image data streams, including satellite and UAV imagery, even in data-poor or newly monitored areas.

From a policy perspective, reduced dependence on labelled data is particularly relevant for adaptive management frameworks, where monitoring priorities may evolve in response to emerging environmental risks, regulatory changes, or shifts in maritime activity patterns. Faster deployment and updating of monitoring systems can support more timely regulatory responses and flexible implementation of monitoring mandates at local, national, and regional scales.

5.2. Enhancing robustness for regulatory and governance applications

The hybrid integration of structured local descriptors with global DINO features enables consistent, reliable identification of vessel types and environmental features, thereby supporting maritime traffic supervision and compliance-oriented monitoring workflows. In regulatory and operational contexts, the hybrid integration of structured local sparse descriptors and global DINO features provides robust, accurate, and consistent classification performance, enabling reliable identification of maritime targets, including vessel types and environmental features, and thereby supporting surveillance and compliance-related applications. Marine and coastal imagery increasingly underpins regulatory functions related to maritime traffic management,

port operations, environmental protection, and safety enforcement. However, classification errors caused by reflections, haze, cluttered backgrounds, and small targets can undermine regulatory confidence in automated monitoring outputs. The improved robustness observed across challenging maritime datasets indicates that combining structured local encoding with global self-supervised semantic representations can mitigate these sources of error, supporting more reliable surveillance outputs for governance-relevant tasks such as vessel-type monitoring, nearshore activity assessment, and long-term evaluation of human pressures on coastal and marine environments.

More consistent and reliable classification results can contribute to fairer and more transparent enforcement practices by reducing both false positives that impose unnecessary regulatory burdens and false negatives that weaken compliance objectives. In this way, the proposed framework aligns technical performance improvements with broader governance goals.

5.3. Strengthening transparency, accountability, and institutional trust

Beyond classification accuracy, the interpretability of HS-LNLSC-DINO outputs provides traceable and transparent feature representations, thereby enhancing auditability and fostering institutional trust in marine and coastal governance. These practical validations underscore the added value of the hybrid sparse-semantic framework in operationally relevant management contexts. Transparency and accountability are central principles of contemporary ocean and coastal governance, particularly where automated systems inform regulatory actions, risk assessments, or public reporting. Compared with purely deep learning-based approaches, the structured sparse encoding component of the proposed framework provides interpretable representations of local feature contributions, improving the traceability of classification outcomes. Such interpretability supports internal audits, post-event evaluations, and external scrutiny when monitoring outputs are used to justify management interventions or regulatory decisions.

From an institutional standpoint, enhanced transparency can strengthen trust in automated monitoring systems among regulators, port authorities, coastal managers, and other stakeholders. This is especially important in governance settings where decisions must be communicated across organisational boundaries or defended in legal, administrative, or public forums.

5.4. Practical deployment considerations and future policy-relevant directions

The proposed framework adopts a modular design that combines frozen self-supervised features with a lightweight classifier, limiting training complexity and reducing the risk of overfitting. This design makes the framework suitable for deployment in operational settings where computational resources and technical capacity may be constrained. In practice, the feature extraction and classification pipeline can be integrated as a modular component within broader monitoring systems, including satellite and UAV data ingestion, image categorisation workflows, and event summarisation platforms.

Nevertheless, several considerations remain relevant for policy- and management-oriented deployment. In practical monitoring workflows, geographic variability, seasonal changes, and sensor heterogeneity can affect classification performance. While HS-LNLSC-DINO demonstrates robustness across multiple datasets, systematic evaluation under these domain-shift conditions remains necessary. The current study does not include cross-region transfer, seasonal robustness, or sensor-shift experiments, nor does it address open-set recognition or uncertainty-aware prediction. In addition, the current evaluation focuses primarily on image-level classification rather than downstream decision outcomes, such as risk prioritisation or regulatory effectiveness. Finally, operational decision making often requires explicit representation of uncertainty, particularly in risk-sensitive contexts related to safety, environmental protection, and enforcement.

