Monocular Vision-Based 3D Ship Detection: From Image Pixels to Spatial Perception

Zhe Kang [1,2,3], Feng Ma [1,2*], He Li [3], Chen Chen [4,5], Zaili Yang [3], Jin Wang [3]

1. State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan 430063, PR China

2. Intelligent Transportation Systems Research Center, Wuhan University of Technology, Wuhan 430063, PR China

3. School of Engineering, Liverpool John Moores University, Liverpool, L3 3AF, UK

4. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, PR China

5. Nanjing Smart Water Transportation Technology Co. Ltd, Nanjing 210028, PR China

Abstract: In this paper a novel monocular 3D object detection algorithm, namely MonoStrong, is constructed as a new solution to enhancing the navigation safety for general and autonomous ships in particular with the assistance of visual detections. Initially, a channel attention-enhanced feature network is designed to adaptively strengthen feature channels corresponding to critical geometric structures, improving the representation of complex ship characteristics, particularly for feature-sparse ships at medium and long distances. Subsequently, a multi-scale feature fusion module with attention mechanisms is formulated to enhance the robustness of perception through cross-level feature interaction, further strengthening feature representation for ships with scale variations. An optimized 2D–3D bounding box association loss function is then constructed to effectively reduce errors caused by rapid heading variations at short distances and localization drift at extended ranges. A multi-scenario 3D visual dataset for ships, called Kitti3D-Ship, is constructed to validate the proposed technique. The results demonstrate that MonoStrong outperforms existing methods in ship attitude estimation, particularly in long-range, high-density, and multi-scale scenarios. The method offers a reliable solution for intelligent maritime perception and contributes to the safe operation of ships.

Keywords: Maritime safety, Monocular vision; 3D object detection; Channel attention; Feature fusion

Corresponding authors at: State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan 430063, PR China (F. Ma). E-mail addresses: martin7wind@whut.edu.cn (F. Ma)

1. Introduction

Ship navigation safety lies in accurate perception of the dynamic environment, such as spatial information in terms of nearby ships (Bläser et al., 2024). Precise perception represents the foundation of collision avoidance and navigation risk prevention, which calls for accurate target observation (for target identification), size estimation (for maneuverability assessment), and heading monitoring (for early warning of abnormal behaviors) (Liu et al., 2025). The development of spatial perception algorithms, however, transforms the ship operation from "Reactive Avoidance" to a "Proactive Prediction" manner.

Multi-modal sensor fusion, including LiDAR, radar, Automatic Identification System (AIS), and visual sensing have been the fundamentals for navigation perception. To be specific, LiDAR provides high-precision point cloud and functions in less than 500 meters, and its effectiveness is limited by increased cost in the mid-to-long-range perception scenarios; Radar operates at all-weather conditions but suffers from high false alarm rates in complex weather conditions (Ma et al., 2024); AIS system generates comprehensive monitoring data for navigation perception, but, in real application scenarios, it is challenged by data errors, update delays, vulnerability to spoofing attacks, and high maintenance costs (Li et al., 2022; Xin et al., 2025).

Visual perception, simulating the human lookout mechanism in maritime navigation, offers unique advantages of navigation perception, like instantaneous spatial awareness, intuitive judgment of ship behavior, and distinct adaptability to lighting changes. Powered by deep learning-based feature extraction and pattern recognition, vision systems are able to exploit salient visual cues of ships to achieve robust target separation in complex backgrounds/environments, at relatively low hardware cost (Meng et al., 2023). Specifically, 2D perception focuses on analyzing targets within the image (e.g., detection, classification, segmentation), enabling identification and localization in the camera view, while, 3D perception aims to reconstruct spatial dimensions (e.g., estimating bearing, distance, size, and motion in real-world coordinates). Although network-based 2D ship detection shows promising performance, its inability to provide rich motion dynamics severely constrains practical deployment in shipboard perception systems. The 3D perception, however, provides crucial geometric understanding for situational awareness and collision avoidance, it is noted that Monocular 3D object detection has achieved notable progress with promising performance on

Corresponding authors at: State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan 430063, PR China (F. Ma). E-mail addresses: martin7wind@whut.edu.cn (F. Ma)

general-purpose datasets such as KITTI (Ge Geiger et al., 2012), nuScenes (Caesar et al., 2020), and Waymo (Sun et al., 2020), etc. However, its application to ship perception faces significant challenges as follows, wanting new solutions to be found:

(1) The inherent physical characteristics of ships (e.g., aspect ratio variations) induce scale sensitivity and feature ambiguity. Methods like MonoCon (Liu et al., 2022) and MonoFlex (Zhang et al., 2021) rely on size priors deviating from real ship parameters, causing gradient vanishing in 3D bounding box regression. Further exacerbated by scarce ship samples in general datasets, model overfitting limits cross-scenario generalization.

(2) Distant ship detection suffers from sparse pixels and the uncertainty in monocular depth, while close-range scenarios face multi-scale ships and high inter-class variation. Specifically, ship type diversity and varying movement distances complicate feature extraction, significantly challenging spatial localization and heading prediction.

To this end, this paper proposes a novel detection algorithm, namely MonoStrong, to realizing the monocular 3D object detection of ships. The rest of this paper is organized as follows: Section 2 reviews state-of-the-art methods in vision-based perception studies. Section 3 presents the proposed ship 3D detection algorithm. Section 4 provides the results. Section 5 concludes the research.

2. State of the art

This section analyses the development of vision-based 3D object detection and feature extraction techniques such that to provide a framework for 3D detection of ships.

