Journal of *Marine Science and Engineering*

MDPI

Article

VIOS-Net: A Multi-Task Fusion System for Maritime Surveillance Through Visible and Infrared Imaging

Junquan Zhan [1,†], Jiawen Li [1,2,3,†], Langtao Wu [1], Jiahua Sun [1] and Hui Yin [1,2,3,*]

[1] Naval Architecture and Shipping College, Guangdong Ocean University, Zhanjiang 524005, China; junquanzhan@stu.gdou.edu.cn (J.Z.); jiawen-li@gdou.edu.cn (J.L.); wulangtao2@stu.gdou.edu.cn (L.W.); jiahuasun@stu.gdou.edu.cn (J.S.)
[2] Technical Research Center for Ship Intelligence and Safety Engineering of Guangdong Province, Zhanjiang 524005, China
[3] Guangdong Provincial Key Laboratory of Intelligent Equipment for South China Sea Marine Ranching, Zhanjiang 524005, China
* Correspondence: huiyin@gdou.edu.cn
† These authors contributed equally to this work and should be considered co-first authors.

Abstract: Automatic ship monitoring models leveraging image recognition have become integral to regulatory applications within maritime management, with multi-source image co-monitoring serving as the primary method for achieving comprehensive, round-the-clock surveillance. Despite their widespread use, the existing models predominantly train each data source independently or simultaneously train multiple sources without fully optimizing the integration of similar information. This approach, while capable of all-weather detection, results in the underutilization of data features from related sources and unnecessary repetition in model training, leading to excessive time consumption. To address these inefficiencies, this paper introduces a novel multi-task learning framework designed to enhance the utilization of data features from diverse information sources, thereby reducing training time, lowering costs, and improving recognition accuracy. The proposed model, VIOS-Net, integrates the advantages of both visible and infrared data sources to meet the challenges of all-weather, all-day ship monitoring under complex environmental conditions. VIOS-Net employs a Shared Bottom network architecture, utilizing both shared and specific feature extraction modules at the model's lower and upper layers, respectively, to optimize the system's recognition capabilities and maximize data utilization efficiency. The experimental results demonstrate that VIOS-Net achieves an accuracy of 96.20% across both visible and infrared spectral datasets, significantly outperforming the baseline ResNet-34 model, which attained accuracies of only 4.86% and 9.04% in visible and infrared data, respectively. Moreover, VIOS-Net reduces the number of parameters by 48.82% compared to the baseline, achieving optimal performance in multi-spectral ship monitoring. Extensive ablation studies further validate the effectiveness of the individual modules within the proposed framework.

Keywords: ship classification; deep learning; multi-task learning; maritime target recognition; all-weather all-day; data fusion; feature sharing

check for updates

Received: 23 March 2025
Revised: 27 April 2025
Accepted: 29 April 2025
Published: 6 May 2025

Citation: Zhan, J.; Li, J.; Wu, L.; Sun, J.; Yin, H. VIOS-Net: A Multi-Task Fusion System for Maritime Surveillance Through Visible and Infrared Imaging. *J. Mar. Sci. Eng.* **2025**, *13*, 913. https://doi.org/10.3390/jmse13050913

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ship monitoring constitutes a critical component of ship management [1–5], playing a pivotal role in ensuring maritime safety, environmental protection, resource management, operational efficiency, and the prevention of criminal activities [6–12]. Effective ship monitoring is essential for safeguarding maritime safety, as it enables the timely detection of

maritime accidents and violations through the intelligent monitoring and identification of vessel targets, thereby preventing cargo mishaps and other maritime incidents. This function is particularly crucial in increasingly complex meteorological and marine environments, where the ability to capture abnormal behaviors and dangerous situations through early warning and preventive measures becomes even more pronounced. Moreover, ship monitoring contributes significantly to marine environmental protection by enabling the control and monitoring of pollutant emissions from vessels. This capability helps reduce marine pollution and preserves the health of the marine ecosystem. By facilitating the real-time monitoring and statistical analysis of pollutant discharges, such as exhaust gases and sewage, ship monitoring supports regulatory authorities in taking timely actions to mitigate environmental contamination. Ship monitoring also serves as an effective tool for the real-time management of marine resources, including fisheries, oil and gas, and mineral resources. By tracking the number and regional distribution of fishing vessels, it aids in the sustainable development and conservation of fishery resources. Additionally, monitoring vessel transport routes and cargo types enables the efficient management of marine mineral and oil and gas resources. In the context of port operations, ship monitoring provides a foundation for enhancing operational efficiency. It enables the accurate tracking and identification of vessel positions within ports, thereby equipping port management personnel with the necessary information for informed decision making and scheduling. This capability is instrumental in improving both port management and transportation efficiency. Furthermore, ship monitoring plays a crucial role in maritime law enforcement and crime prevention. Through vessel identification and route tracking, it assists maritime law enforcement officers in detecting and addressing violations and criminal activities, such as illegal immigration, drug trafficking, and smuggling, thereby enhancing the effectiveness of investigations and interventions. In maritime safety supervision, the real-time monitoring and positioning of vessels help prevent dangerous forced landings, thereby safeguarding lives and property.

With the continuous advancement of science and technology, ship monitoring based on sensor imaging (including satellite imagery, UAV-based aerial photography, and coastal CCTV systems) has emerged as a primary method for achieving all-weather, real-time detection of maritime conditions. While our deep learning model demonstrates strong adaptability across these heterogeneous image sources through transfer learning techniques, its performance remains contingent upon image resolution quality and stable data transmission conditions. The proposed framework is suitable for port surveillance systems with dedicated computing infrastructure, though deployment on resource-constrained edge devices may require additional optimization.

In the realm of ship image monitoring, multiple information sources are utilized to ensure comprehensive real-time surveillance, with visible light, infrared imagery, and synthetic aperture radar (SAR) being the most commonly employed technologies [13]. These sensors, operating on different imaging principles, generate image data with distinct characteristics.

Visible light sensors excel in capturing color, texture, and shape information, resulting in high-resolution images that offer a vivid, intuitive representation of the maritime environment. This capability enables the clear identification of vessels and their spatial relationships within the scene. In contrast, infrared imaging operates by detecting radiation differences between the target and its background, making it less susceptible to variations in lighting conditions. This unique imaging mechanism allows infrared sensors to penetrate certain materials, such as fog and smoke, making them particularly effective for night vision, target detection in extreme environments, and thermal signature identification [14]. On the other hand, SAR operates independently of external light sources by emitting mi-

crowave signals and receiving the signals reflected from the Earth's surface. This distinctive working mechanism enables SAR to penetrate clouds, rain, and fog, providing continuous monitoring regardless of lighting and weather conditions. SAR's ability to function in any environment, day or night, makes it an invaluable tool for uninterrupted maritime surveillance [15]. However, no sensor system is without limitations. Visible light sensors are highly dependent on ambient light, rendering them vulnerable to variations in lighting conditions, such as shadows, occlusions, and low light, which can compromise image quality. Infrared images, while effective in certain conditions, often lack color information and typically have lower resolution compared to visible light images, complicating the tasks of ship identification and scene interpretation. SAR images, due to their complex imaging mechanism, suffer from prolonged imaging times and reduced intuitive appeal. They do not provide direct color information on the Earth's surface and require sophisticated processing and interpretation techniques [16].

By leveraging the combined use of two or more information sources, it is possible to achieve all-weather, round-the-clock ship monitoring, thereby enhancing maritime safety and operational efficiency. In contrast to SAR imagery, the temporal, spatial, feature, and processing consistency of data acquired from visible and infrared sensors significantly improves the training outcomes of deep learning-based ship monitoring systems. Our methodology specifically addresses pre-registered multi-source data scenarios, leveraging spatiotemporally synchronized imaging systems that maintain angular perspective consistency while tolerating minor parallax offsets inherent in practical maritime observation platforms. Visible and infrared images captured simultaneously ensure that the dynamic scene depicted is consistent across both modalities. This temporal consistency facilitates the effective processing of dynamic temporal information during training, allowing for the full utilization of the high-resolution and distinct features of visible imaging during daylight hours. At the same time, it mitigates the limitations of visible imaging in low-light conditions by transitioning to infrared imaging, which excels in night vision environments and in detecting thermal signatures for accurate ship classification. The residual architecture selection directly supports this configuration through its capacity for preserving cross-modal positional relationships—shallow convolutional layers capture alignment-sensitive low-level features while deeper layers maintain geometric correspondence through identity mapping connections. Moreover, when visible light and infrared images are captured from the same angle, the geometric relationships within the scene remain consistent, and the ship's position is identical in both images. This consistency provides a crucial spatial reference for data fusion and model training, with ResNet-34's skip connections proving particularly effective in preserving these geometric priors throughout the network depth. Our architectural analysis reveals that compared to deeper alternatives, ResNet-34 optimally balances positional sensitivity with contextual abstraction—its intermediate feature maps maintain sufficient spatial resolution (112×112 at residual block 3) to resolve typical parallax offsets ≤ 15 pixels observed in our dataset, while avoiding excessive downsampling that could degrade alignment-critical features. This consistency reduces errors in the fusion process and serves as a key factor in the simultaneous training of multiple data sources. It also plays a pivotal role in enhancing the accuracy and reliability of monitoring and identification. The geometric structure of ship images obtained from different sensors at the same moment and angle is consistent, preserving fundamental morphological features such as edges, contours, and basic shapes. This feature consistency is instrumental in enabling the model to better understand and integrate multiple sources of information for effective multi-task learning. Given the similarity in features, consistent or similar preprocessing and data augmentation strategies can be employed across multiple data sources to enhance data diversity while maintaining geometric consistency. This approach

promotes feature sharing across different data sources and fosters cross-source feature learning, thereby optimizing the model's performance.

In summary, the high correlation between ship information obtained from visible and infrared sensors, coupled with the complementary nature of the data they provide, makes it advantageous to combine these two sources. By integrating the information from both sensors, we can leverage their respective strengths to enhance the overall monitoring capabilities.

Deep learning-based multi-task automatic ship recognition technology not only facilitates the learning of common features from multi-source images obtained through different sensors but also transforms the traditional approach into a unified reinforcement framework. This integrated method integrates visible detection and infrared classification tasks, allowing for the fusion of features from individual information sources. As a result, the performance of both visible detection and infrared classification tasks is significantly enhanced when compared to single-source approaches. By combining multiple sources of information, this approach achieves higher monitoring outcomes, achieving state-of-the-art performance compared to the existing ship recognition frameworks.

Inspired by the success of multi-task learning, we aim to improve real-world visible and infrared ship recognition tasks through mutual reinforcement within a unified architecture. Unlike the existing models that treat these tasks independently, this paper introduces a unified multi-task ship recognition model, VIOS-Net, which facilitates interaction between multiple information sources by learning a set of bidirectionally complementary features from ship images relevant to both tasks. This mutual learning process enhances each task's specific features, allowing the high-resolution imaging properties of visible light to be shared and integrated with the penetrating capabilities of infrared imaging. This is achieved using a multi-layer convolutional neural network with shared modules across multiple information sources, where a shared feature extractor is employed at the bottom layer. Two Unique Feature Extractors (UFEs) are positioned at the top to accommodate different task-specific representations and their corresponding parameters. Given that the variation in information content between visible and infrared sources might affect the effectiveness of joint reinforcement learning, we address this by fusing information from multiple sources and designing weighting factors for each source during training. By integrating multiple information sources, the simultaneous learning of visible and infrared ship recognition not only reduces the risk of overfitting—often present when training on a single source—but also addresses the challenge of differing information content between sources, thereby further enhancing recognition accuracy. The learned representations are more compact than those constructed from single-source surface features and demonstrate higher accuracy than the results obtained by merely fusing multiple information sources with linear weighted decisions. The experimental results indicate that the combined visible and infrared ship monitoring system significantly improves the performance of the multi-task ship recognition system VIOS-Net for each information source. This approach yields superior recognition results compared to models that rely solely on single-source information or that simply fuse multiple information sources without integrated decision making.

The experimental results demonstrate that our newly proposed multi-task deep ship recognition system, VIOS-Net, effectively establishes associations between multiple information sources, enabling joint reinforcement recognition. This significantly enhances the model's learning efficacy and strengthens its ability to address the challenges posed by poor sensor imaging due to varying light, weather, and sea state conditions in real-world marine environments. Additionally, this paper offers four main contributions:

- Integrated Multi-Task Fusion Learning Neural Network (VIOS-Net): This study introduces a multi-task fusion learning neural network (VIOS-Net) based on visible

and infrared imaging for ship classification. To the best of our knowledge, this is the first model to integrate visible and infrared multi-information sources into a unified framework for complementary feature fusion and joint reinforcement within a deep ship recognition model. Our model achieves effective ship recognition using only an optical camera and an infrared imager, greatly enhancing recognition accuracy over single-sensor systems. This approach is particularly suited to the complex and dynamic maritime environment, ensuring around-the-clock maritime safety.