Addressing these limitations represents an important direction for future research. Incorporating cross-region adaptation, open-set recognition, and uncertainty-aware classification could further enhance the policy relevance of automated image analysis, supporting risk-informed decision making and precautionary management approaches. Such developments would strengthen the alignment of image-based monitoring tools with the principles of adaptive, transparent, and accountable ocean and coastal governance.

Overall, the proposed framework contributes to marine and coastal management in an indirect yet structured manner by reinforcing the upstream data interpretation layer within monitoring systems. This enhanced data foundation supports downstream processes, including indicator construction, surveillance analysis, and decision-support activities, thereby establishing a coherent pathway from image analysis to governance applications. It should be noted, however, that the current evaluation primarily focuses on classification performance at the image level, rather than on downstream decision-making outcomes such as risk prioritisation, regulatory effectiveness, or operational decisions.

6. Conclusions

This paper developed a hybrid sparse-semantic image classification framework (HS-LNLSC-DINO) to support robust marine and coastal image interpretation in operationally realistic and data-constrained settings. By integrating structured local representations derived from HS-LNLSC sparse coding with global semantic features extracted from a self-supervised DINO Vision Transformer, the proposed framework bridges fine-grained geometric modelling and high-level semantic understanding. This unified representation is well-suited to the visual complexity commonly encountered in marine and coastal environments, including variable illumination, background clutter, and limited availability of labelled data.

From a methodological standpoint, the proposed framework enhances the stability and interpretability of sparse coding by jointly preserving locality, non-negativity, and neighbourhood consistency, while introducing semantic guidance to improve discriminative capability. In parallel, self-supervised DINO features provide transferable global semantics without reliance on extensive labelled datasets. The feature-level fusion of these complementary representations, combined with a lightweight linear classifier, results in a

practical and computationally efficient analytical pipeline. Such characteristics are particularly relevant for marine and coastal monitoring systems that must operate under resource constraints and integrate with existing observational infrastructures.

Extensive evaluation on four public benchmark datasets and three maritime image datasets demonstrates consistent and robust performance improvements compared with representative sparse coding-based, semantic-aware, and deep-feature baselines. The proposed framework achieves stable gains across diverse conditions, including challenging maritime scenarios, and remains effective under limited training data. Ablation analyses further confirm the complementary value of structured local modelling and self-supervised semantic representations, highlighting the importance of combining interpretability and semantic richness in complex marine image classification tasks.

Beyond methodological performance, the findings have direct implications for marine and coastal management practice. More reliable and interpretable image classification can strengthen environmental monitoring, maritime traffic observation, and coastal surveillance by improving the consistency and transparency of automated analytical outputs. In management and governance contexts where monitoring results inform regulatory oversight, operational decision making, or stakeholder communication, the ability to balance robustness, data efficiency, and interpretability is particularly important. The proposed framework therefore contributes to the broader objective of enabling evidence-based ocean and coastal management through improved utilisation of visual data.

Several directions for future research emerge from this work. It should be emphasised that the primary contribution of this study lies in enhancing the reliability and interpretability of upstream image-based analyses within marine and coastal monitoring systems. While these improvements strengthen the foundation for downstream decision-support activities, evaluation of decision-level outcomes-such as risk prioritisation, regulatory effectiveness, or decision-support workflows-remains beyond the scope of the current work and is identified as a key focus for future research. First, extending the framework towards end-to-end optimisation while retaining interpretability could further enhance operational performance. Second, incorporating adaptive and uncertainty-aware fusion strategies would better support upstream information extraction within monitoring and decision-support workflows under domain shifts associated with seasonal variation, geographic heterogeneity, and sensor diversity. Future research will further investigate cross-region transfer, seasonal generalisation, sensor-shift adaptation, open-set recognition, and uncertainty-aware prediction to enhance robustness and reliability in heterogeneous, real-world monitoring environments. Finally, future studies should expand evaluation beyond image-level classification to management-oriented applications, such as fine-grained vessel categorisation, maritime scene understanding, and cross-region or cross-season generalisation within real-world monitoring workflows. Advancing along these directions will further strengthen the contribution of image-based analytics to resilient, transparent, and adaptive marine and coastal management.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements

This work presented in this study is financially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 42407114 and No. 52571407.