2.1. Vision-Based 3D Object Detection

Monocular 3D object detection has become a solution for cost-effective and efficient perception in autonomous driving (Wang et al., 2021). However, the core challenge is the irreversible loss of depth information from a single image, which leads to ambiguity in 3D spatial localization. To this end, end-to-end mapping, depth-assisted optimization, and implicit prior guidance approaches are developed. Specifically, the end-to-end mapping links image pixels to 3D parameters through global modeling. For instance, MonoATT employs a lightweight network to adaptively identify key image regions and dynamically extract multi-granularity primitive features, enhancing geometric cues for foreground objects (Zhou et al., 2023); Depth-assisted optimization uses pre-trained depth estimation models or temporal information to achieve explicit supervision. BA-Det, for instance, leverages multi-frame trajectory information from long video sequences and extracts temporal region-of-interest (RoI)

features to achieve global optimization (He et al., 2023); Implicit prior guidance integrates geometric constraints or domain knowledge as a basis to enhance spatial reasoning. Notably, BEVHeight replaces depth estimation with pixel-wise height prediction to prevent localization failure caused by negligible depth differences between distant object tops and the ground (Yang et al., 2023). Moreover, to address the high cost of 3D annotation, Weak-Mono3D introduces a weakly supervised framework that requires only 2D bounding boxes and orientation labels for training, reducing the annotation burden (Tao et al., 2023).

Stereo vision techniques leverage the geometric constraints of binocular cameras to build a physically interpretable framework for depth perception. The main approaches can be categorized into: (1) 2D detection-driven methods reconstruct 3D objects by aligning RoI features, achieving high computational efficiency at the cost of reliance on stereo matching (Qin et al., 2021; Pon et al., 2020); (2) pseudo-LiDAR methods convert depth maps generated via stereo matching into LiDAR-like point clouds, enabling the reuse of existing point cloud detection algorithms (Qian et al., 2020; Meng et al., 2024); (3) differentiable volumetric feature models (e.g., DSGN++) directly encode geometric and semantic information in 3D space, optimizing the propagation of 2D features to 3D through sliding slicing operations (Chen et al., 2022). In addition, multi-view detection techniques, based on surround-view camera systems, achieve panoramic perception by projecting multi-view features into a unified bird's-eye view (BEV) space (Philion et al., 2020; Li et al., 2023; Zou et al., 2024). Although stereo or multi-view vision techniques offer advantages in spatial completeness and feature complementarity, their adoption is hindered by challenges such as complex multi-camera synchronization and calibration, computational burden from high-resolution BEV features, and cost caused by sensor redundancy.

2.2. Visual Feature Extraction

Traditional vision feature techniques (Kelvin et al., 2010; Felzenszwalb et al., 2010) require substantial expert knowledge while demonstrating constrained generalization, especially in complex scenarios. In contrast, contemporary deep learning-based methods, including ResNet (He et al., 2016) and Transformer (Vaswani et al., 2017) have significantly improved feature extraction capabilities in instance segmentation and object detection (Sigirci et al., 2022).

Object detection methods can be broadly categorized into two-stage (Girshick, 2015; He et al., 2017) and one-stage approaches (Duan et al., 2019; Zhao et al., 2024). Due to the application of novel computational networks, one-stage algorithms have gradually become mainstream. For example, Object Detection with Transformers (DETR) achieves end-to-end

detection, eliminating some conventional post-processing steps (Carion et al., 2020); YOLOv8 uses a cross-stage block to incorporate dynamic label assignment, achieving accuracy comparable to two-stage models (Sohan et al., 2024). In addition, the PP-YOLO series balance speed and performance through deformable convolution and optimized feature pyramid networks (Long et al., 2020). Meanwhile, rotation-aware detection models like H2RBox-v2 address complex annotations via weakly supervised learning, promoting the deployment of detection algorithms in applications (Yu et al., 2023).

In terms of deep learning-based visual perception, the early YOLO series evolved based on the Darknet backbone. To be specific, YOLOv2 adopted Darknet19, a 19-layer network with batch normalization, enabling fast inference but showing limited localization accuracy in complex scenes (Redmon et al., 2017); YOLOv3 was upgraded to Darknet53, incorporating residual connections to enhance the representational power of deeper networks while maintaining a good balance between speed and accuracy (Redmon et al., 2018); YOLOv4 with CSPDarknet53 optimizes information flow through a cross-stage partial (CSP) design, improving classification and detection efficiencies (Bochkovskiy et al., 2020). For lightweight solutions, MobileNetV3-Large employed depthwise separable convolutions to reduce parameters, balancing performance and computational cost (Howard et al., 2019). On the other hand, for deep network optimization, ResNet-50/101 alleviated the gradient vanishing problem via residual structures, enhancing perceptual capabilities. ResNeXt further introduced grouped convolutions, improving accuracy through multi-path feature fusion (Xie et al., 2017). The evolution of these networks reflects a multidimensional exploration: from speed prioritization and lightweight adaptation to in-depth optimization. These advancements in 2D perception have not only pushed the boundaries of image understanding but also laid a solid foundation for more complex 3D detection (Fan et al., 2024; Zhang et al., 2023).

2.3 Novel contributions

To this end, this paper proposes a novel detection algorithm, named MonoStrong, to realizing the monocular 3D object detection of ships. The novel contributions include:

(1) Design a channel attention-enhanced feature network to adaptively strengthen key visual feature channels related to 3D ship detection. It enables a comprehensive geometric structural feature capture and improves the capability of representation of ships with complex/various shapes.

(2) Develop a multi-scale feature fusion network with an attention mechanism to enhance discriminative feature representation. This allows the precise fusion of cross-scale feature mapping, thus improving long-range ship positioning.

(3) Optimize the loss functions for 2D and 3D bounding box regression to enhance gradient propagation continuity and improve the heading angle estimation.

(4) Construct a multi-scenario 3D visual perception dataset, named Kitti3D-Ship, to advance ship spatial localization and heading prediction.

3. METHODOLOGY

MonoStrong (see Figure 1) inherits the data processing workflow from MonoLSS (Li et al., 2024) and introduces channel attention modules to strengthen 3D object feature extraction. Meanwhile, through improving the calculation of various loss components, the method effectively boosts the accuracy of core metrics such as depth estimation, object size, and heading prediction.