- Superior Performance on Real-World Datasets: Using two real-world ship-monitoring datasets (visible and infrared), VIOS-Net achieves the best performance among the existing networks, with an accuracy of 96.199% on both datasets. Compared to the baseline ResNet-34 model, VIOS-Net shows significant improvements—25.688% in visible image recognition accuracy and 28.047% in infrared image recognition accuracy. Furthermore, compared to the simple linear weighted decision fusion of multiple information sources, VIOS-Net demonstrates a 6.00% improvement in accuracy on the visible dataset and a 10.80% improvement on the infrared dataset, highlighting the model's generalization capabilities and its reduced risk of overfitting.
- Impact of Weighting Coefficients on Performance: Through experimentation, we identified the influence of weighting coefficients on model accuracy. The accuracy of both visible and infrared tasks reached 96.199%, and adjustments to the weighting settings during training affected the highest accuracy by 4.150% for visible information sources and 4.650% for infrared information sources. The use of weight coefficients allows the model to more effectively transfer learned features from one task to another, addressing the limitations of neural networks in learning underlying features from a single task.
- Effectiveness of Transfer Learning and Data Augmentation: Our experiments confirmed the effectiveness of transfer learning and data augmentation in improving model performance. Specifically, the use of transfer learning increased accuracy by 5.505% compared to not using it, while data augmentation led to a 2.097% increase in accuracy. These techniques enhance the model's ability to generalize and improve overall recognition performance.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of the related work. Section 3 offers an in-depth description of our fusion monitoring system designed for visible and infrared data sources. In Section 4, we present the experiments conducted and analyze the results obtained. Section 5 evaluates the effectiveness of the various components within our proposed architecture. Finally, Section 6 summarizes the findings and discusses potential directions for future research.

2. Related Works

This section provides a review of related works, which can be classified into two main types: (1) Traditional-Based Vessel Monitoring Method and (2) Image Information Source Vessel Monitoring Method.

2.1. Traditional-Based Vessel Monitoring Methods

Traditionally, radio communication devices aboard ships have been integral to maritime monitoring systems, playing a crucial role in facilitating effective shipping management by port authorities. Early research predominantly centered on enhancing the accuracy of maritime monitoring through the acquisition and optimization of data from radar systems, radio communication, and Automatic Identification System (AIS) equipment. These studies investigated various techniques to improve data collection and processing. As scientific and technological advancements have progressed, deep learning techniques

have been increasingly incorporated into traditional ship monitoring systems. The integration of deep learning has significantly optimized the analysis of ship monitoring data, thereby reducing the need for human intervention and further improving the accuracy of shipping management.

Early maritime monitoring relied on shipboard radio devices (AIS, radar, and GPS) with three evolutionary milestones: the 1940s radar era established basic vessel tracking [17,18], 1990s digital communication enabled real-time reporting [19,20], and recent AI integration enhanced data processing [21,22]. Modern systems combine deep learning with legacy technologies: CNNs improve radar target detection in clutter [23], graph networks enhance spatiotemporal correlation [24], and transformer models detect AIS anomalies [25]. However, these device-dependent approaches face intrinsic limitations—over 25% of maritime accidents involve equipment failures according to EMSA 2024 reports [26].

In summary, while the aforementioned methods have proven effective in utilizing intelligent communication devices aboard ships for monitoring a limited number of vessels, several challenges remain. When these devices malfunction, maritime management agencies face significant obstacles in obtaining timely and accurate information about ship activities at sea. This limitation poses potential risks to maritime safety and regulatory oversight, as the inability to effectively monitor and manage ship behavior may introduce latent hazards that compromise both safety and compliance.

2.2. Image Information Source Vessel Monitoring Method

2.2.1. Single-Source-Based Vessel Monitoring Method

Deep learning techniques have significantly advanced the field of image recognition, and this progress is increasingly being integrated into traditional ship monitoring methodologies. The transition from monitoring conventional radio communication equipment to the analysis of ship images marks a pivotal development in maritime surveillance. This shift is largely driven by the superior capabilities of deep learning models in image recognition, where training on extensive datasets substantially enhances the precision and accuracy of ship identification. Unlike traditional methods, deep learning excels in extracting and discerning representative features from images, thereby improving classification and recognition.

The dominance of single-source approaches in maritime surveillance stems from their conceptual simplicity and direct sensor-specific optimization. The current methodologies can be categorized into three technical streams based on sensor modality and learning strategies: visible spectrum modeling, infrared thermal analysis, and SAR signal processing.

The visible technical stream, pioneered by Liu et al.'s dual-attention network [27], prioritizes RGB feature extraction through background-aware architectures, multi-scale pyramid networks (GLPM's 7.2% accuracy gain [28]), and enhanced attention mechanisms (Xu et al.'s spatial-channel modules [29]). While achieving 89.3% mean accuracy in daylight conditions [30], these methods suffer severe performance degradation (<62% AP) under low-light scenarios due to photometric limitations.

The infrared technical stream, represented by Wu et al.'s SRCANet [31], employs thermal gradient encoding (Li et al.'s low-rank decomposition [32]), salience-guided detection (Chen's morphological reconstruction [33]), and lightweight adaptation (YOLOv5-based nighttime detectors [34]). Though effective in darkness (84.5% recall rate [35]), fog/rain causes > 30% false positives from atmospheric thermal noise.

In the SAR technical stream, advanced through Zeng's dual-polarized CNN [36], key innovations include speckle-robust convolutions (Li's Faster R-CNN variant [37]) and transfer learning for small datasets (2.1× sample efficiency [38]). Despite high resolution (0.5 m precision [39]), SAR struggles with dense ship clusters (<70% detection rate in [40]).

The accumulated evidence underscores a critical insight: the physical limitations of individual sensors create irreversible information bottlenecks that no algorithmic refinement can overcome. This realization motivates our shift toward adaptive multi-source fusion, as detailed in Section 2.2.2.

2.2.2. Multi-Source-Based Vessel Monitoring Method

The evolution of maritime monitoring systems reveals an essential paradox: while single-sensor approaches dominate the current implementations (infrared/visible/SAR), their inherent limitations fundamentally contradict the dynamic complexity of marine environments. Traditional methods relying on single-source data exhibit three critical failures: spectral bias where modality-specific features (e.g., thermal signatures in infrared) overshadow complementary information; environmental fragility causing performance degradation under lighting/weather variations; and scale rigidity limiting adaptation to multi-resolution targets. These deficiencies originate from neural networks' tendency to overfit sensor-specific patterns while ignoring cross-modal relationships.

As the demands on maritime management technology evolve, image classification algorithms grounded in computer vision are playing an increasingly critical role. To overcome challenges such as incomplete information, feature extraction biases, and limited generalization capabilities, researchers have developed methods that train models on data from multiple information sources simultaneously, resulting in more effective monitoring outcomes. Recent advancements address these challenges through three architectural paradigms: parallel modality processing, hybrid hierarchical fusion, and attention-based integration. Exemplified by Li et al. [41]'s dual-stream network for satellite/contour images, this approach achieves 12.6% accuracy gains but incurs 38% computational overhead due to separate processing branches. While transfer learning mitigates domain gaps, their fixed-weight fusion (0.5:0.5 ratio) fails to adapt to environmental changes. Similarly, Aziz et al. [42] introduced a recognition method based on a multi-source convolutional neural network (CNN) for visible and infrared spectra, although the method's ship recognition accuracy remained relatively low. Chang et al. [40,43]'s YOLOv3 variants demonstrate how multi-scale feature extraction (through SPP modules) improves small ship detection (F1-score +9.3%). However, their exhaustive cross-sensor interactions lead to quadratic complexity growth ($O(n^2)$), making real-time processing infeasible with >3 sensors. Additionally, Chen et al. [44]'s parametric edge convolution reduces resource consumption by 22%, yet their static attention mechanism cannot handle heterogeneous sensor sampling rates—a critical limitation in moving vessel scenarios.

The aforementioned studies provide a practical framework for continuous, all-weather ship monitoring to enhance ship monitoring accuracy, marking significant progress in maritime monitoring research. However, many of these approaches predominantly rely on single-source images—such as infrared, visible light, or SAR imagery—for ship identification, often neglecting the potential benefits of integrating information from multiple sources. Such integration could substantially improve the accuracy of ship identification at sea. Moreover, challenges persist in multi-source ship recognition methods, including suboptimal quality in serial feature fusion and high algorithmic complexity. However, four fundamental limitations persist: isomorphic feature treatment disregarding sensor-specific characteristics; fixed fusion weights failing to adapt to environmental dynamics; computation redundancy from exhaustive cross-sensor interaction; and lack of self-supervised alignment for modality gaps.

Additionally, the existing research has largely focused on deep learning tasks without sufficiently addressing the challenges inherent in multi-source ship recognition methods based on deep learning. This oversight significantly hinders the practical applicability

of these models within maritime management systems. Addressing these limitations is crucial for developing more effective and feasible solutions for maritime monitoring and management.

3. VIOS-Net System

The overall workflow of the VIOS-Net network proposed in this paper is divided into three main stages: the preprocessing and augmentation unit (PAU), the shared feature extractor (SFE), the Unique Feature Extractor (UFE), and the Result Output Module, as illustrated in Figure 1 below. The first stage, termed the Data Preprocessing and Augmentation stage, involves the preprocessing and enhancement of infrared and visible light datasets from multiple information sources. In the PAU, input data are processed according to the specific characteristics of each information source. Subsequently, images of the same ship, captured by different sensors from multiple information sources at the same angle, are simultaneously fed into two channels—comprising a visible subnet and an infrared subnet—both of which are based on the ResNet-34 architecture. These two subnets together form the SFE, with shared weights between them during the training process, enabling the learning of effective classification features from the preprocessed and augmented images. The SFE Block performs matrix transformations on the two datasets, producing feature data that integrate shared information from multiple sources. These shared feature data are then input into the UFE, a fully connected layer corresponding to the visible and infrared subnets, where the SoftMax function is applied to extract effective categorical features, yielding predicted ship category labels at the output layer. The error between the actual ship category labels and the predicted labels is calculated, and a weight backpropagation algorithm is employed to iteratively update the weights and biases of the UFE in VIOS-Net, optimizing the implementation of the visible and infrared subnets. During the model performance evaluation phase, after training is completed, multi-source images are fed into the visible and infrared sub-networks separately to extract effective classification features. The optimal model is then used to test visible and infrared images, with the SoftMax function independently classifying each, resulting in final maritime vessel image recognition outcomes for the visible and infrared subnets. The proposed VIOS-Net system demonstrates the ability to perform joint visible and infrared monitoring with high accuracy, providing a practical and reliable solution for maritime monitoring.

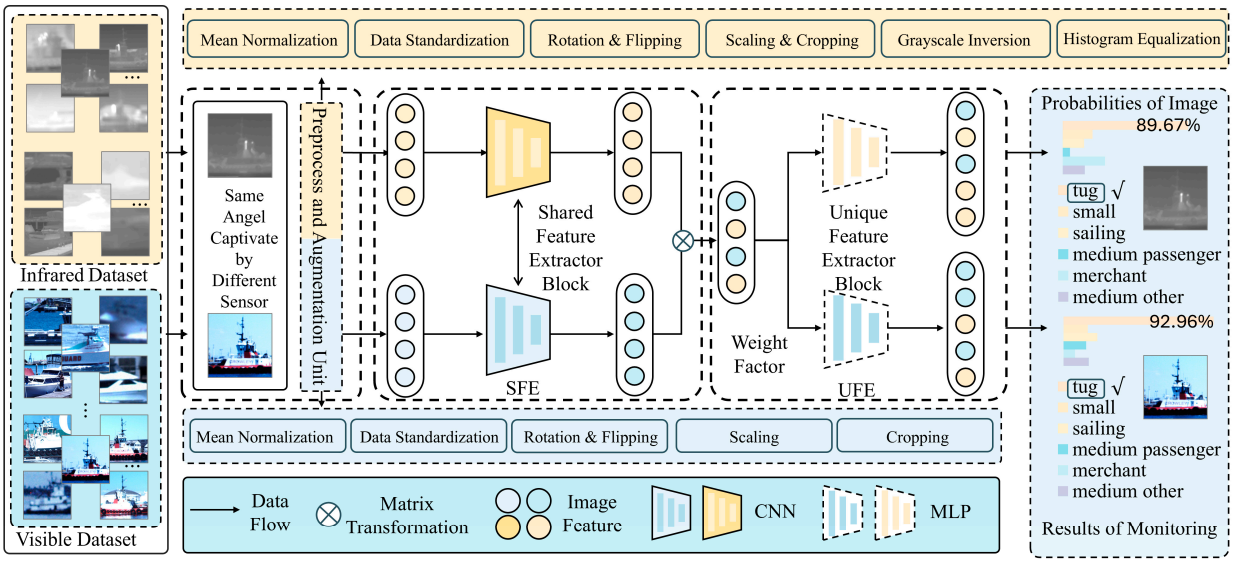


Figure 1. Process architecture diagram of VIOS-Net.