Data availability

https://github.com/shiying820212/HS-LNLSC-DINO.git.

References

Abu Bakar, N.N., Bazmohammadi, N., Vasquez, J.C., Guerrero, J.M., 2025. Seaside port operation optimization and energy management system with integrated seaport microgrid and cold ironing. Next Energy 9, 100439. https://doi.org/10.1016/j.nxener.2025.100439

Adetunji, A.S., Vasanthan, C., Glomsrud, J.A., Galeazzi, R., Rokseth, B., 2025. Safety Assurance for Autonomous Ships Using Contract-Based Design and Simulation-Based Testing. IFAC-PapersOnLine, 16th IFAC Conference on Control Applications in Marine Systems, Robotics and Vehicles CAMS 2025 59, 722–727. https://doi.org/10.1016/j.ifacol.2025.11.720

Agrillo, E., Alessi, N., Angelini, P., Buffi, F., Carli, E., Casella, L., Cutini, M., Filipponi, F., Fratarcangeli, C., Massimi, M., Mercatini, A., Pezzarossa, A., Pretto, F., Sarmati, S., Tartaglione, N., Attorre, F., 2025. Enhancing Natura 2000 habitat monitoring: A framework for biodiversity conservation assessment. Ecological Indicators 181, 114403. https://doi.org/10.1016/j.ecolind.2025.114403

Bakirci, M., 2025. Advanced ship detection and ocean monitoring with satellite imagery and deep learning for marine science applications. Regional Studies in Marine Science 81, 103975. https://doi.org/10.1016/j.rsma.2024.103975

Cai, D., He, X., Han, J., Huang, T.S., 2011. Graph Regularized Nonnegative Matrix Factorization for Data Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 1548–1560. https://doi.org/10.1109/TPAMI.2010.231

Cao, C., Yi, H., Xiang, H., He, P., Hu, J., Xiao, F., Gao, X., 2024. Accelerated Sparse-Coding-Inspired Feedback Neural Architecture Search for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing 62, 1–14. https://doi.org/10.1109/TGRS.2024.3363777

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660.

Chen, H., Xie, K., Wang, H., Zhao, C., 2018. Scene image classification using locality-constrained linear coding based on histogram intersection. Multimed Tools Appl 77, 4081–4092. https://doi.org/10.1007/s11042-017-4830-7

Cheng, C., Peng, J., Cui, W., 2023. A Two-Stage Convolutional Sparse Coding Network for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters 20, 1–5. https://doi.org/10.1109/LGRS.2023.3245210

Dong, S., Feng, J., Fang, D., 2024. A Novel Multiscale Contrastive Learning Network for Fine-Grained Ocean Ship Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17, 9989–10005. https://doi.org/10.1109/JSTARS.2024.3399310

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

Galgani, F., Lusher, A.L., Strand, J., Haarr, M.L., Vinci, M., Molina Jack, E., Kagi, R., Aliani, S., Herzke, D., Nikiforov, V., Primpke, S., Schmidt, N., Fabres, J., De Witte, B., Solbakken, V.S., van Bavel, B., 2024. Revisiting the strategy for marine litter monitoring within the european marine strategy framework directive (MSFD). Ocean & Coastal Management 255, 107254. https://doi.org/10.1016/j.ocecoaman.2024.107254

Gao, S., Tsang, I.W.-H., Chia, L.-T., Zhao, P., 2010. Local features are not lonely – Laplacian sparse coding for image classification, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Presented at the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3555–3561. https://doi.org/10.1109/CVPR.2010.5539943