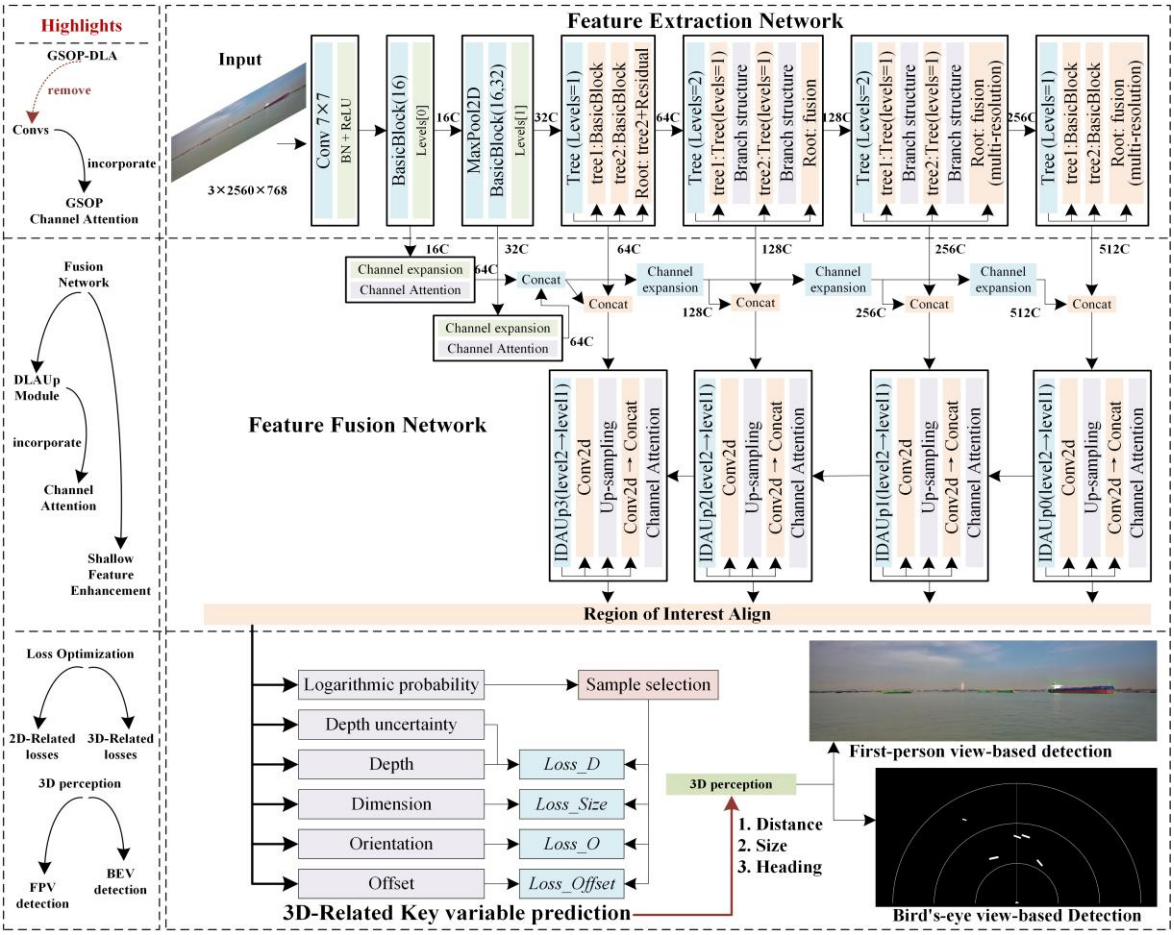


Figure 1. Overall structure of MonoStrong

3.1. Feature extraction network

Ships face challenges such as low contrast, dynamic backgrounds, and small-scale targets (e.g., distant targets), which reduces the ability of standard Deep Layer Aggregation (DLA) networks in distinguishing key semantic features from background noise. To this end, a Global Second-order Pooling (GSOP) channel attention mechanism (Gao et al., 2019) is introduced to dynamically calibrate channel weights and optimize feature representation. Its key advantage lies in suppressing redundant channel responses related to non-target backgrounds (e.g., waves, reflections), while enhancing activations corresponding to discriminative ship structures (e.g., waterlines, cabin edges and bow). This enables the network to focus on highly distinctive regions, as shown in Figure 2. The integration of the GSOP mechanism into the DLA network enhances feature representation by modeling higher-order statistical dependencies between channels through global second-order pooling. The improved architecture strengthens the network's ability in representing 3D object attributes in dense or small-scale targets scenarios.