The following sections present a comprehensive description of the VIOS-Net network model, covering the Problem Statement, preprocessing and augmentation unit, shared feature extractor, and Unique Feature Extractor.

3.1. Problem Statement

We defined the ship monitoring task as a hybrid task, which can be divided into a visible light monitoring subtask and an infrared monitoring subtask. The visible light monitoring subtask is defined as follows: given a ship's visible light image, the model determines the type of ship depicted in the image. The visible light dataset is denoted as $I_i^v, \forall i \in \{1, \ldots, \delta\}$, where I_i^v is the i-th image in the dataset, and δ represents the total number of images in the visible light ship image dataset. The output of this subtask is denoted as $f^v : I_i^v \rightarrow \hat{y_i^v}$, where $\hat{y_i^v}$ denotes the predicted labels of the ship's visible images. Similarly, the infrared monitoring subtask uses a dataset denoted as $I_i^r, \forall i \in \{1, \ldots, \lambda\}$, where I_i^r is the i-th image in the dataset, and λ denotes the total number of images in the infrared ship image dataset. The output of this subtask is denoted as $f^r : I_i^r \rightarrow \hat{y_i^r}$, where $\hat{y_i^r}$ represents the predicted labels of the ship's infrared images.

3.2. Preprocessing and Augmentation Unit

The preprocessing and augmentation unit (PAU) aims to enhance the quality of the images, enable the model to learn various features and changes in the data, and ultimately improve model performance. Additionally, the transformations applied during data augmentation generate diverse training samples and reduce the model's dependency on the training data, making the model more robust, less susceptible to noisy data, and better at mitigating overfitting. This, in turn, improves the model's generalization ability. The images generated by preprocessing and data augmentation are shown in Figure 2.

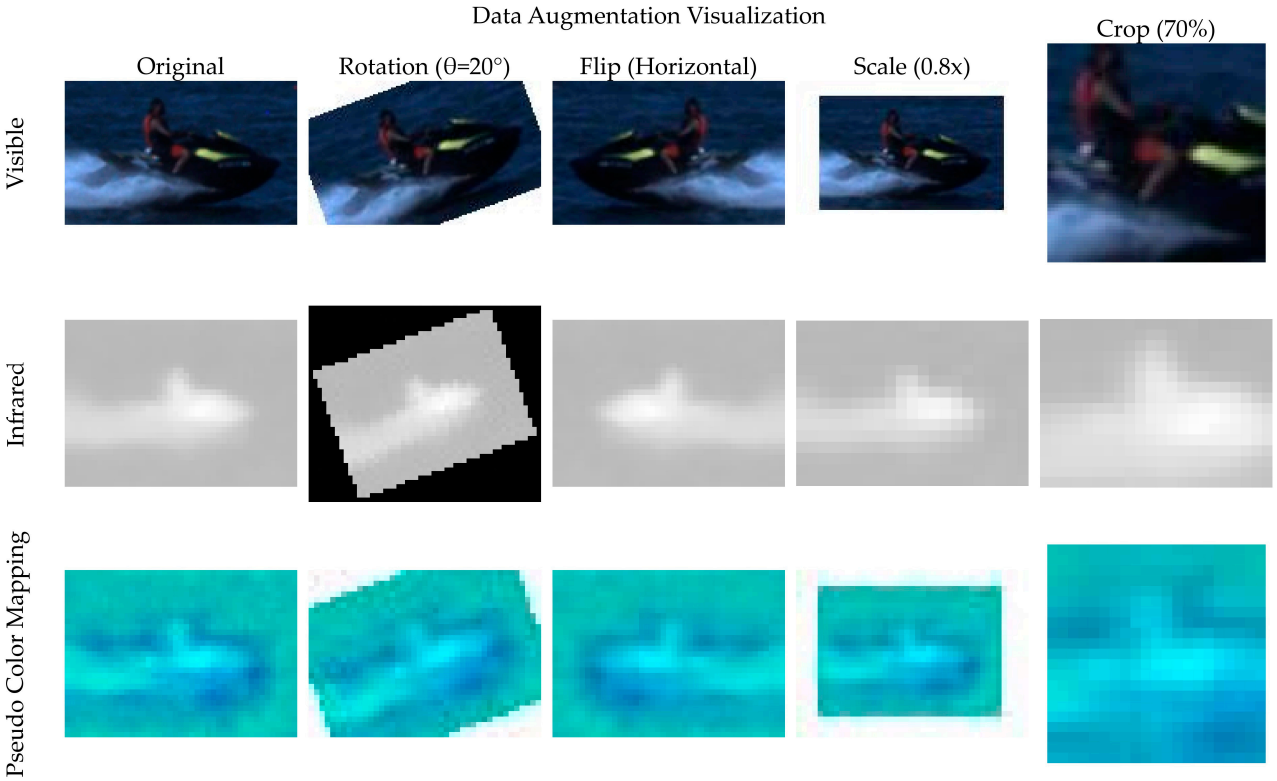


Figure 2. The images are generated based on Algorithms 1 and 2.

Algorithm 1 Preprocess and Augmentation for Visible Light Images Algorithm

Input: Outboard Profile Visible Light Ship Inspection Dataset $I_i^v, \forall \in \{1, \cdots, \delta\}$;
Require: Random rotation function, Random Rotation(); Flipping function, Flipping(); Scaling function, Scaling(); Cropping function, Cropping().

#Mean normalization
$\bar{I}_{\text{mean}}^{\text{v}} = \frac{1}{\delta}\sum_{\text{i=1}}^{\delta} I_i^v$
for i **in** $1 \ldots \delta$ **do**
  $I_i^v = I_i^v - \overline{I}_{mean}^v$
end for
 # Data standardization
$\sigma = \sqrt{\frac{1}{\delta}\sum_{i=1}^{\delta}\left(I_i^v\right)^2}$ # σ is the standard deviation
for i **in** $1 \ldots \delta$ **do**
  $I_i^v = I_i^v / \sigma$
end for
 # Augmentation
SetRandomSeed(seed_value)
for i **in** $1 \ldots \delta$ **do**
  $I_i^{v1} \leftarrow Ramdom\ Rotation\left(I_i^v, \theta \in [-30°, 30°]\right)$ # Simulate vessel orientation variations under wave disturbances
  $I_i^{v2} \leftarrow HorizontalFlipping\left(I_i^v\right)$ # Leverage port-starboard symmetry in ship structures
  $I_i^{v3} \leftarrow Scaling\left(I_i^v, l \in [0.8, 1.2]\right)$ # Emulate multi-scale observation from varying distances
  $I_i^{v4} \leftarrow Cropping\left(I_i^v, W \geq 0.7w, H \geq 0.7h\right)$ # Simulate partial occlusion in camera-captured scenarios
end for
$I_i^v = I_i^{v1} \cup I_i^{v2} \cup I_i^{v3} \cup I_i^{v4}$# Merge the original dataset of four data augmentation methods

Output: preprocess and augmented visible light dataset I_i^v

Algorithm 2 Preprocess and Augmentation for Infrared Images Algorithm

Input: Outboard Profile Infrared Ship Inspection Dataset $I_i^r, \forall \in \{1, \cdots, \lambda\}$;
Require: Random rotation function, RandomRotation(); Flipping function, Flipping(); Scaling function, Scaling(); Cropping function, Cropping();Grayscale inversion function(); Contrast adjusting function();Pseudo-color mapping function, Pseudo ColorMapping().

 # Mean normalization
$\overline{I}_{mean}^r = \frac{1}{\lambda}\sum_{i=1}^{\lambda} I_i^r$
for i **in** $1 \ldots \lambda$ **do**
  $I_i^r = I_i^r - \overline{I}_{mean}^r$
end for
 # Data standardization
$\sigma\prime = \sqrt{\frac{1}{\lambda}\sum_{i=1}^{\lambda}\left(I_i^r\right)^2}$ # σ' *is the standard deviation*
for i **in** $1 \ldots \lambda$ **do**
  $I_i^r = I_i^r / \sigma'$
end for
 # Augmentation
SetRandomSeed(seed_value)
for i **in** $1 \ldots \lambda$ **do**
  $I_i^{r1} \leftarrow RamdomRotation\left(I_i^r, , \theta \in [-30°, 30°]\right)$ # Simulate vessel orientation variations under wave disturbances
  $I_i^{r2} \leftarrow HorizontalFlipping\left(I_i^r\right)$ # Leverage port-starboard symmetry in ship structures
  $I_i^{r3} \leftarrow Scaling\left(I_i^r, , l \in [0.8, 1.2\ \right)$ # Emulate multi-scale observation from varying distances
  $I_i^{r4} \leftarrow Cropping\left(I_i^r, W \geq 0.7w, H \geq 0.7h\right)$ # Simulate partial occlusion in camera-captured scenarios
end for
$I_i^r = I_i^{r1} \cup I_i^{r2} \cup I_i^{r3} \cup I_i^{r4}$ # Merge the original dataset of four data augmentation methods
for i **in** $1 \ldots \lambda$ **do**
  $I_i^{r'} = GrayscaleInversion\left(I_i^r\right)$
  $I_i^{r''} = ContrastAdjusting(\mathrm{I_i^r})$
  $I_i^r = PseudoColorMapping\left(I_i^r, I_i^{r'}, I_i^{r''}\right)$
end for

Output: preprocess and augmented infrared dataset I_i^r

3.2.1. Preprocessing and Augmentation Unit for Visible Light Images

For the visible light image dataset, two preprocessing steps were performed: mean normalization and data standardization. Additionally, we implemented physics-constrained data augmentation through four maritime-compliant transformations: rotation, flipping, scaling, and cropping. The union operation (line 18) aggregates these mutually exclusive augmented samples into the training set, where each image undergoes only one transformation per iteration. This approach expands data diversity while maintaining maritime physical realism. The algorithm of the preprocessing and augmentation unit for the visible light dataset is described in Algorithm 1.

1. Zero-Centering Normalization

Zero-Centering normalization adjusts the pixel value distribution of an image to be centered around zero, reducing data bias and improving model training. This technique helps the model converge more quickly and prevents gradient saturation when activation functions are used in neural network layers, thus enhancing training efficiency and performance.

2. Data Standardization

Data standardization modifies each channel's data so that the mean is zero and the standard deviation is one. This process ensures that the normalized data falls within the same range as the activation function, specifically between 0 and 1. Consequently, fewer non-zero gradients occur during the training of the visible light monitoring module, enabling faster learning by the network's neurons.

3. Rotating

Rotating an image can simulate the appearance of a ship at different angles, enhancing the robustness of the model to changes in orientation while increasing the diversity of the dataset. The rotation angles are constrained to $\pm 30^{\circ}$ to mimic realistic ship pitching motions in maritime environments. The formula for rotation is as follows:

$$\begin{bmatrix} w' \\ h' \end{bmatrix} = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \begin{bmatrix} \frac{w}{2} \\ \frac{h}{2} \end{bmatrix} (\theta \in [-30^{\circ}, 30^{\circ}]) \tag{1}$$

where the given image size h is for height and w for width, the midpoint of the image is $\left(\frac{w}{2}, \frac{h}{2}\right)$, the angle of rotation is θ, and the new position of the midpoint is (w', h').

4. Horizontal flipping

Horizontal flipping an image increases the amount and diversity of training data, addressing the lack of samples in certain directions within the dataset. Specifically, while vertical flipping is deliberately excluded due to its contradiction with fundamental buoyancy principles in vascular dynamics, horizontal flipping maintains physical plausibility. The formula for horizontal flipping is as follows:

$$x' = w - x \tag{2}$$

$$y' = y \tag{3}$$

where (x, y) are arbitrary coordinate points in the given image, and (x', y') are the flipped coordinates.

5. Scaling

Random scaling of the image can simulate the change in the size of the ship at different distances and help the model to identify the ship effectively at different scales, the scaling formula for the image is as follows:

$$x' = l \cdot x \ (l \in [0.8,\ 1.2]) \tag{4}$$

$$y' = l \cdot y (l \in [0.8, 1.2]) \tag{5}$$

where (x, y) are the coordinates of a point in the image, (x', y') is the scaled coordinate, and l is the scaling factor.

6. Cropping

Cropping involves selecting a random sub-image from an image, allowing the model to extract image data from different regions and enhancing its ability to learn diverse image information. The cropped image includes the following pixels from the original image:

$$x \in [x_0, x_0 + W),\ W \in [0.7w, w] \tag{6}$$

$$y \in [y_0, y_0 + H),\ H \in [0.7h, h] \tag{7}$$

where $x_0 \in [0, w - W)$, $y_0 \in [0, h - H)$, (w, h) is the size of the original image, (W, H) is the size of the cropped image, and (x_0, y_0) are the coordinates of the upper left corner.