Ghaban, W., Ahmad, J., Siddique, A.A., Alshehri, M.S., Saghir, A., Saeed, F., Ghaleb, B., Rehman, M.U., 2025. Sustainable Environmental Monitoring: Multistage Fusion Algorithm for Remotely Sensed Underwater Super-Resolution Image Enhancement and Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18, 3640–3653. https://doi.org/10.1109/JSTARS.2024.3522202

Graves, A., 2012. Long Short-Term Memory, in: Supervised Sequence Labelling with Recurrent Neural Networks, Studies in Computational Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 37–45. https://doi.org/10.1007/978-3-642-24797-2_4

Han, H., Liu, S., Gan, L., 2015. Non-negativity and dependence constrained sparse coding for image classification. Journal of Visual Communication and Image Representation 26, 247–254. https://doi.org/10.1016/j.jvcir.2014.12.002

He, J., Huang, M., Wang, Y., Zhang, Y., Yu, H., 2025. A holistic approach to resilient port management: Synergizing space template and crane deployment via benders decomposition algorithm. Ocean & Coastal Management 269, 107806. https://doi.org/10.1016/j.ocecoaman.2025.107806

He, L., Zhou, Y., Yang, H., Su, L., Ma, J., 2026. A deep learning-based method for marine oil spill detection and its application in UAV imagery. Marine Pollution Bulletin 222, 118889. https://doi.org/10.1016/j.marpolbul.2025.118889

Hoyer, P.O., 2002. Non-negative sparse coding, in: Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing. Presented at the the 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 557–565. https://doi.org/10.1109/NNSP.2002.1030067

Kanjir, U., Greidanus, H., Oštir, K., 2018. Vessel detection and classification from spaceborne optical images: A literature survey. Remote Sensing of Environment 207, 1–26. https://doi.org/10.1016/j.rse.2017.12.033

Khatri, T.K., Wei, K.T., Sharif, K.Y., 2026. An optimized deep learning framework for detecting and localizing mooring lines in marine images using Faster R-CNN. Ocean Engineering 343, 123559. https://doi.org/10.1016/j.oceaneng.2025.123559

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Presented at the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), pp. 2169–2178. https://doi.org/10.1109/CVPR.2006.68

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791. https://doi.org/10.1038/44565

Lin, Z., Liu, R., Su, Z., 2011. Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., Bach, F., 2008. Supervised Dictionary Learning, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Min, H., Liang, M., Luo, R., Zhu, J., 2016. Laplacian regularized locality-constrained coding for image classification. Neurocomputing 171, 1486–1495. https://doi.org/10.1016/j.neucom.2015.07.084

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the Fisher Kernel for Large-Scale Image Classification, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010. Springer, Berlin, Heidelberg, pp. 143–156. https://doi.org/10.1007/978-3-642-15561-1_11

Puskic, P.S., Cramer, I., Church, E., Deery, E., Egger, M., Fox, N., de Haan, W.P., Lebreton, L., Liconti, A., Sanchez-Vidal, A., Wolter, H., 2026. Highly engaged marine users can help monitor marine plastic pollution in under accessed environments. Ocean & Coastal Management 272, 107983. https://doi.org/10.1016/j.ocecoaman.2025.107983

Rangel-Buitrago, N., Giarrizzo, T., Brabo, L., Silva Filho, F.J., Cooper, J.A.G., Neal, W.J., 2026. Updating coastal beach classification: A cluster-based typology for contemporary human

use and management. Ocean & Coastal Management 273, 108064. https://doi.org/10.1016/j.ocecoaman.2025.108064

Rasiwasia, N., Vasconcelos, N., 2008a. Scene classification with low-dimensional semantic spaces and weak supervision, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6. https://doi.org/10.1109/CVPR.2008.4587372

Rasiwasia, N., Vasconcelos, N., 2008b. Image retrieval using query by contextual example, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08. Association for Computing Machinery, New York, NY, USA, pp. 164–171. https://doi.org/10.1145/1460096.1460124