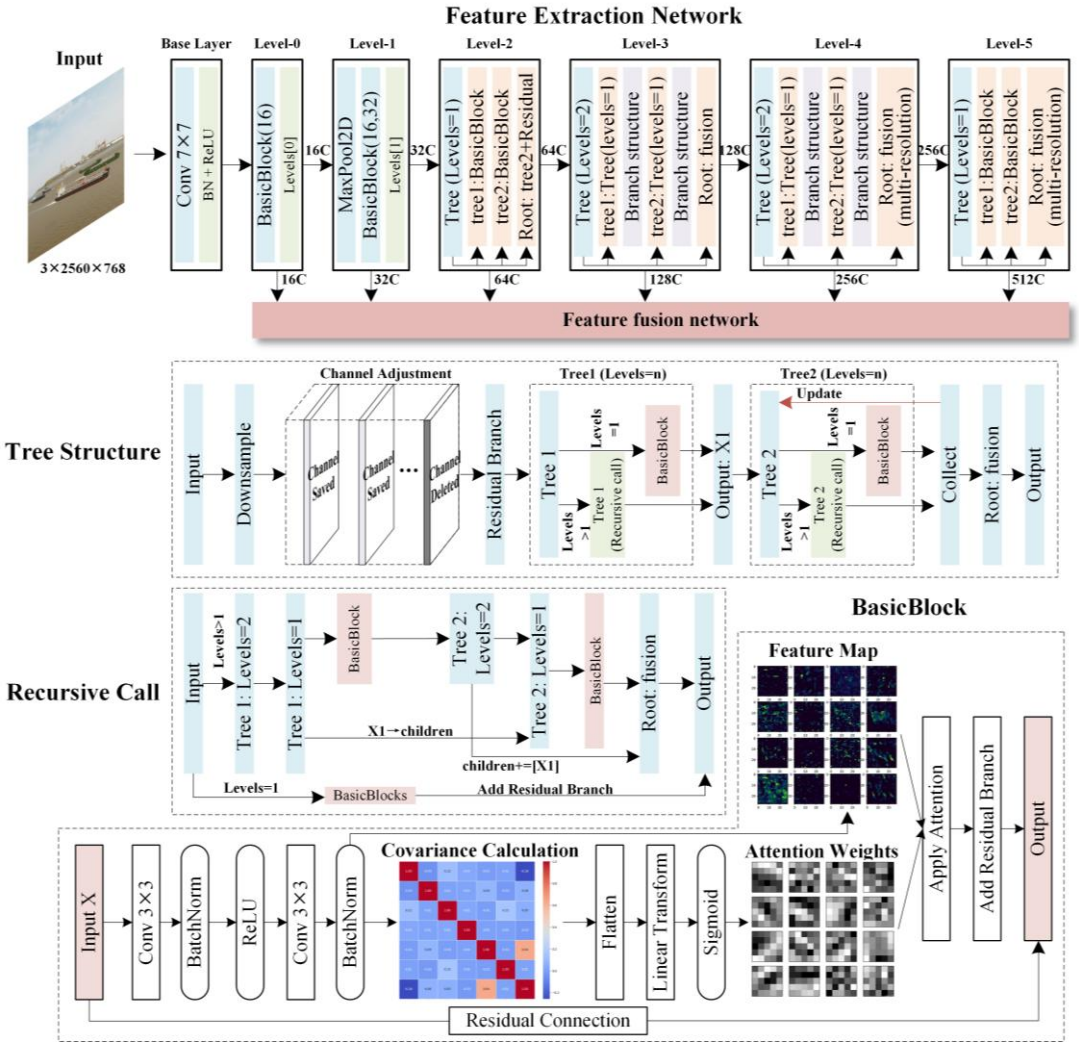


Figure 2. GSOP-DLA Network.

Given an input 3D tensor $X \in \mathbb{R}^{h' \times w' \times c'}$, where $h' \times w'$ represents the spatial dimensions and c' is the number of channels, a 1×1 convolution is applied to reduce the channel dimension, resulting in $Y \in \mathbb{R}^{h' \times w' \times c} (c < c')$. The global second-order pooling operation captures second-order statistical dependencies between features by computing the covariance matrix in channels, as:

$$\Sigma = \frac{1}{h'w'} \sum_{i=1}^{h'} \sum_{j=1}^{w'} Y_{i,j} Y_{i,j}^{T} \tag{1}$$

Here, $Y_{i,j} \in \mathbb{R}^{c}$ represents the feature vector at spatial position (i, j). Subsequently, a convolutional layer combined with a Sigmoid activation function generates the channel weight vector w, which is finally multiplied with the original input tensor X, as:

$$w = \sigma(W_{ori} \cdot vec(\Sigma) + b) \tag{2}$$

$$X_{out} = X \cdot w \tag{3}$$

where $\sigma(\cdot)$ is the standard Sigmoid function, W_{ori} is the original weight matrix of the convolutional layer, vec(Σ) is the operation of flattening the covariance matrix, b is the preset bias, and X_{out} is the final output tensor.

This process completes the recalibration of features by explicitly modeling higher-order statistical relationships between channels, thus overcoming the limitations of traditional convolutions that rely solely on local linear transformations, corresponding to the first new contribution in Section 2.3.

3.2. Feature fusion network

The standard DLAUp network, which used in MonoLSS, achieves progressive multi-scale feature fusion via an inverse hierarchical IDAUp structure and grouped transposed convolutions, providing high-generalization multi-scale feature expressions for 3D object detection in general road scenario. However, in dynamic water environments, ship features are often obscured by low-light areas and dynamic backgrounds, where the standard DLAUp network's uniform feature weighting strategy struggles to distinguish valuable semantic clues from noise.

To address these challenges, this paper introduces a channel attention module, based on the Mish activation function (Misra, 2019), into the DLAUp framework. It adjusts the contribution of cross-level feature channels, suppressing redundant feature responses dominated by background noise (such as water ripples and hull reflections) while enhancing the activation intensity of key target components. For instance, in low-contrast scenarios, the attention mechanism boosts the weight of hull edge channels, improves the separation between

contour features and the background, see Figure 3. The Mish-activated Squeeze-and-Excitation (SE) mechanism models high-order statistical correlations among channels related to depth and scale, and enhances the representation of geometric features for ship targets. Moreover, the smooth gradient property of the Mish function strengthens the network's ability in extracting hierarchical features.