3.2.2. Preprocessing and Augmentation Unit for Infrared Images

For the infrared (IR) image set, we applied two preprocessing steps: Zero-Centering normalization and data standardization. Additionally, six maritime-constrained data augmentation techniques were implemented: rotation, flipping, scaling, cropping, grayscale inversion, and contrast adjustment. We also achieved pseudo-color mapping by channel splicing the original IR image with the grayscale-inverted and contrast-adjusted images. The algorithm for the preprocessing and augmentation unit for the Infrared Dataset is detailed in Algorithm 2.

1. Basic Preprocessing and Data Augmentation for Infrared Images

The preprocessing and data augmentation techniques applied to infrared images are the same as those used for visible light images, including rotation, horizontal flipping, scaling, and cropping.

2. Pseudo-color Mapping

Pseudo-color mapping significantly enhances the informativeness and visual interpretability of infrared images, facilitating improved analysis and processing by both human observers and computer algorithms. This technique involves enhancing contrast and inverting the grayscale within the infrared data processing unit to enrich the visual content of the image. As a result, the features within the image become more prominent, simplifying observation, analysis, and interpretation, while also making the image more accessible for computer processing and deep learning applications. The pseudo-color mapping formula is expressed as follows:

$$I_i^r \leftarrow concat\left[I_i^r,\quad I_i^{r'},\quad I_i^{r''}\right] \tag{8}$$

In this formula, I_i^r on the left side of the concat function denotes the original infrared image. $I_i^{r'}$ represents the infrared image after grayscale inversion, while $I_i^{r''}$ refers to the infrared image after histogram equalization. The result of the concat function is a

new infrared image that has undergone the stitching process, effectively integrating the information from all three versions into a single, enhanced dataset.

- Grayscale Inversion

Grayscale inversion facilitates the exchange of light and dark regions within the original image, thereby enhancing the level of detail available for pseudo-color mapping. This transformation also modifies the visual presentation of the image, making distinct features more easily identifiable. The formula for grayscale inversion is given by the following:

$$I_i^{r'}(x,y) = 255 - I_i^r(x,y) \tag{9}$$

where (x,y) represents any coordinate point within the image, while $I_i^r(x,y)$ denotes the gray value at this coordinate in the original image. The resulting $I_i^{r'}(x,y)$ is the gray value at the corresponding point in the grayscale-inverted image $I_i^{r'}$.

- Histogram Equalization

Histogram equalization is employed to redistribute the gray levels within an image, effectively enhancing contrast and addressing issues such as the blurring of boundaries between the monitored ship and its background. This adjustment makes the ship more prominent within the image. The contrast transformation process is outlined as follows, beginning with the computation of the gray-level histogram:

$$n_g = \sum I_g \tag{10}$$

In this equation, n_g represents the frequency of occurrence for a given gray level g within the image. Here,g is the gray value at pixel $I(x,y)$ in the ith image I_i^r, where $g \in (0,255)$. I_g is a matrix that shares the same dimensions as the ith image I_i^r, and $I_g(x,y)$ is defined as follows:

$$I_g(x,y) = \begin{cases} 1 & \text{if} \quad I(x,y) = i \\ 0 & \text{else} \end{cases} \tag{11}$$

Next, the probability density function (PDF) is calculated using the following formula:

$$p(g) = \frac{n_g}{w \times h} \tag{12}$$

where $w \times h$ represents the total number of pixels in the image, with w and h corresponding to the image's width and height, respectively. Following this, the cumulative distribution function (CDF) is determined as follows:

$$c(g) = \sum_{j=0}^{g} p(j) \tag{13}$$

Finally, gray level mapping is performed, wherein the original image is mapped from its gray level $I_i^r(x,y)$ to the new gray level $I_i^{r''}(x,y)$ using the CDF as follows:

$$I_i^{r''}(x,y) = \lfloor 255 \cdot c(g) \rfloor I_i^r(x,y) \tag{14}$$

where $\lfloor \cdot \rfloor$ represents the floor function, which rounds down to the nearest integer, yielding the histogram-equalized image $I_i^{r''}$.

3.3. Shared Feature Extractor

The infrared (IR) and visible light images I_i^r and I_i^v, obtained from the preprocessing and augmentation units, are subsequently fed simultaneously into the shared feature

extractor (SFE) for feature extraction. To accurately recognize a set of ship images relevant to both tasks, we employed the ResNet-34 model as the backbone, utilizing its residual network for robust feature extraction. The data are fed into the SFE through dual channels, as represented by the following formula:

$$s = SFE(I_i^r, I_i^v) \tag{15}$$

where s denotes the extracted feature, SFE stands for the shared feature extractor, and I_i^r and I_i^v represent the infrared and visible images, respectively.

3.3.1. Residual Networks Block Unit

These image representations are processed through multiple shared residual building blocks to achieve deeper feature extraction. The structure of these residual blocks is depicted in Figure 3, which includes a convolutional layer, a batch normalization layer, and an activation layer. Each residual building block follows two potential paths: the direct path Residual(x) and the shortcut path Residual$(x) + x$. The latter is implemented through feedforward neural networks with "shortcuts," allowing the bypass of one or more layers, thereby facilitating efficient information transfer and reducing the risk of model overfitting.

After processing through the residual building blocks, the information is passed to the subsequent layer, the Average Pooling layer, for dimensionality reduction. This processed information is then forwarded to the Unique Feature Extractor (UFE) for specialized feature extraction. The detailed algorithm for the residual networks block unit in the shared feature extractor is provided in Algorithm 3.

Algorithm 3 Residual Networks for Shared Feature Extractor

Input: Preprocessed visible light image set
$I_i^v, \forall \in \{1, \cdots, \delta\}$; *Infrared image set* $I_i^r, \forall \in \{1, \cdots, \lambda\}$;
Require: basic block of residual networks, Residual(); convolution layer, Conv(); max pooling function, Maxpool(); avgpooling function, Avgpool(), Fully Connected Layer, FC()

 # Visible Image Feature Extracting
$s_i^v \leftarrow Conv(I_i^v)$
$s_i^v \leftarrow Maxpool(s_i^v)$
for i **in** [3,4,6,3] **do** #[3,4,6,3] is the number of the residual basic block
  for j **in** $1 \ldots i$ **do**
    $s_i^v \leftarrow Residule(s_i^v)$
  endfor
endfor
$s_i^v \leftarrow Avgpool(s_i^v)$
 # Infrared Images Feature Extracting
$s_i^r \leftarrow Conv(I_i^r)$
$s_i^r \leftarrow Maxpool(s_i^r)$
for i in [3,4,6,3] do #[3,4,6,3] is the number of the residual basic block
  for j **in** $1 \ldots i$ **do**
    $s_i^r \leftarrow Residule(s_i^r)$
  end for
end for
$s_i^r \leftarrow Avgpool(s_i^r)$

Output: Infrared images feature s_i^r; visible light images feature s_i^v

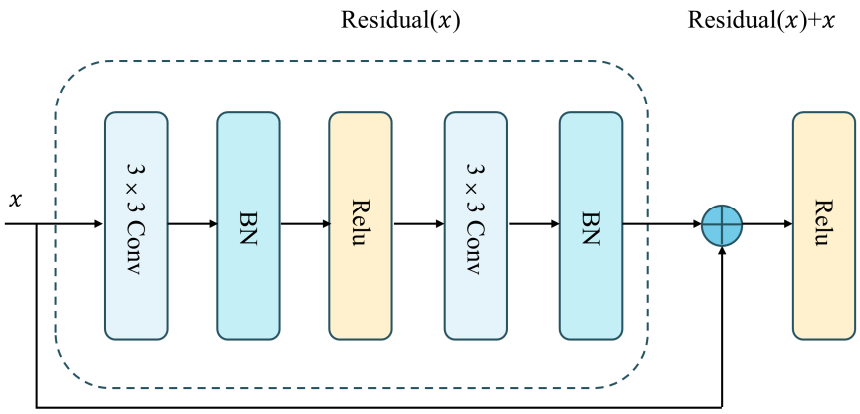


Figure 3. The diagram of the residual block architecture.

3.3.2. Shared Feature Extracting Block Unit

Within the Shared Feature Extracting Block Unit, a set of related visible and infrared image information is concurrently fed into the SFE from both channels to obtain a more sophisticated representation. The core layer of this unit, the convolutional layer, which is integral to the convolutional neural network (CNN), scans the input data using convolutional kernels to extract high-level features. These extracted features are then processed through a max-pooling layer to identify the most representative features. The formula governing the convolutional layer operation is as follows:

$$s_{ij}^{r} = \sum_{m}\sum_{n}(I_i^r)_{i+m} \times w_{m,n} \tag{16}$$

$$s_{ij}^{v} = \sum_{m}\sum_{n}(I_i^v)_{i+m} \times w_{m,n} \tag{17}$$

where s denotes the extracted feature, I is the input image to the convolutional layer, w represents the weights of the convolutional kernel, i and j are the dimensions of the extracted information, and m and n are the dimensions of the convolutional kernel. After processing the IR and visible images through the SFE, their features are extracted and subsequently fused. The fusion process is represented by the following formula:

$$s_{shared} = s_{ij}^{r} \cdot s_{ij}^{v} \tag{18}$$

In this equation, s_{shared} represents the shared feature derived from the fusion of visible and infrared image features, while s_{ij}^{r} and s_{ij}^{v} denote the features obtained from the SFE for the infrared and visible images, respectively.

3.4. Unique Feature Extractor

To systematically address platform-induced spatiotemporal misalignments between modalities, the Unique Feature Extractor (UFE) operates in tandem with the shared feature extractor (SFE) through a dual-branch architecture. While SFE focuses on multi-source data invariant patterns, UFE explicitly handles modality-specific distortion correction and feature recalibration through adaptive attention mechanisms—crucial for maintaining representation integrity under practical parallax conditions.

3.4.1. Unique Feature Extracting Block Unit

To systematically address platform-induced spatiotemporal misalignments between modalities, two sets of fully connected layers are employed to perform unique feature extraction for both visible and infrared images, tailored to the respective tasks. The UFE

block processes multi-modal features through parallel transformation paths: employing FC_r with temperature-aware kernel initialization to enhance thermal signature resolution and utilizing FC_v with spectral sensitivity normalization for RGB feature adaptation. The extraction process is governed by the following formulas:

$$s_{ij}^{r} \leftarrow FC_r(s_{shared}) \tag{19}$$

$$s_{ij}^{v} \leftarrow FC_v(s_{shared}) \tag{20}$$

$$\hat{y_i^r} = Softmax\left(W_r \cdot s_{ij}^{r} + b_r\right) \tag{21}$$

$$\hat{y_i^v} = Softmax\left(W_v \cdot s_{ij}^{v} + b_v\right) \tag{22}$$

where s_{shared} represents the data output from the shared feature extractor, while W_r and b_r denote the weight and bias for the infrared image extractor. Similarly, W_v and b_v represent the weight and bias of the visible light image extractor within the Unique Feature Extractor (UFE). Though sharing the input s_{shared}, these fully connected layers contain disjoint parameters optimized for distinct sensor physics.

3.4.2. Unique Feature Extracting Weight Factor

In multi-task training involving multiple information sources, the visible light and infrared images originate from different sensors, which inevitably results in varying image information being captured by these sensors. As the deep learning model is trained on images from these diverse sources, it can lead to discrepancies in how the model attends to the information from each source, as well as differences in the training progress for the two tasks.

To address this, we introduced different weight coefficients during backpropagation for the tasks under varying information sources. This approach optimizes the model's accuracy across both tasks by adjusting the focus of the learning process. In traditional multi-task training, each task is typically assigned the same weight during training. However, by configuring appropriate weighting coefficients and learning rates for each task, the model can be directed to focus more or less on certain tasks, thereby enhancing or diminishing its ability to capture task-specific features. This method helps the model develop a deeper and more precise task-specific representation. Moreover, due to the model's shared representation learning nature, these adjustments not only impact the specific task for which the learning rate is configured but also positively influence other tasks, allowing the model to balance the learning across multiple tasks. The loss function for predicting results and calculating label loss is given by the following:

$$\mathcal{L}_r = \beta_r \cdot \left[-\sum_{i=1}^{N} y_i \log(\hat{y}_i) \right] \tag{23}$$

$$\mathcal{L}_v = (1 - \beta_r) \cdot \left[-\sum_{i=1}^{N} y_i \log(\hat{y}_i) \right] \tag{24}$$

$$\mathcal{L} = \text{mean}(\mathcal{L}_r + \mathcal{L}_v) \tag{25}$$

where $\mathcal{L}_v$ and $\mathcal{L}_r$ represents the losses between the predicted result and the true label for visible and infrared images, respectively. $\mathcal{L}$ represents the average loss of $\mathcal{L}_v$ and $\mathcal{L}_r$, where β_r serves as the weight factor for the infrared data source. The formula for updating the weights for unique feature extraction is as follows:

$$w \leftarrow w - \alpha \frac{\partial \mathcal{L}}{\partial W} \tag{26}$$

$$b \leftarrow b - \alpha \frac{\partial \mathcal{L}}{\partial b} \tag{27}$$

In these equations, w denotes the weight, b is the bias, and α is the learning rate. This approach enables the model to extract specific features from images derived from different information sources effectively.