Ruan, H., Xu, Z., Yang, Z., Lu, Y., Qin, J., Chen, T., 2025. Learning Semantic-aware Representation in Visual-Language Models for Multi-label Recognition with Partial Labels. ACM Trans. Multimedia Comput. Commun. Appl. 21, 79:1-79:19. https://doi.org/10.1145/3708991

Shen, F., Zeng, G., 2019. Semantic image segmentation via guidance of image classification. Neurocomputing 330, 259–266. https://doi.org/10.1016/j.neucom.2018.11.027

Shi, J., Chen, J., Wan, Z., Zhou, S., Jun, Y., Shu, Y., 2025. The impact of low-sulfur marine fuel policy on air pollution in global coastal cities. Sustainable Horizons 14, 100130. https://doi.org/10.1016/j.horiz.2024.100130

Shi, W., Zheng, W., Xu, Z., 2025. Ship-Yolo: A Deep Learning Approach for Ship Detection in Remote Sensing Images. Journal of Marine Science and Engineering 13. https://doi.org/10.3390/jmse13040737

Shi, Y., Wan, Y., Wang, X., Li, H., 2025. Incorporation of Histogram Intersection and Semantic Information into Non-Negative Local Laplacian Sparse Coding for Image Classification. Mathematics 13, 219. https://doi.org/10.3390/math13020219

Shu, Z., 2026. Hybrid quantum sparse coding and dynamic convolution capsule network for enhanced image classification. Pattern Recognition 169, 111974. https://doi.org/10.1016/j.patcog.2025.111974

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

Smirnov, E.A., Timoshenko, D.M., Andrianov, S.N., 2014. Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks. AASRI Procedia, 2nd AASRI Conference on Computational Intelligence and Bioinformatics 6, 89–94. https://doi.org/10.1016/j.aasri.2014.05.013

Teixeira, E.H., Mafra, S.B., De Figueiredo, F.A.P., 2025. InaTechShips: A validation study of a novel ship dataset through deep learning-based classification and detection models for maritime applications. Ocean Engineering 326, 120823. https://doi.org/10.1016/j.oceaneng.2025.120823

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H., 2021. Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th

International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 10347–10357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. ukasz, Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Vedaldi, A., Zisserman, A., 2012. Efficient Additive Kernels via Explicit Feature Maps. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 480–492. https://doi.org/10.1109/TPAMI.2011.153

Vehmaa, A., Lanari, M., Jutila, H., Mussaari, M., Pätsch, R., Telenius, A., Banta, G., Eklöf, J., Jensen, K., Krause-Jensen, D., Quintana, C.O., von Numers, M., Boström, C., 2024. Harmonization of Nordic coastal marsh habitat classification benefits conservation and management. Ocean & Coastal Management 252, 107104. https://doi.org/10.1016/j.ocecoaman.2024.107104

Wang, H., Han, B., Shu, C., Ouyang, Z., 2025. An integrated assessment of urban habitat quality based on the InVEST–IUEMS model. Ecological Frontiers. https://doi.org/10.1016/j.ecofro.2025.10.011

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained Linear Coding for image classification, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Presented at the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3360–3367. https://doi.org/10.1109/CVPR.2010.5540018

Wang, Q., Zhang, H., Xi, S., 2024. China's law and policy framework for maritime safety regulation of alternative fuel ships in the decarbonization transition. Marine Policy 163, 106142. https://doi.org/10.1016/j.marpol.2024.106142

Wu, J., Rehg, J.M., 2009. Beyond the Euclidean distance: Creating effective visual codebooks using the Histogram Intersection Kernel, in: 2009 IEEE 12th International Conference on Computer Vision. Presented at the 2009 IEEE 12th International Conference on Computer Vision, pp. 630–637. https://doi.org/10.1109/ICCV.2009.5459178