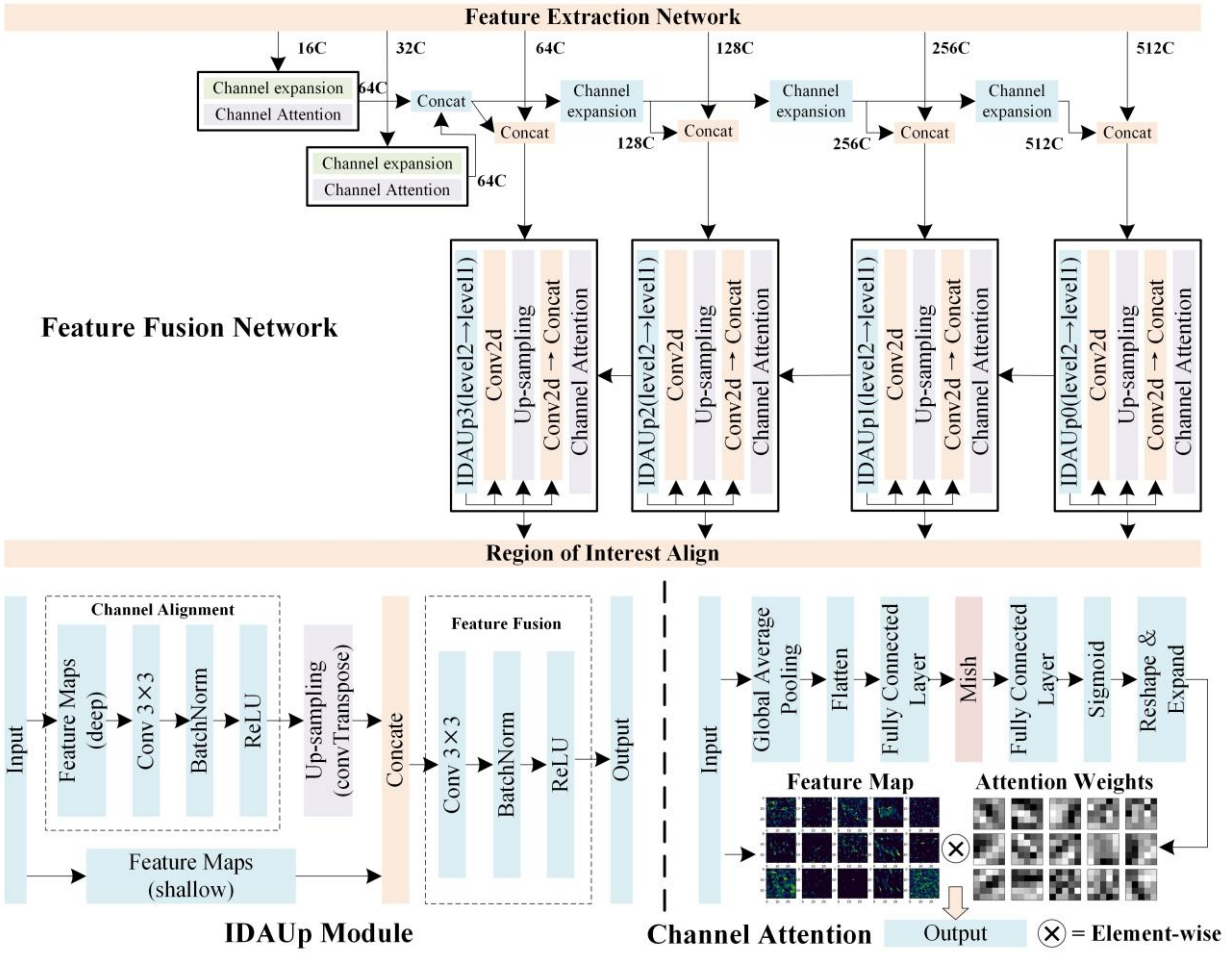


Figure 3. DLAup-SE+ network.

In the DLAup-SE+ module, a channel attention mechanism is constructed to dynamically calibrate feature channel weights through dynamic convolution to enhance multi-scale feature representation. It uses feature map of the *k*-th layer F_k and Global Average Pooling (GAP) to generate channel weight vector w_k via a two-layer fully connected network with Mish activation:

$$w_k = \sigma(W_2 \cdot Mish(W_1 \cdot GAP(F_k))) \tag{4}$$

where W_1 and W_2 are weight parameters in two fully connected layers, $\sigma(\cdot)$ is the Sigmoid function, and the generated weight vector $w_k \in [0,1]$ performs channel-wise feature recalibration through element-wise multiplication, ultimately generating F_{out}, formulated as:

$$F_{out} = F_k \cdot w_k \tag{5}$$

This operation enhances the response intensity of critical channels (such as ship contours) by driving $(w_k \to 1)$, while suppressing background noise interference (such as water surface reflections) by driving $(w_k \to 0)$.

Furthermore, multi-scale features are seamlessly integrated through recursive fusion via the DLAup module. In summary, the resulting output feature map effectively combines multi-scale information through adaptive hierarchical weighting, simultaneously preserving fine-grained shallow details and rich deep semantic representations. This comprehensive feature integration significantly enhances ship detection robustness and target localization precision, reflecting the second contribution in Section 2.3.

3.3. Loss Function Optimization

The optimization of loss functions determines the model's training direction and improves convergence efficiency. MonoStrong employs a "2D+3D" joint optimization loss function framework. Specifically, the 2D bounding box regression loss aims to simultaneously optimize size prediction and offset localization to improve image feature-associated modelling. It designs an improved smooth loss function and an adaptive learning mechanism by the following.

The size prediction error and offset prediction error of the 2D bounding box are constrained using an improved *Smooth L1* loss function. Knowing the prediction value *p* and target value *t*, the loss function L_{smooth} is:

$$L_{smooth}(p,t) = \begin{cases} 0.5 \cdot (p-t)^2 / \beta, & |p-t| < \beta \\ |p-t| - 0.5 \cdot \beta, & otherwise \end{cases} \tag{6}$$

Here, $\beta = 1.0$ is a smooth transition threshold. The quadratic term suppresses gradient oscillations for small errors, while reverting to a linear form for large errors to maintain optimization stability. It is to achieve a balance between gradient smoothness and convergence efficiency.

To further enhance the model's learning capability for occluded and small-scale targets, a dynamic weighting function is introduced whose weight value is generated through a non-linear mapping of the prediction error, as:

$$w(p,t) = \frac{2}{1 + e^{-\gamma |p-t|}} - 1 \tag{7}$$

where, γ is a preset regulatory factor. The weighting function maps the absolute error $|p-t|$ within [0, 1], assigning higher weights to larger errors to amplify their gradient contributions during backpropagation. Accordingly, the size loss L_{size} and offset loss L_{offset} are computed as:

$$L_{size} = \frac{1}{N}\sum_{i=1}^{N} w(p_{size}^{(i)}, t_{size}^{(i)}) \cdot L_{smooth}(p_{size}^{(i)}, t_{size}^{(i)}) \tag{8}$$

$$L_{offset} = \frac{1}{N}\sum_{i=1}^{N} w(p_{offset}^{(i)}, t_{offset}^{(i)}) \cdot L_{smooth}(p_{offset}^{(i)}, t_{offset}^{(i)}) \tag{9}$$

Here, *N* denotes the number of valid samples. $(p_{size}^{(i)}, t_{size}^{(i)})$ represents the predicted and ground truth dimensions of the target 2D bounding box. Similarly, $(p_{offset}^{(i)}, t_{offset}^{(i)})$ represents the predicted and ground truth offsets of the target. The dynamic weighting mechanism enables the model to automatically focus more on samples with larger prediction errors. The weighted sum of both terms forms the total loss $L_{2D} = L_{size} + L_{offset}$. Such an adaptive strategy enhances localization robustness in complex scenarios.