4. Experiments and Results

4.1. Dataset

Due to the limited availability of infrared-visible related datasets in the maritime domain, we utilized two publicly available datasets that were balanced and augmented for both visible and infrared images. These datasets were used to train the VIOS-Net model, enabling the evaluation of its performance. The dataset categorizes ship targets into six types: medium-sized "other" ships, medium-sized passenger ships, merchant ships, sailboats, small boats, and tugboats. Given the imbalanced data distribution in both the infrared and visible light datasets, as shown in Table 1, with sailboats and small boats accounting for 24.6% and 39.1%, respectively, and other categories representing less than 20%, we addressed the issue of data imbalance to mitigate model training bias. Initially, we employed data augmentation techniques to achieve a more balanced dataset distribution across categories, as shown in Table 2. Following augmentation, the data were relatively evenly distributed and subsequently split into training and testing datasets at an 8:2 ratio, as depicted in Table 3. This resulted in 3052 training images and 763 testing images. An example of the dataset is illustrated in Figure 4.

Table 1. Distribution of datasets prior to data balancing.

Dataset	Category	Train	Test	All
Visible Dataset	medium 'other'	148	37	185
	medium passenger	112	28	140
	merchant	144	36	180
	sailing	330	82	412
	small	524	131	655
	tug	80	20	100
Infrared Dataset	medium 'other'	148	37	185
	medium passenger	112	28	140
	merchant	144	36	180
	sailing	330	82	412
	small	524	131	655
	tug	80	20	100

Table 2. Comparation of dataset distribution pre- and post-data balancing.

Category	Original Data	Balanced Data	Category	Original Data	Balanced Data
medium 'other'	185	650	medium 'other'	185	650
medium passenger	140	560	medium passenger	140	560
merchant	180	650	merchant	180	650
sailing	412	650	sailing	412	650
small	655	655	small	655	655
tug	100	650	tug	100	650

Table 3. Distribution of datasets after data balancing.

Dataset	Visible		Infrared	
Category	**Train Set1**	**Test Set1**	**Train Set2**	**Test Set2**
medium 'other'	520	130	520	130
medium passenger	448	112	448	112
merchant	520	130	520	130
sailing	520	130	520	130
small	524	131	524	131
tug	520	130	520	130

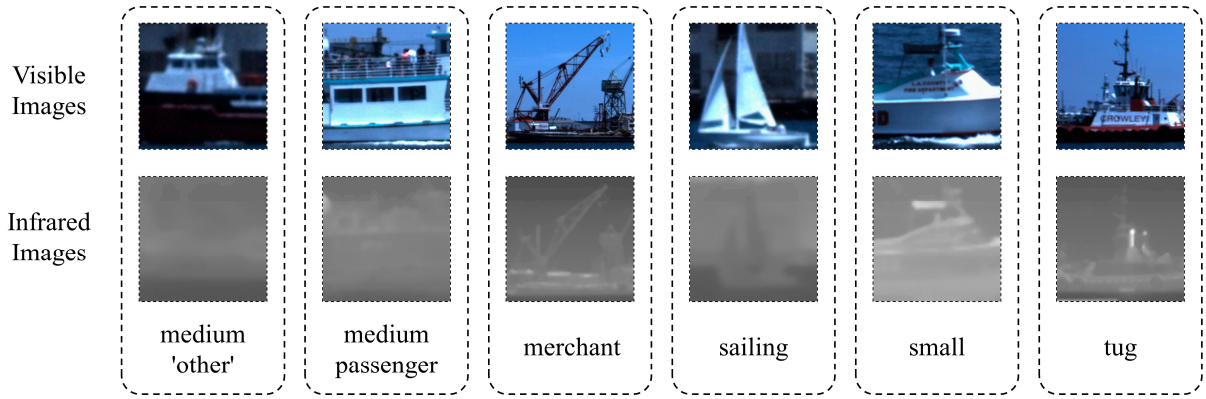


Figure 4. Example images from the VAIS dataset.

The performance of the VIOS-Net system was quantitatively evaluated using several key metrics, including accuracy, precision, recall, and F1-score. These metrics are essential for assessing the effectiveness of the model in classification tasks. The evaluation metrics are defined as follows:

As a classification problem, six evaluation metrics are used in this paper to evaluate the performance of VIOS-Net, Accuracy, Recall, Precision, and F1-score (F1), these formulas are shown below:

$$\text{Accuracy} = (TP + TN)/(TP + FP + FP + FN) \tag{28}$$

$$\text{Recall} = TP/(TP + FN) \tag{29}$$

$$\text{Precision} = TP/(TP + FP) \tag{30}$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \tag{31}$$

In these formulas, TP (True Positive) refers to the number of samples correctly predicted as positive by VIOS-Net, FP (False Positive) is the number of samples incorrectly predicted as positive, TN (True Negative) represents the number of samples correctly predicted as negative, and FN (False Negative) denotes the number of samples incorrectly predicted as negative.

4.2. Experiment Setup

The VIOS-Net system was constructed using the ResNet-34 model, which was pre-trained on the ImageNet dataset through transfer learning and served as the backbone network for the monitoring module. The entire model was optimized using a proposed multivariate cross-entropy loss function, incorporating weighting coefficients with a ratio of 0.4:0.6 for infrared and visible tasks to enhance task-specific optimization. The optimizer chosen was Averaged Stochastic Gradient Descent (ASGD). Empirically, we set the batch size to 32, the learning rate to 0.0001, and the number of training iterations to 60.

The experiments were conducted on a system running a 64-bit Windows 11 operating system, powered by a 12th generation Intel (R) Core (TM) i7-12700 processor at 2.10 GHz, equipped with 32 GB of RAM and an NVIDIA GeForce RTX 3060 graphics card. The deep learning framework employed was PyTorch 1.13, with PyCharm 2024.1.1 (Community Edition) as the primary software tool, and Python 3.9 as the programming language.

4.3. Performance Comparison

This section presents a comparative analysis of our model against the existing baseline models, including widely used deep learning models for classification tasks such as AlexNet, VGG-13, VGG-16, GoogLeNet, ResNet-18, and ResNet-34, utilizing the VAIS dataset. All the comparative models were initialized with ImageNet pre-trained weights. The VIOS-Net framework not only outperformed traditional baselines but also surpassed the improved CNN and Dual CNN models across all the evaluated metrics. The comparison metrics—accuracy, precision, recall, and F1-score—are summarized in Table 4. The VIOS-Net framework demonstrated state-of-the-art performance on the VAIS dataset, underscoring its capability for high-accuracy monitoring and strong generalization across different datasets. The experimental outcome is presented in the last row of Table 4.

Table 4. Comparison of quantitative evaluation across different methods.

Method	Visible Images (%)				Infrared Images (%)			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Alex Net [45]	88.976	89.500	88.650	88.820	86.239	86.380	86.210	85.950
VGG-11 [46]	83.202	83.790	83.300	83.040	84.142	84.600	84.220	83.780
VGG-13 [46]	84.514	89.980	84.340	84.820	84.535	84.460	84.430	84.400
GoogLeNet [47]	88.451	89.630	88.510	88.550	85.714	79.990	78.790	78.200
ResNet-18 [48]	88.189	88.500	88.050	88.140	88.467	88.520	88.450	88.320
ResNet-34 [48]	91.339	91.620	91.410	91.280	87.156	86.920	87.220	86.960
MobileNetV2 [49]	86.089	86.790	85.920	85.970	84.928	84.700	85.020	84.700
MobileNetV3 Small [49]	90.814	91.190	90.780	90.820	85.452	85.380	85.390	85.100
ShuffleNetV2_x0_5 [50]	89.238	89.780	89.280	89.250	87.139	88.230	87.190	87.210
SqueezeNet [51]	86.352	86.800	86.380	85.910	76.016	75.200	75.840	74.860
ConvNeXt [52]	89.501	91.050	89.210	89.470	84.928	85.630	84.980	84.880
Improved CNN [39]	90.200	91.600	89.100	90.000	85.400	82.900	85.800	83.500
Dual CNN [39]	93.600	95.500	92.000	93.500	93.600	95.500	92.000	93.500
VIOS-Net	**96.199**	**96.214**	**96.241**	**96.227**	**96.199**	**96.238**	**96.263**	**96.250**

Among the evaluated models, ResNet-34 emerged as the best-performing backbone network for image recognition, which may be attributed to ResNet's architecture that effectively mitigates the issues of gradient vanishing and exploding in deep neural network training. This allows for the successful training of very deep networks without performance degradation. As evidenced in Table 4, VIOS-Net achieved a substantial improvement in accuracy metrics, with an 8.010% and 7.732% increase for visible and infrared imaging, respectively, and an 8.059% and 7.879% improvement in F1-scores compared to ResNet-18. The multi-task learning approach employed by VIOS-Net facilitated complementary information sharing and feature extraction through the shared feature extractor (SFE), significantly enhancing recognition performance. Additionally, multi-task learning improved the model's generalization ability and reduced the risk of overfitting by leveraging shared features across multiple tasks.

As depicted in Figures 5 and 6, the ROC curves for the visible light and infrared monitoring task indicate that VIOS-Net achieved 96.20% accuracy in ship classification tasks. The macro average and micro average ROC curve values surpassed 0.99 for visible

and infrared image recognition. Compared to the other networks, the superior monitoring performance of VIOS-Net in both the infrared and visible light detection tasks highlights its robust capability to provide stable performance across diverse ship recognition scenarios.

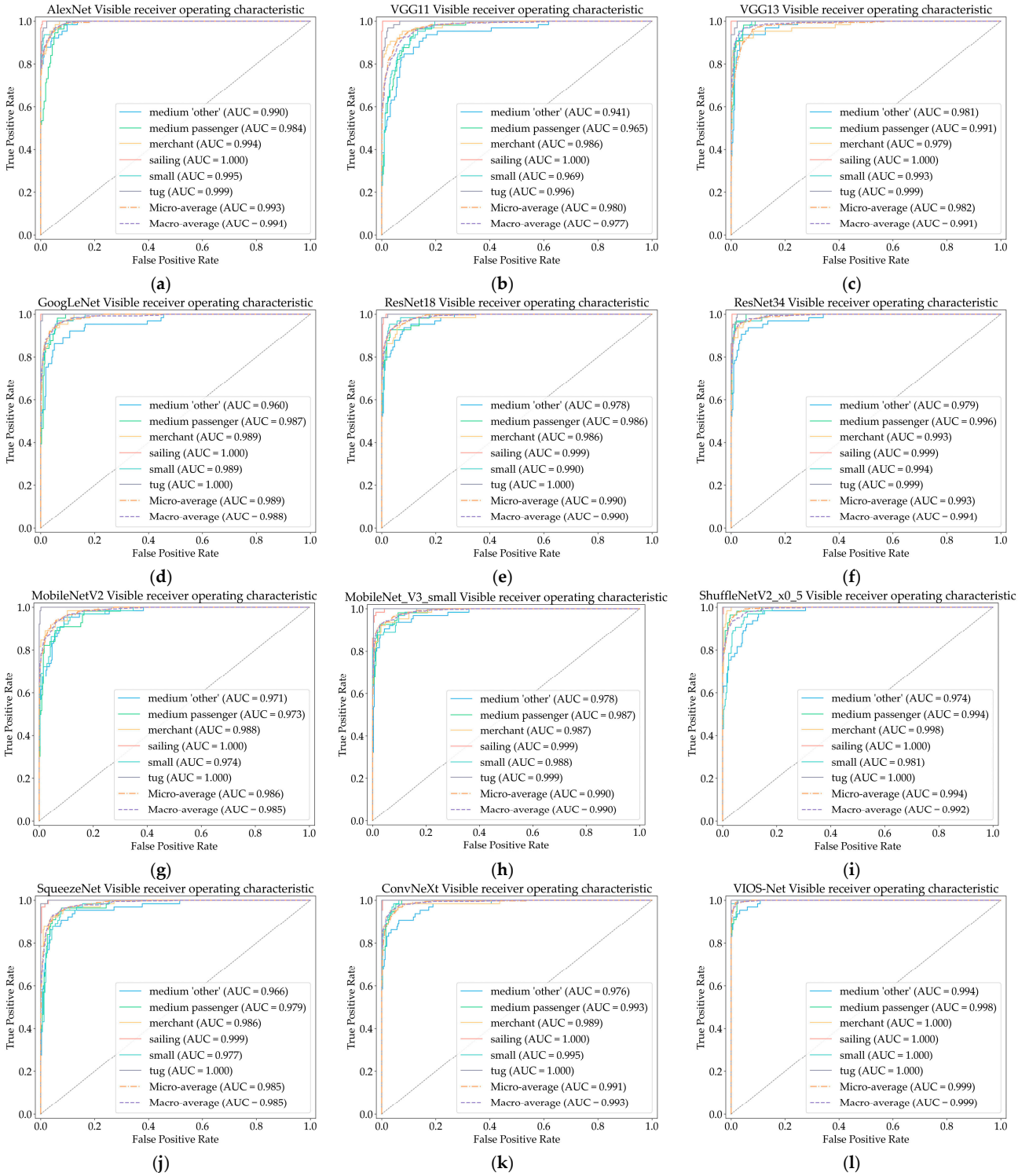


Figure 5. ROC curves across different methods on the visible light dataset: (**a**) AlexNet; (**b**) VGG-11; (**c**) VGG-13; (**d**) GoogLeNet; (**e**) ResNet-18; (**f**) ResNet-34; (**g**) MobileNet-V2; (**h**) MobileNet-V3Small; (**i**) ShuffleNet-V2_x0_5; (**j**) SqueezeNet; (**k**) ConvNeXt; (**l**) VIOS-Net(ours).