Xiao, X., Wang, P., Ge, Y., Luo, J., Chen, H., He, Y., Zhang, D., Li, Y., Fang, C., Lin, H., 2025. GeoKG-HSA: A framework for habitat suitability assessment with geospatial knowledge graphs. International Journal of Applied Earth Observation and Geoinformation 144, 104921. https://doi.org/10.1016/j.jag.2025.104921

Xie, R.-D., He, Z.-F., Li, B., Liu, B., Hu, J.-Y., 2025. Semantic-Aware Representation Learning via Conditional Transport for Multi-Label Image Classification. https://doi.org/10.48550/arXiv.2507.14918

Xu, L., Li, X., Yan, R., Chen, J., 2025a. How to support shore-to-ship electricity constructions: Tradeoff between government subsidy and port competition. Transportation Research Part E: Logistics and Transportation Review 201, 104258. https://doi.org/10.1016/j.tre.2025.104258

Xu, L., Shen, C., Chen, J., 2025b. The impact of the Maritime Silk Road Initiative on the carbon intensity of the participating countries. Maritime Economics & Logistics 27, 331–349. https://doi.org/10.1057/s41278-024-00295-z

Xu, L., Wu, J., Yan, R., Chen, J., Fu, S., 2025c. Who predicts better? A comparison of machine learning and econometrics in forecasting CO2 emissions from global shipping. Energy 338, 138967. https://doi.org/10.1016/j.energy.2025.138967

Yang, J., Rao, Y., Fan, H., Dong, J., Yu, H., 2025. Learning Semantic-Aware Point-Line Features for Localization and Reconstruction. IEEE Transactions on Circuits and Systems for Video Technology 35, 11783–11796. https://doi.org/10.1109/TCSVT.2025.3578998

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801. https://doi.org/10.1109/CVPR.2009.5206757

Yang, K., Yang, T., Yao, Y., Fan, S., 2021. A transfer learning-based convolutional neural network and its novel application in ship spare-parts classification. Ocean & Coastal Management 215, 105971. https://doi.org/10.1016/j.ocecoaman.2021.105971

Yang, X., Zeng, Z., Yang, D., 2024. Adaptive Mid-Level Feature Attention Learning for Fine-Grained Ship Classification in Optical Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 62, 1–10. https://doi.org/10.1109/TGRS.2024.3351874

Yorulmaz, M., Susoy, M., 2025. Risk analysis and management for STS operations in ports using an integrated hybrid method. Ocean Engineering 316, 120019. https://doi.org/10.1016/j.oceaneng.2024.120019

Yuan, W. a. N., Jinghui, Z., Zhiping, C., Xiaojing, M., 2019. Non-negative local sparse coding algorithm based on elastic net and histogram intersection. Journal of Computer Applications 39, 706. https://doi.org/10.11772/j.issn.1001-9081.2018071483

Zhang, C., Liu, J., Tian, Q., Liang, C., Huang, Q., 2013. Beyond visual features: A weak semantic image representation using exemplar classifiers for classification. Neurocomputing, Image Feature Detection and Description 120, 318–324. https://doi.org/10.1016/j.neucom.2012.07.056

Zhang, C., Zhu, X., Li, L., Zhang, Y., Liu, J., Huang, Q., Tian, Q., 2015. Joint image representation and classification in random semantic spaces. Neurocomputing 156, 79–85. https://doi.org/10.1016/j.neucom.2014.12.083

Zhang, L., Chen, Z., Zheng, M., He, X., 2011. Robust non-negative matrix factorization. Front. Electr. Electron. Eng. China 6, 192–200. https://doi.org/10.1007/s11460-011-0128-0

Zhang, W., Chen, W., Wu, X., Chen, J., Zhang, Z., 2025. Evaluation and spatiotemporal analysis of low-carbon efficiency of the “ship-port” system. Maritime Policy & Management 1–33. https://doi.org/10.1080/03088839.2025.2514765

Appendix A

Algorithm 2 (Apply an iterative process to Λ in order to compute D)

Input: Non-negative feature matrix X; Sparse coding S; Precision δ.