To address the optimization instability caused by random parameter initialization during early training stages, a warm-up strategy is proposed. In the first $T_{warm} = 10$ training epochs, the loss value increases linearly with training progress, where the ten-epoch duration is empirically determined through experimental validation. This process can be modeled as:

$$L_{2D} = (L_{size} + L_{offset}) \cdot \min(1.0, \frac{T_{warm}}{T_{total}}) \tag{10}$$

where T_{total} indicates the total training epochs. 2D object detection loss focuses on bounding box offsets and sizes within the image plane, while 3D box loss jointly models the target's geometric properties (including depth, offset, dimensions, and orientation). First, depth prediction errors are modeled using *Laplace Uncertainty Loss* as a basis to balance the residual distribution between predicted and ground truth values. Let the predicted depth *d*, the ground truth d^*, and the prediction confidence be σ (where higher σ indicates greater uncertainty). The loss function is defined as:

$$L_{depth} = \frac{\sqrt{2}}{\sigma} \cdot \left| d - d^* \right| + \log(\sigma) \tag{11}$$

Here, the first term $\sqrt{2}/\sigma \cdot |d - d^*|$ adjusts the residual weight by confidence σ, causing high-uncertain samples and generating weaker gradients. The second term $\log(\sigma)$ acts as a regularization term to prevent σ from growing excessively. To further enhance semantic consistency in depth prediction, attention mask mechanism is introduced in later training stages.

High-response regions are extracted from the attention map M_{attn}, focusing the model on salient features, as:

$$L_{depth} = \frac{1}{N}\sum_{i=1}^{N} L_{depth}^{(i)} \cdot M_{attn}^{(i)} \tag{12}$$

Here, $L_{depth}^{(i)}$ is the depth loss value of the *i-th* target calculated based on Equation (11). And M_{attn} is generated via Gumbel-Softmax, which suppresses background noise and improves the geometric accuracy of depth estimation.

The prediction errors for 3D offsets o_{3d} and dimensions s_{3d} are defined using an improved weighted loss function. For a predicted value *p* and its corresponding ground truth *t*, the loss function is formulated as:

$$L_{smooth}(p,t) = \begin{cases} 0.5 \cdot (p-t)^2 / \beta, & |p-t| < \beta \\ |p-t| - 0.5 \cdot \beta, & otherwise \end{cases} \tag{13}$$

Here, $\beta = 1.0$ is a common parameter threshold derived from standard mathematical conventions for loss smoothing. Additionally, a dynamic loss adjustment mechanism is constructed, which is expressed as:

$$w(p,t) = \frac{2}{1+e^{-\gamma|p-t|}} - 1 \ (\gamma = 1.0) \tag{14}$$

The offset loss $L_{offset3d}$ and size loss L_{size3d} are independently calculated by:

$$L_{offset3d} = \frac{1}{N}\sum_{i=1}^{N} w(o_{3d}^{(i)}, o_{3d}^{*,(i)}) \cdot L_{smooth}(o_{3d}^{(i)}, o_{3d}^{*,(i)}) \tag{15}$$

$$L_{size3d} = \frac{1}{N}\sum_{i=1}^{N} w(s_{3d}^{(i)}, s_{3d}^{*,(i)}) \cdot L_{smooth}(s_{3d}^{(i)}, s_{3d}^{*,(i)}) \tag{16}$$

Here, $(o_{3d}^{(i)}, o_{3d}^{*,(i)})$ represents the predicted and ground truth values of the target's 3D offsets, respectively. Similarly, $(s_{3d}^{(i)}, s_{3d}^{*,(i)})$ represents the predicted and ground truth values of the target's 3D sizes, respectively.

To mitigate parameter instability during early training, a warm-up strategy is applied, linearly scaling the loss over the first $T_{warm} = 10$ epochs:

$$L_{offset3d} = L_{offset3d} \cdot \min(1.0, \frac{T_{warm}}{T_{total}}) \tag{17}$$

$$L_{size3d} = L_{size3d} \cdot \min(1.0, \frac{T_{warm}}{T_{total}}) \tag{18}$$

Here, the choice of $T_{warm} = 10$ is based on empirical observations that parameter gradients typically stabilize within this timeframe. In addition, T_{total} indicates the total training epoch.

The heading angle prediction is achieved through joint classification and regression. The heading angle is discretized into *K=24* directions, providing 15-degree angular resolution (360°/24) that balances directional precision with computational efficiency, with the classification prediction of $c \in \mathbb{R}^K$. The loss function consists of classification loss and regression loss. For the classification loss, label-smoothed cross-entropy is adopted to mitigate overfitting:

$$L_{cls} = -\sum_{k=1}^{K} (\alpha \cdot \delta(k, k^*) + \frac{1-\alpha}{K}) \log c_k \tag{19}$$

Here, α is set to 0.9 based on experimental experience, which provides an effective balance between maintaining confidence in the correct direction while allowing reasonable uncertainty for neighboring angular classes, and *k* represents the predicted class, and k^* is the ground-truth class. $\delta(\cdot)$ denotes the Kronecker delta function, and c_k represents the predicted probability for the *k* direction.