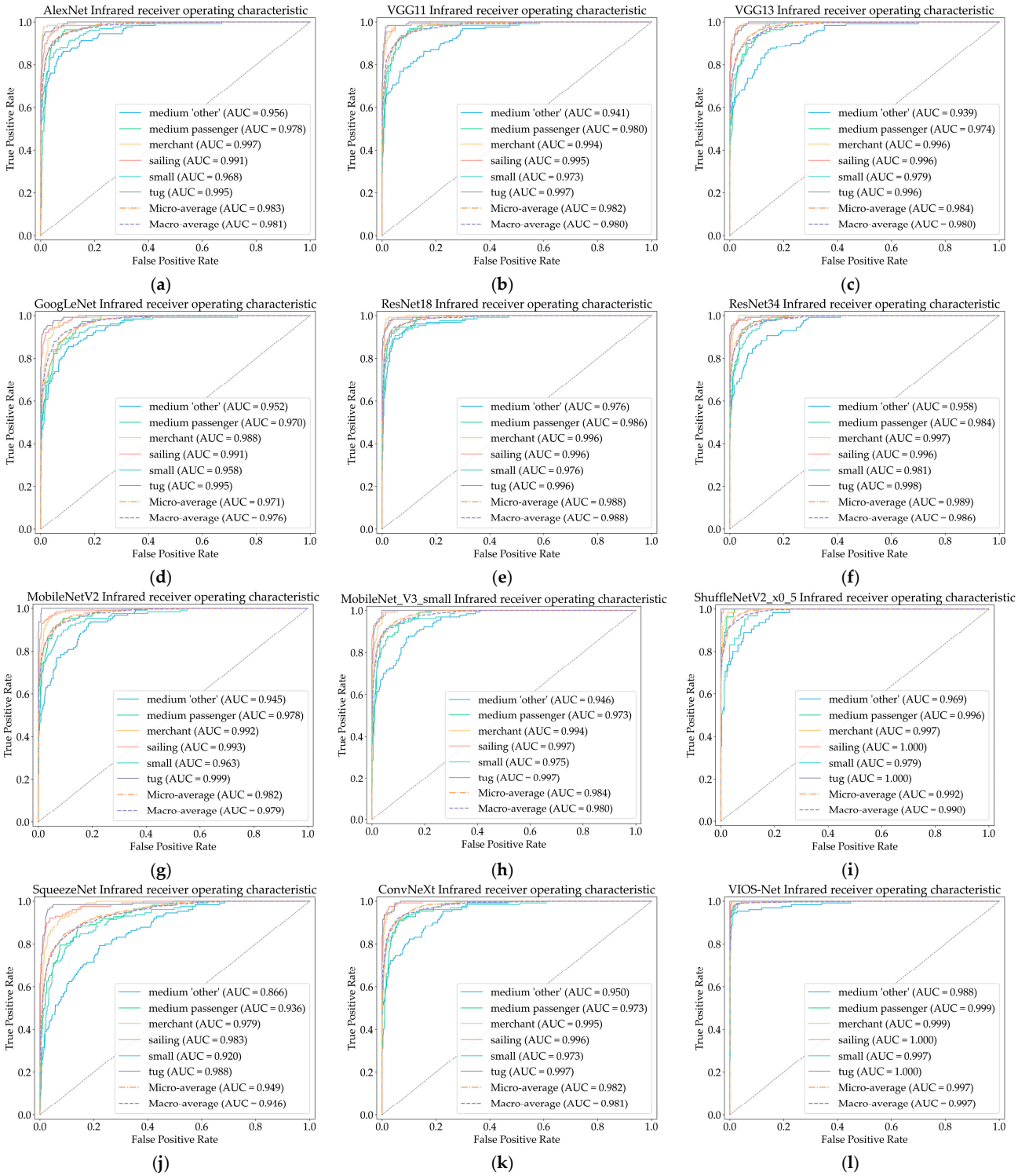


Figure 6. ROC curves across different methods on the infrared dataset: (**a**) AlexNet; (**b**) VGG-11; (**c**) VGG-13; (**d**) GoogLeNet; (**e**) ResNet-18; (**f**) ResNet-34; (**g**) MobileNet-V2; (**h**) MobileNet-V3Small; (**i**) ShuffleNet-V2_x0_5; (**j**) SqueezeNet; (**k**) ConvNeXt; (**l**) VIOS-Net(ours).

5. Quantitative Analysis

To validate the effectiveness of the various components in the VIOS-Net architecture, several ablation experiments were conducted using different variants of the model. These variants include VIOS-Augmentation—a variant of VIOS-Net that excludes data augmentation; VIOS-Transfer—a variant of VIOS-Net that excludes the transfer learning module;

VIOS-Infrared—the infrared image single-task network within VIOS-Net; VIOS-Visible—the visible light single-task network within VIOS-Net; VIOS-SFE—a variant of VIOS-Net that excludes the Shared Feature Extracting module; VIOS-Weight—a variant of VIOS-Net that removes the effect of weight coefficients.

5.1. Preprocessing and Augmentation Unit Impact Study

To demonstrate the improvement brought about by the transfer learning and data augmentation strategies, an experiment was conducted comparing the performance of VIOS-Net with and without data augmentation. The results are presented in Table 5 and Figure 7. In the VIOS-Augmentation model, which lacks preprocessing and data augmentation, the original dataset I_i^r was directly input into VIOS-Net without any pseudo-color mapping or data augmentation. The results indicate that the absence of preprocessing and data augmentation led to a 2.097% decrease in accuracy on the visible light dataset and a 1.966% decrease in the infrared dataset compared to VIOS-Net. These findings underscore the effectiveness of the preprocessing and data augmentation strategies employed in VIOS-Net.

Table 5. Comparison of quantitative evaluation on data augmentation.

	Visible Images (%)				Infrared Images (%)			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
With Data Augmentation Technique	96.199	96.214	96.241	96.227	96.199	96.238	96.263	96.250
Without Data Augmentation Technique	94.102	94.484	94.132	94.308	94.233	94.498	94.281	94.389

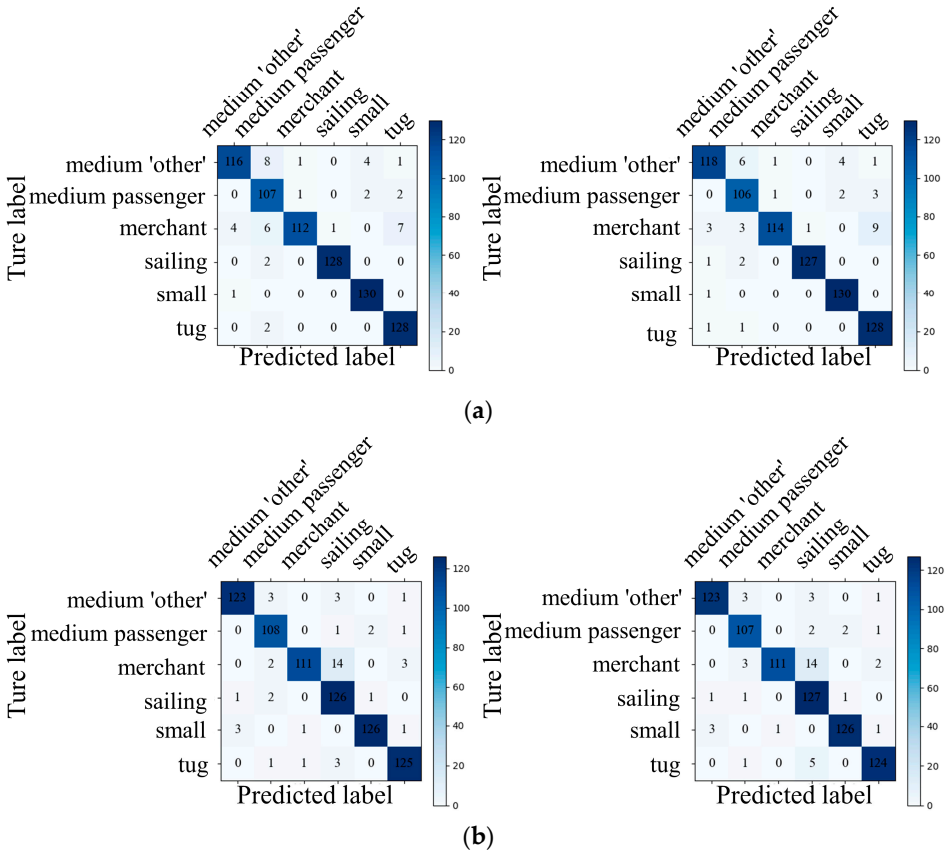


Figure 7. Data augmentation impact result: (**a**) training result with data augmentation technique (left: visible light dataset; right: infrared dataset); (**b**) training result without data augmentation technique (left: visible light dataset; right: infrared dataset).

5.2. Transfer Learning Impact Study

To assess the impact of transfer learning on VIOS-Net, the transfer learning module was excluded, resulting in the VIOS-Transfer model. As demonstrated in Table 6 and Figure 8, the absence of transfer learning led to a decrease in prediction accuracy by 5.505% for visible light data and 5.374% for infrared data. Additionally, the AP values of the PR curves declined by 0.3 and 0.2, respectively. These results underscore the crucial role of transfer learning in enhancing information flow within the fusion network, thereby emphasizing its significance in improving the overall performance of VIOS-Net. As depicted in Figure 9, VIOS-Net exhibits distinct learning behaviors under three configurations: (1) baseline (without augmentation/transfer learning), (2) transfer learning only, and (3) combined augmentation-transfer learning.

Table 6. Comparison of quantitative evaluation on transfer learning.

	Visible Images (%)				Infrared Images (%)			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
With Transfer Learning Technique	96.199	96.214	96.241	96.227	96.199	96.238	96.263	96.250
Without Transfer Learning Technique	90.694	91.042	90.324	90.682	90.825	91.150	90.475	90.811

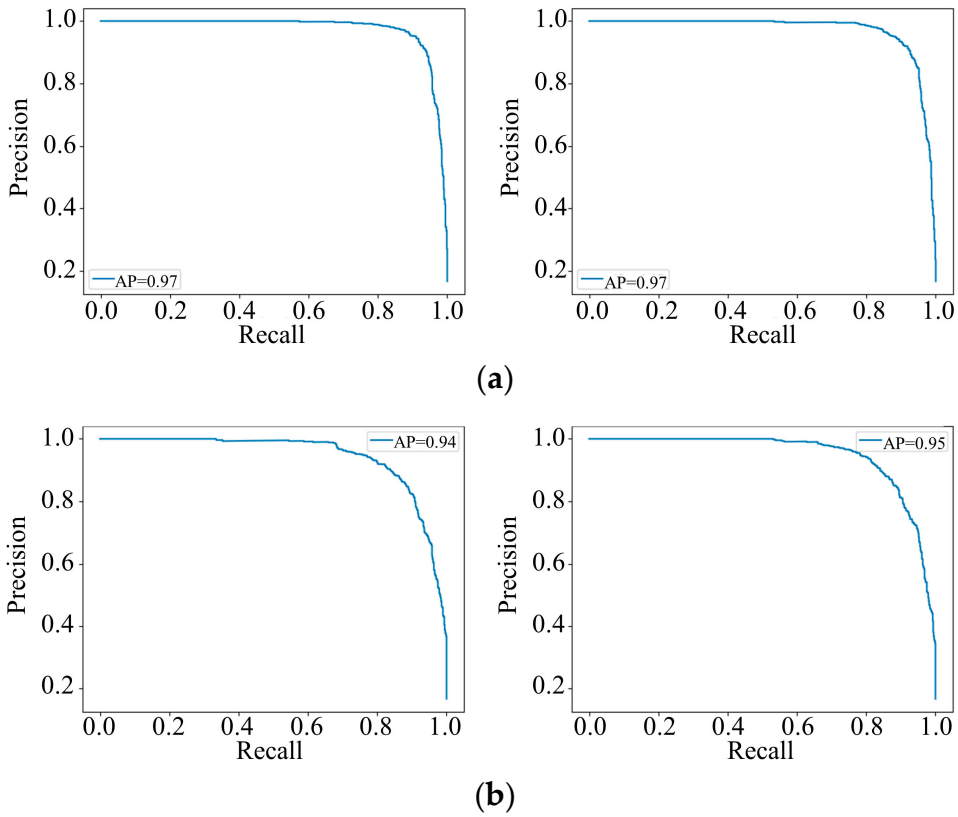


Figure 8. Impact of transfer learning: (**a**) training results with transfer learning technique (left: visible light dataset; right: infrared dataset); (**b**) training results without transfer learning technique (left: on the infrared dataset; right: on the visible light dataset).