Output: Diagonal matrix Λ ; Dictionary *D*.

Step 1. Initialise: Set Λ^0, $i=0$, and δ as the initial precision.

Step 2. Convergence Check: While convergence criteria are not met, proceed with the following steps:

Step 3. Set $h_0 = \nabla f(\Lambda^0)$, if $\|h_0\|_2 \le \delta$ then

Step 4. Λ^0 denotes the desired extremum.

Step 5. else

Step 6. d^0=-h_0 .

Step 7. end if

Step 8. Let $h_{i+1} = \nabla f(\Lambda^{i+1})$, if $\|h_{i+1}\|_2 \le \delta$ or $\left\|\Lambda^{i+1} - \Lambda^i\right\|_F \le \delta \left\|\Lambda^1 - \Lambda^0\right\|_F$ then

Step 9. Λ^{i+1} denotes the desired extremum.

Step 10. else

Step 11. d^{i+1}=-$h_{i+1} + \beta_i d^i$, $\beta_i = \|h_{i+1}\|_2^2 / \|h_i\|_2^2$.

Step 12. end if

Step 13. Identify the most suitable step size ξ_i through a rough search method along with one dimension, i.e., $f(\Lambda^i + \xi_i d^i) = \min_{\xi} f(\Lambda^i + \xi d^i)$.

Step 14. $\Lambda^{i+1} = \Lambda^i + \xi_i d^i$.

Step 15. Let $i = i+1$,go back to Step 8.

Step 16. end while

Step 17. Go back to Λ , and acquire dictionary *D*using Eq. (6).

Appendix B

Algorithm 3 (The most suitable approximation D^* for appropriate sparseness of *D*)

Input: A randomly sampled column vector *d* from matrix *D*.

Output: The closest non-negative vector d^* within D^*.

Step 1. Calculate the sparseness S_{d^*} for the column vector d^* using Eq. $sparseness(d^*) = \dfrac{\sqrt{M} - \|d^*\|_1 / \|d^*\|_2}{\sqrt{M} - 1}$ and $l_1 = \|d^*\|_1 / \|d^*\|_2 = \sqrt{M} - (\sqrt{M} - 1)S_{d^*}$ .Here, *M* denotes the dimension of d^*.

Step 2. Project the vector *d* into the space defined by the l_1 constraint, $d_k^* = d_k + \dfrac{l_1 - \|d\|_1}{M}, \forall k$, i.e., let $\|d^*\|_1 = l_1$.

Step 3. Initialise a set $W = \{\}$ containing the indices of negative elements within *d*.

Step 4. While the vector d^* does not meet the non-negativity constraint and the required sparseness condition, namely, $\|d^*\|_2^2 = \|d\|_2^2$, continue the iteration process:

Step 5. Compute the midpoint $m_k = \begin{cases} \dfrac{l_1}{M - length(W)}, i \notin W \\ 0, k \in W \end{cases}$ within the non-negative space defined by the l_1 constraint.

Step 6. Solve the quadratic Eq. $\|d^*\|_2^2 = \|m + \alpha(d^* - m)\|_2^2$ to get the non-negative solution α , and update d^* according to Eq. $d^* = m + \alpha(d^* - m)$.

Step 7. If all elements of d^* are non-negative proceed to Step 8. Otherwise, go to Step 9.

Step 8. Go back to d^*.

Step 9. If any elements of d^* remain negative:

Step 10. For each negative element, set it to zero via $W = W \bigcup \{k : d_k^* < 0\}$, and set $d_k^* = 0, \forall k \in W$.

Step 11. Recalculate the projection of d^*, ensuring d^* invariant within the non-negative space defined by the l_1 constraint, i.e., $d_k^* = d_k^* + (l_1 - \|d^*\|_1) / (M - length(W))$.

Step 12. Repeat the process by returning to Step 5.

Step 13. Continue iterating until the vector d^* satisfies both the non-negativity and sparseness conditions.