For the regression loss, a learnable uncertainty parameter σ is introduced, where a higher value indicates greater uncertainty. The weighted L_{reg} is defined as:

$$L_{reg} = \frac{1}{N} \sum_{i=1}^{N} (e^{-\sigma} \left| \Delta\theta^{(i)} - \Delta\theta^{*,(i)} \right| + \sigma) \tag{20}$$

Here, *N* is the total number of samples, $\Delta\theta^{(i)}$ denotes the predicted angle offset for the *i-th* sample, and $\Delta\theta^{*,(i)}$ represents its ground-truth value. The heading loss combines classification and regression through adaptive weighting:

$$L_{heading} = \lambda \cdot L_{cls} + (1-\lambda) \cdot L_{reg} \tag{21}$$

Here, $\lambda = sigmoid(0.1 \cdot (\bar{k} - 6))$ is a balance factor, and $\bar{k}$ is the average classification score.

The total 3D bounding box loss is formulated as a weighted sum of the sub-losses:

$$L_{3D} = L_{depth} + L_{offset3d} + L_{size3d} + L_{heading} \tag{22}$$

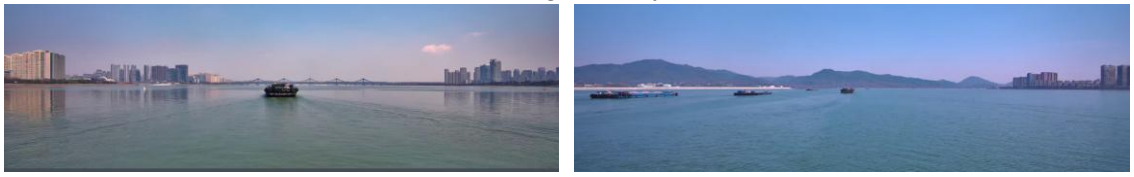
4. CASE STUDY

4.1. Dataset

A monocular 3D vision perception dataset designed for inland ships, Kitti3D-Ship, is applied in this case study. Adhering strictly to the standard KITTI format, Kitti3D-Ship integrates multi-perspective data from shore-based fixed observation points and moving ship platforms. The images were collected successively from February 2025 to April 2025 across the Beijing-Hangzhou Grand Canal (e.g., Hangzhou section, Zhenjiang section, Yangzhou

section), Huai River waters, and the Yangtze River (e.g., Jiangsu section) in China. Typical image examples are shown in Figure 4.

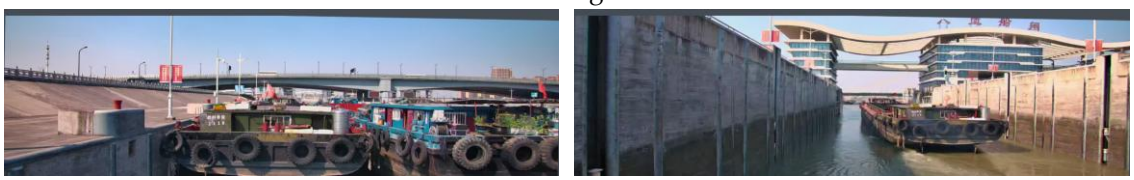
a. Bridge vicinity



b. Unobstructed water area

c. Port waters

d. Berthing area



e. Lock channel

Figure 4. Images of various scenarios in the Kitti3D-Ship dataset

For robust all-weather data collection, low-light and visible-light cameras are deployed for multi-spectral imaging. This setup, however, introduces cross-scene pixel variations. Hence, bilinear interpolation is applied to align image density distributions, thereby enhancing consistency over multi-source images. The dataset comprises 6,950 annotated samples, split into training (60%), validation (30%), and test (10%) sets. Each sample includes a monocular RGB image paired with corresponding labels and calibration files, featuring high-precision 3D ship bounding boxes, heading angles, dimensions, and sensor calibration matrices. The proposed dataset serves as a valuable resource for advancing 3D perception in maritime navigation and autonomous ship research.

4.2. Training optimization

To maintain general visual representation capabilities while adapting to ship features, the shallow convolutional layers of the backbone network were frozen, allowing only fine-tuning of higher-level semantic feature extraction layers. It ensures stable convergence and gradual adaptation to the specific feature distribution of ship targets. An adaptive cosine annealing decay mechanism was implemented to adjust the learning rate following a smooth cosine curve. Formally:

$$\eta_t = \eta_{\min} + 0.5(\eta_{\max} - \eta_{\min})(1 + \cos(\frac{t}{T}\pi)) \quad (23)$$

where T is the total number of training epochs, t represents the current training epoch, with $\eta_{\max}$ and $\eta_{\min}$ are preset upper/lower bounds. This mechanism maintains a high initial rate for rapid convergence, refines gradients via cosine curvature mid-training, and late-stage training fixes the rate at $\eta_{\min}$ to optimal parameters.

To enhance hierarchical learning of ship fine-grained features, a three-phased feature decoupling training framework is proposed. To be detailed, phase 1 trains on a small amount of image samples to establish macro-feature representations (e.g., hull contours and aspect ratios); phase 2 improves generalization by incorporating multi-source sensor images; Phase 3 tackles hard samples (e.g., lighting variations, occlusions, glare) to learn robust features in various scenarios.

4.3. Training

The experiments were conducted on Ubuntu 20.04 with 4×NVIDIA A100 GPUs (120GB VRAM), using CUDA 11.7.0, PyTorch 1.13.1, and Python 3.8. The key preset configurations of methods are listed in Table 1.

Table 1. Initial Training Parameters

Parameters	Value / Setting
Initial Learning Rate	0.0001
Learning Rate Decay	0.1
Weight Decay	0.000001
Batch Size	32
Image Input Size	2560×768
Warmup Enabled	True
Training Epochs	1400
LR Decay Milestones	[400, 800, 1200]

As shown in Figure 5, MonoStrong's training losses (including depth loss, 2D/3D box offset & size loss, and orientation loss) converged to values under the above parameter settings. Notably, the gradient dynamics remained highly coordinated, demonstrating stable training and improved learning effectiveness.