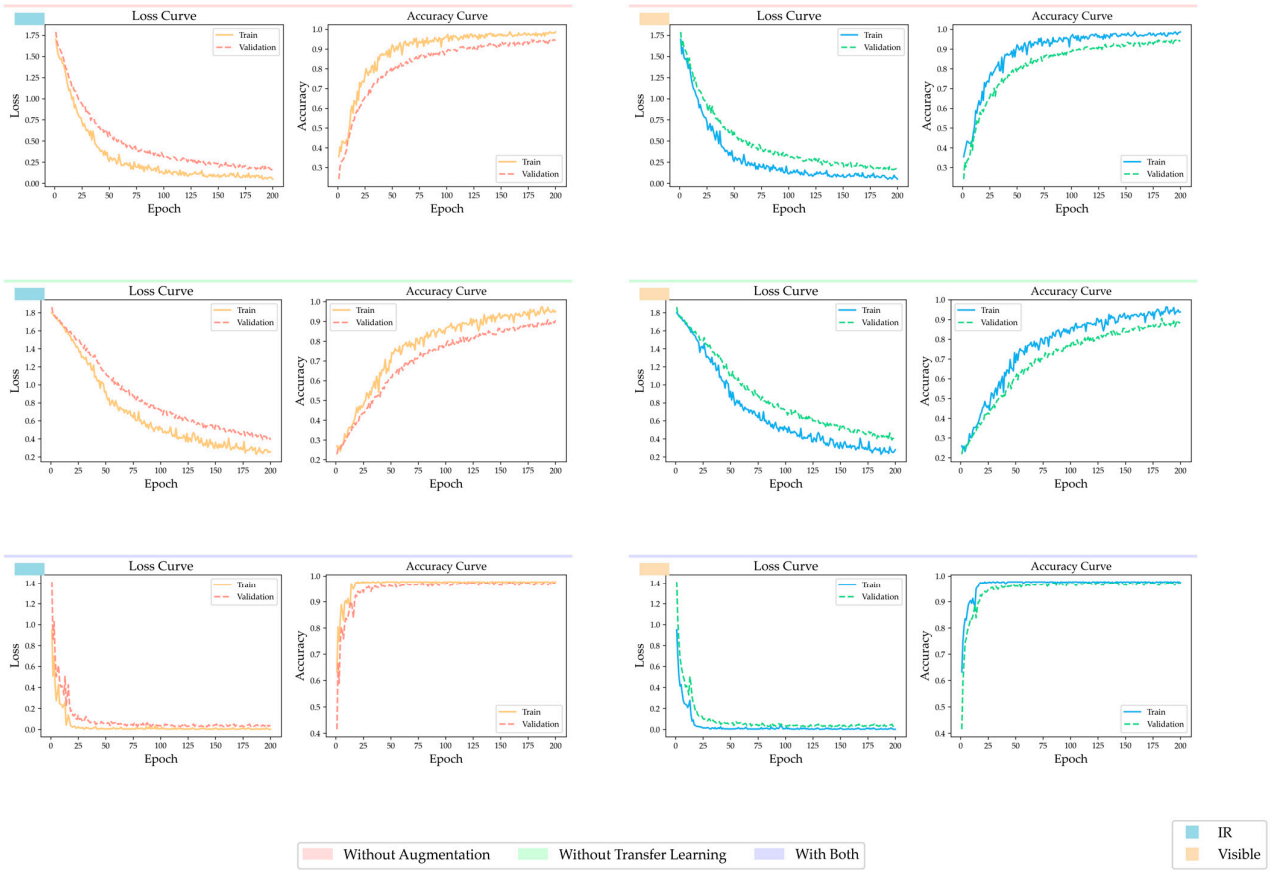


Figure 9. Training and validation curves of VIOS-Net on infrared and visible data sources (without data augmentation, without transfer learning, and with both techniques).

5.3. Shared Feature Extraction Impact Study

As depicted in Table 7, VIOS-Net achieved a 25.688% improvement in accuracy over the VIOS-Visible variant and a 28.047% improvement over the VIOS-Infrared variant. In terms of the F1-score, VIOS-Net showed a 23.678% improvement over VIOS-Visible and a 27.075% improvement over VIOS-Infrared. These results indicate that both visible and infrared features are crucial for accurate target recognition, and that VIOS-Net effectively enhances fusion network performance by extracting shared feature representations from both datasets.

To further validate the efficiency of the shared feature extractor (SFE) in reducing model complexity, an additional experiment was conducted. In this experiment, the SFE was removed, resulting in the VIOS-SFE model. The findings, as presented in Table 7, show that removing the SFE from ResNet-34 increased the number of parameters by 48.818% compared to VIOS-Net. VIOS-Net's SFE module allows for more efficient feature extraction with fewer parameters, reducing model complexity, training time, and overfitting risks. Additionally, fewer parameters translate to lower computational and storage costs during both the training and inference phases. Among the tested models, VIOS-Net strikes the best balance between parameter reduction and accuracy improvement.

Table 7. Comparison of parameters across different methods.

Method	Alex Net	VGG-11	VGG-13	GoogLeNet	ResNet-18	ResNet-34	MobileNetV2	MobileNetV3 Small	ShuffleNetV2_x0_5	SqueezeNet	ConvNeXt	VIOS-Net (OURS)
Visible	61.10M	128.14M	128.32M	6.99M	11.68M	21.79M	3.96M	662.032k	641.46k	6.49M	2.81M	11.11M
Infrared	61.10M	128.14M	128.32M	6.99M	11.68M	21.79M	3.96M	662.032k	641.46k	6.49M	2.81M	11.11M
Total Parameters	122.20M	256.38M	256.65M	13.99M	23.39M	43.59M	7.92M	1.32M	1.28M	12.98M	5.62M	22.31M
Acc	87.607	83.672	84.525	87.083	88.328	89.247	85.509	88.133	88.189	81.184	87.215	96.199
Parameter/Acc (M/%)	1.39	3.06	3.04	0.16	0.26	0.49	0.09	0.01	0.01	0.16	0.06	0.23

As the scale of visual models continues to expand, their parameter sizes frequently reach billions or even trillions, imposing extremely high demands on computational and storage resources. Against this backdrop, the improvement in parameter efficiency is manifested in three key dimensions. In terms of computational cost, a lower parameter density (Params/Accuracy) can significantly reduce the amount of computation, effectively alleviating the computational pressure brought by large-scale models. Regarding storage efficiency, reducing parameter density not only avoids high-cost backpropagation in large-scale backbone networks, thereby substantially saving GPU memory to meet the storage constraints of edge computing devices (such as the shipborne NVIDIA Jetson Nano), but also, in practical applications involving large-scale models and numerous downstream tasks, changes the situation where full fine-tuning requires storing massive model weights for each task. By only storing a small number of parameters, it significantly reduces storage burden and costs. In terms of training efficiency, compared with the current mainstream lightweight methods, VIOS-Net achieves a parameter density of 0.232M/% while maintaining 96.199% classification accuracy, which demonstrates the superiority of this method in balancing accuracy and cost. These cost advantages are of great significance for practical deployment, enabling better promotion of the application and implementation of related technologies in real-world scenarios.

5.4. Weight Coefficient Impact Study

The impact of adjusting the weighting coefficients on the performance of VIOS-Net was examined using the VIOS-Weight variant. As shown in Table 8 and Figure 10, altering the weight settings affected model accuracy by up to 4.150% for the visible light dataset and 4.650% for the infrared dataset. Figure 10 illustrates the weighting pattern, which shows that by fine-tuning the weight coefficients, the model can better manage the learning progress of both tasks, leading to optimal performance in both visible and infrared image recognition tasks.

Table 8. Comparison of quantitative evaluation of variations in weight coefficients.

Weight Coefficients Pair	0.1,0.9	0.2,0.8	0.3,0.7	0.4,0.6	0.5,0.5	0.6,0.4	0.7,0.3	0.8,0.2	0.9,0.1
Visible accuracy (%)	92.049	95.569	96.059	96.199	96.019	95.739	95.739	95.589	92.049
Infrared accuracy (%)	91.549	95.269	95.879	96.199	95.899	95.669	95.549	95.489	95.349

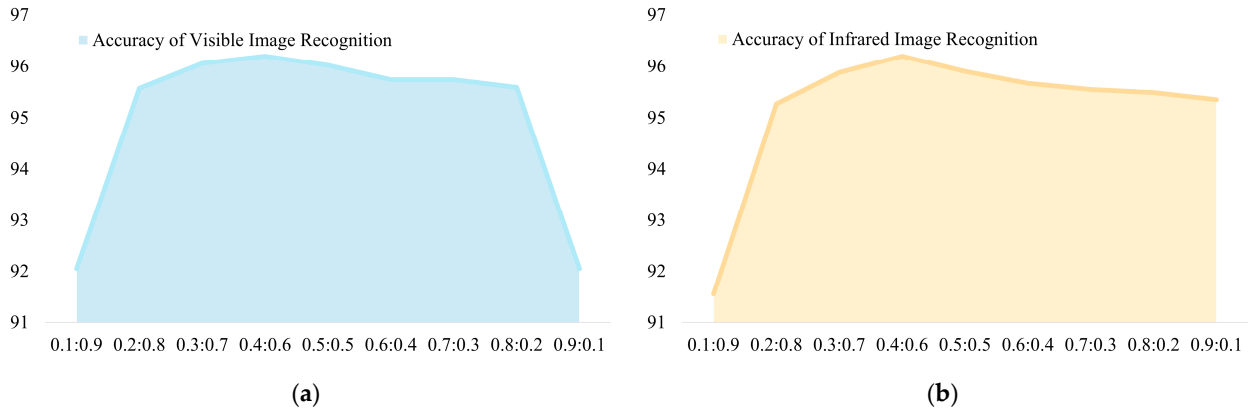


Figure 10. Pattern and impact of weight coefficient (**a**) on the visible light dataset and (**b**) on the infrared dataset.

6. Conclusions

In this paper, we proposed a novel automatic ship monitoring model that leverages the fusion of multiple information sources to enhance maritime management effectiveness under various lighting conditions and adverse weather such as snow, rain, frost, dust, and haze. This model addresses the challenges posed by traditional visible light cameras in accurately identifying ship types. Specifically, the proposed method integrates a preprocessing and data augmentation module, a shared feature extraction module, and a specific feature extraction module.

The preprocessing and data augmentation module is designed to enhance the quality of the training data, thereby improving the model training process and its overall effectiveness. The shared feature extraction module extracts two sets of highly correlated features from multi-source data, enabling the complementary exchange of information between data sources through joint augmentation and weight sharing. This bidirectional feature extraction process significantly improves accuracy by 25.688% and 28.047% compared to separate training on visible and infrared datasets while reducing the number of parameters of the ResNet-34 backbone network by 48.818%. The specific feature extraction module further optimizes the fused features from both information sources, tailoring the dual-channel network for superior performance. Additionally, this study introduces a task weighting factor, uncovering the relationship between accuracy and the task weighting setup, thereby achieving targeted specific feature optimization. This approach effectively compensates for the limitations of visible light cameras in real-world environments, contributing valuable theoretical insights toward the realization of all-weather maritime ship management.

To further enhance the efficiency of maritime management, future research will explore two key directions: Firstly, we will investigate the incorporation of adaptive augmentation learning techniques during the fusion process of multiple information sources. This will allow the model to dynamically adjust data weights and feature extraction strategies in response to environmental changes, thereby improving the model's robustness and generalization capabilities. Secondly, we will optimize the sensor arrangement and fusion strategies for image data within the same modality to enhance data complementarity and consistency, ultimately improving the overall performance of the model. Furthermore, the feature-sharing module will be extended to sensor data from different modalities, such as radar and sonar, to develop a more comprehensive multi-modal deep learning model capable of enhancing ship monitoring in more complex environments.

Thanks to the information-sharing and joint augmentation capabilities of our proposed VIOS-Net, it maintains excellent feature extraction performance even when processing data from multiple sources. This significantly mitigates the limitations of optical sensors under varying lighting conditions and adverse weather, offering valuable insights for the achievement of continuous, all-weather ship monitoring.

Author Contributions: Conceptualization, methodology, resources, software, validation, supervision, and writing—original draft preparation, J.Z.; methodology, conceptualization, software, validation, supervision, and writing—review and editing, J.L.; data curation and validation, J.S.; validation and software, L.W.; funding acquisition, methodology, and supervision, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Young Innovative Talents Grants Program of Guangdong Province (Grant No. 2022KQNCX024), the Ocean Young Talent Innovation Program of Zhanjiang City (Grant No. 2022E05002), the National Natural Science Foundation of China (Grant No. 52171346), the Natural Science Foundation of Guangdong Province (Grant No. 2021A1515012618), the program for scientific research start-up funds of Guangdong Ocean University, the College Student Innovation Team of Guangdong Ocean University (Grant No. CXTD2024018), the Student Innovation Team of

Port Industry TechAcademy (Grant No. GHCY2024008), and the Innovation and Entrepreneurship Training Program for College Students (Grant No. CXXL2024221).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Wang, Z. biSAMNet: A Novel Approach in Maritime Data Completion Using Deep Learning and NLP Techniques. *J. Mar. Sci. Eng.* **2024**, *12*, 868. [CrossRef]
2. Zhang, S.; Chen, J.; Wan, Z.; Yu, M.; Shu, Y.; Tan, Z.; Liu, J. Challenges and Countermeasures for International Ship Waste Management: IMO, China, United States, and EU. *Ocean Coast. Manag.* **2021**, *213*, 105836. [CrossRef]
3. Li, J.; Sun, J.; Li, X.; Yang, Y.; Jiang, X.; Li, R. LFLD-CLbased NET: A Curriculum-Learning-Based Deep Learning Network with Leap-Forward-Learning-Decay for Ship Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 1388. [CrossRef]
4. Sun, J.; Li, J.; Li, R.; Wu, L.; Cao, L.; Sun, M. Addressing unfamiliar ship type recognition in real-scenario vessel monitoring: A multi-angle metric networks framework. *Front. Mar. Sci.* **2025**, *11*, 1516586. [CrossRef]
5. Feng, Y.; Yin, H.; Zhang, H.; Wu, L.; Dong, H.; Li, J. Independent Tri-Spectral Integration for Intelligent Ship Monitoring in Ports: Bridging Optical, Infrared, and Satellite Insights. *J. Mar. Sci. Eng.* **2024**, *12*, 2203. [CrossRef]
6. Li, X.; Li, D.; Liu, H.; Wan, J.; Chen, Z.; Liu, Q. A-BFPN: An Attention-Guided Balanced Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2022**, *14*, 3829. [CrossRef]
7. Li, W.; Ning, C.; Fang, Y.; Yuan, G.; Zhou, P.; Li, C. An Algorithm for Ship Detection in Complex Observation Scenarios Based on Mooring Buoys. *J. Mar. Sci. Eng.* **2024**, *12*, 1226. [CrossRef]
8. Połap, D.; Włodarczyk-Sielicka, M.; Wawrzyniak, N. Automatic Ship Classification for a Riverside Monitoring System Using a Cascade of Artificial Intelligence Techniques Including Penalties and Rewards. *ISA Trans.* **2022**, *121*, 232–239. [CrossRef]
9. Baek, W.-K.; Kim, E.; Jeon, H.-K.; Lee, K.-J.; Kim, S.-W.; Lee, Y.-K.; Ryu, J.-H. Monitoring Maritime Ship Characteristics Using Satellite Remote Sensing Data from Different Sensors. *Ocean Sci. J.* **2024**, *59*, 8. [CrossRef]
10. Šaparnis, L.; Rapalis, P.; Daukšys, V. Ship Emission Measurements Using Multirotor Unmanned Aerial Vehicles: Review. *J. Mar. Sci. Eng.* **2024**, *12*, 1197. [CrossRef]
11. Jiang, X.; Li, J.; Huang, Z.; Huang, J.; Li, R. Exploring the performance impact of soft constraint integration on reinforcement learning-based autonomous vessel navigation: Experimental insights. *Int. J. Nav. Archit. Ocean. Eng.* **2024**, *16*, 100609. [CrossRef]
12. Li, J.; Jiang, X.; Zhang, H.; Wu, L.; Cao, L.; Li, R. Multi-joint adaptive control enhanced reinforcement learning for unmanned ship. *Ocean Eng.* **2025**, *318*, 120121. [CrossRef]
13. Kim, H.; Kim, D.; Park, B.; Lee, S.-M. Artificial Intelligence Vision-Based Monitoring System for Ship Berthing. *IEEE Access* **2020**, *8*, 227014–227023. [CrossRef]
14. Wang, Y.; Wang, B.; Huo, L.; Fan, Y. GT-YOLO: Nearshore Infrared Ship Detection Based on Infrared Images. *J. Mar. Sci. Eng.* **2024**, *12*, 213. [CrossRef]
15. Wang, R.; Miao, K.; Deng, H.; Sun, J. Application of Computer Image Recognition Technology in Ship Monitoring Direction. In *New Approaches for Multidimensional Signal Processing*; Kountchev, R., Mironov, R., Nakamatsu, K., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 270, pp. 233–241. ISBN 9789811685576.
16. Zhang, Z.; Zhang, L.; Wu, J.; Guo, W. Optical and Synthetic Aperture Radar Image Fusion for Ship Detection and Recognition: Current State, Challenges, and Future Prospects. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 132–168. [CrossRef]
17. Wang, J.; Li, S. Maritime Radar Target Detection in Sea Clutter Based on CNN With Dual-Perspective Attention. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
18. Wang, J.; Liu, J.; Zhao, W.; Xiao, N. Performance Analysis of Radar Detection for Fluctuating Targets Based on Coherent Demodulation. *Digit. Signal Process* **2022**, *122*, 103371. [CrossRef]
19. Chen, X.; Su, N.; Huang, Y.; Guan, J. False-Alarm-Controllable Radar Detection for Marine Target Based on Multi Features Fusion via CNNs. *IEEE Sens. J.* **2021**, *21*, 9099–9111. [CrossRef]
20. Su, N.; Chen, X.; Guan, J.; Huang, Y. Maritime Target Detection Based on Radar Graph Data and Graph Convolutional Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
21. Jing, H.; Cheng, Y.; Wu, H.; Wang, H. Radar Target Detection With Multi-Task Learning in Heterogeneous Environment. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
22. Li, W.; Wang, K.; You, L.; Huang, Z. A New Deep Learning Framework for HF Signal Detection in Wideband Spectrogram. *IEEE Signal Process. Lett.* **2022**, *29*, 1342–1346. [CrossRef]
23. Lin, H.; Dai, X. Real-Time Multisignal Detection and Identification in Known and Unknown HF Channels: A Deep Learning Method. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 1–29. [CrossRef]

24. Li, C.; Liu, Y.; Wang, Y.; Ye, X. Short-Term Prediction Method of HF Frequency Based on Deep Learning Network. In Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), IEEE, Dublin, Ireland, 6–9 October 2019; pp. 360–363.
25. Li, H.; Jiao, H.; Yang, Z. AIS Data-Driven Ship Trajectory Prediction Modelling and Analysis Based on Machine Learning and Deep Learning Methods. *Transp. Res. Part E Logist. Transp. Rev.* **2023**, *175*, 103152. [CrossRef]
26. European Maritime Safety Agency. Annual Overview of Marine Casualties and Incidents 2024. 2024. Available online: https://www.emsa.europa.eu/publications/reports/item/5352-annual-overview-of-marine-casualties-and-incidents-2024.html (accessed on 27 April 2025).
27. Liu, B.; Xiao, Q.; Zhang, Y.; Ni, W.; Yang, Z.; Li, L. Intelligent Recognition Method of Low-Altitude Squint Optical Ship Target Fused with Simulation Samples. *Remote Sens.* **2021**, *13*, 2697. [CrossRef]
28. Meng, H.; Tian, Y.; Ling, Y.; Li, T. Fine-Grained Ship Recognition for Complex Background Based on Global to Local and Progressive Learning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
29. Xu, M.-Z.; Yao, Z.-X.; Kong, X.-P.; Xu, Y.-C. Ships Classification Using Deep Neural Network Based on Attention Mechanism. In Proceedings of the 2021 OES China Ocean Acoustics (COA), IEEE, Harbin, China, 14 July 2021; pp. 1052–1055.
30. Zhang, T.; Zhang, X.; Shi, J.; Wei, S.; Wang, J.; Li, J.; Su, H.; Zhou, Y. Balance Scene Learning Mechanism for Offshore and Inshore Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
31. Wu, P.; Huang, H.; Qian, H.; Su, S.; Sun, B.; Zuo, Z. SRCANet: Stacked Residual Coordinate Attention Network for Infrared Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
32. Li, Y.; Liu, G.; Xiong, J.; Jia, M.; Zhang, W. Infrared Ship Detection Based on Latent Low-Rank Decomposition and Salient Feature Fusion. In *Proceedings of the 5th International Conference on Computer Information Science and Application Technology (CISAT 2022)*; Zhao, F., Ed.; SPIE: Chongqing, China, 2022; p. 233.
33. Chen, X.; Qiu, C.; Zhang, Z. A Multiscale Method for Infrared Ship Detection Based on Morphological Reconstruction and Two-Branch Compensation Strategy. *Sensors* **2023**, *23*, 7309. [CrossRef]
34. Li, L.; Yu, J.; Chen, F. TISD: A Three Bands Thermal Infrared Dataset for All Day Ship Detection in Spaceborne Imagery. *Remote Sens.* **2022**, *14*, 5297. [CrossRef]
35. Wu, P.; Su, S.; Tong, X.; Guo, R.; Sun, B.; Zuo, Z.; Zhang, J. SARFB: Strengthened Asymmetric Receptive Field Block for Accurate Infrared Ship Detection. *IEEE Sens. J.* **2023**, *23*, 5028–5044. [CrossRef]
36. Zeng, L.; Zhu, Q.; Lu, D.; Zhang, T.; Wang, H.; Yin, J.; Yang, J. Dual-Polarized SAR Ship Grained Classification Based on CNN With Hybrid Channel Feature Loss. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
37. Li, J.; Qu, C.; Shao, J. Ship Detection in SAR Images Based on an Improved Faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), IEEE, Beijing, China, 21–25 May 2017; pp. 1–6.
38. Zhenzhen, L.; Baojun, Z.; Linbo, T.; Zhen, L.; Fan, F. Ship Classification Based on Convolutional Neural Networks. *J. Eng.* **2019**, *2019*, 7343–7346. [CrossRef]
39. Ren, Y.; Wang, X.; Yang, J. Maritime Ship Recognition Based on Convolutional Neural Network and Linear Weighted Decision Fusion for Multimodal Images. *Math. Biosci. Eng.* **2023**, *20*, 18545–18565. [CrossRef] [PubMed]
40. Chang, L.; Chen, Y.-T.; Hung, M.-H.; Wang, J.-H.; Chang, Y.-L. YOLOV3 Based Ship Detection in Visible and Infrared Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, Brussels, Belgium, 11 July 2021; pp. 3549–3552.
41. Li, J.; Yang, Y.; Li, X.; Sun, J.; Li, R. Knowledge-Transfer-Based Bidirectional Vessel Monitoring System for Remote and Nearshore Images. *J. Mar. Sci. Eng.* **2023**, *11*, 1068. [CrossRef]
42. Pan, B.; Jiang, Z.; Wu, J.; Zhang, H.; Luo, P. Ship Recognition Based on Active Learning and Composite Kernel SVM. In *Advances in Image and Graphics Technologies*; Tan, T., Ruan, Q., Wang, S., Ma, H., Di, K., Eds.; Communications in Computer and Information Science; Springer: Berlin, Germany, 2015; Volume 525, pp. 198–207. ISBN 978-3-662-47790-8.
43. Chang, L.; Chen, Y.-T.; Wang, J.-H.; Chang, Y.-L. Modified Yolov3 for Ship Detection with Visible and Infrared Images. *Electronics* **2022**, *11*, 739. [CrossRef]
44. Wang, Y.; Wang, C.; Zhang, H. Combining a Single Shot Multibox Detector with Transfer Learning for Ship Detection Using Sentinel-1 SAR Images. *Remote Sens. Lett.* **2018**, *9*, 780–788. [CrossRef]
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
46. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
47. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, 8–10 June 2015; pp. 1–9.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

49. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.-C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Republic of Korea, 7–10 October 2019; pp. 1314–1324.
50. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11218, pp. 122–138. ISBN 978-3-030-01263-2.
51. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-Level Accuracy with 50× Fewer Parameters and <0.5 MB Model Size. *arXiv* **2016**, arXiv:1602.07360.
52. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. *arXiv* **2022**, arXiv:2201.03545.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.