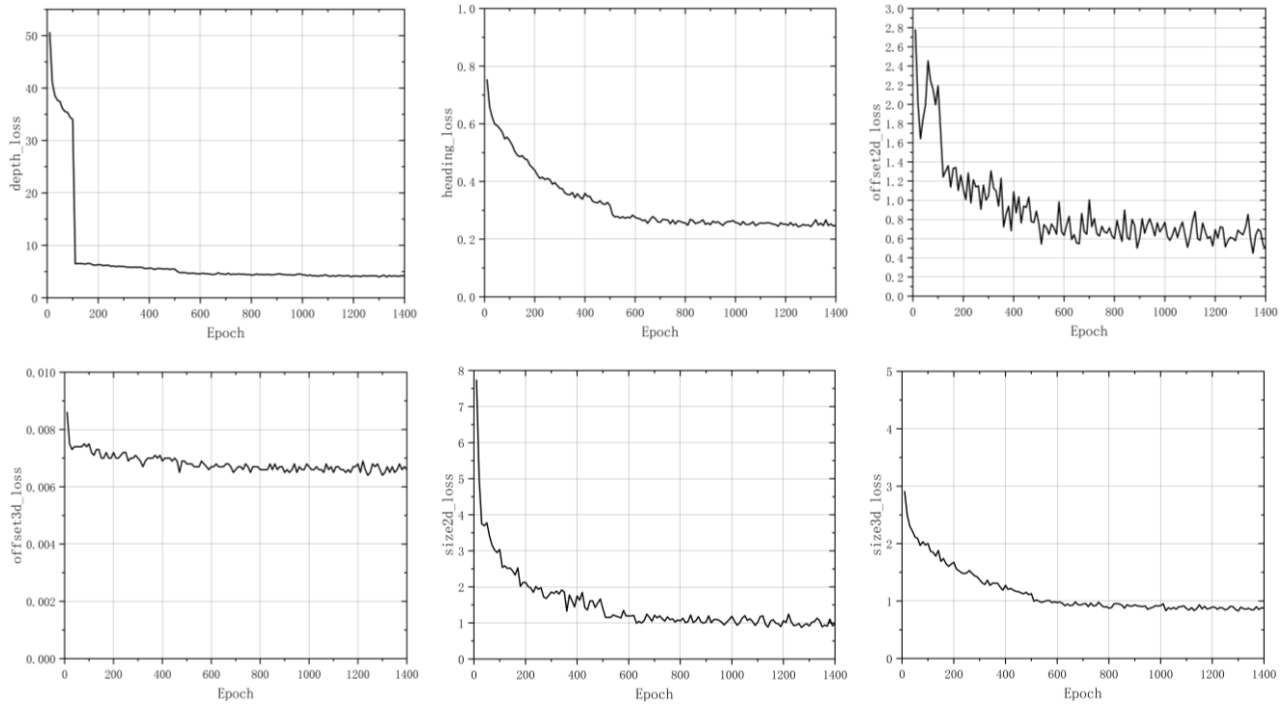


Figure 5. Convergence process of various types of losses

4.4. Comparisons and Discussions

Four core metrics are adopted to verify the performance of the proposed method: 2D detection accuracy (IoU@0.7), heading accuracy, 3D localization (IoU@0.7/0.5/0.3), and BEV localization (IoU@0.7/0.5/0.3) (Li et al., 2024). Additionally, multi-resolution testing to assess scale adaptability, comparative analysis against SOTA methods, ablation studies to quantify module-wise contributions, and visual comparisons highlighting MonoStrong's advantages over baselines.

A. Detection performance with different image resolutions

Input image resolutions affect the accuracy of 3D object detection. The experiment utilized 300 randomly selected images from Kitti3D-Ship (test sub-dataset). The results of MonoStrong under two resolutions (1280×384, 2560×768) are compared with that of MonoLSS, MonoCon, MonoDLE (Ma et al., 2021). 3D localization accuracy (IoU@0.5), BEV localization (IoU@0.7), and heading are indicators of performance, see Table 2. The result indicates that increasing resolution to 2560×768 will improve the performance of all detection models, as evidenced by the 12.97% improvement observed in MonoLSS' 3D localization. Meanwhile,

MonoStrong, leveraging a channel-attention-enhanced feature extraction network, enhanced distance estimation at 2560×768, achieving 89.34% heading accuracy while maintaining 10 FPS on an NVIDIA 3090 GPU. This demonstrates MonoStrong's ability to balance high-resolution geometric modeling with computational constraints.

Table 2. Algorithm Performance in Resolutions (%)

Sizes	Indexes	MonoStrong	MonoCon	MonoDLE	MonoLSS
2560×768	3D	78.11	40.28	33.22	47.10
	BEV	59.41	21.58	18.54	22.27
	Heading	89.34	65.71	55.82	70.24
1280×384	3D	56.97	23.32	22.19	34.13
	BEV	47.52	16.84	9.45	15.28
	Heading	69.74	46.72	41.94	52.14

B. Comparisons

Table 3 lists the 3D perception performance of algorithms on the Kitti3D-Ship (validation sub-dataset) at 2560×768 resolution. MonoStrong achieves a 38.13% localization precision under IoU@0.7, representing a 30% improvement over competing methods. The model attains 59.41% BEV localization accuracy (at IoU@0.7), a 31.14% performance gain compared to MonoLSS. Notably, the model's heading estimation performance surpasses existing methods. This advantage stems from its novel orientation-aware loss function that jointly optimizes global ship structure and local visual features.

Table 3. Performances of algorithms on Kitti3D-Ship Val. Set (%)

Indexes	MonoStrong	MonoFlex	MonoCon	MonoDLE	MonoLSS	Smoke(Liu et al., 2020)
3D (IoU@0.7)	38.13	2.11	8.10	3.37	8.49	3.14
3D (IoU@0.5)	78.11	24.79	40.28	33.22	47.10	27.02
3D (IoU@0.3)	88.57	47.74	64.91	50.46	66.26	46.98
BEV (IoU@0.7)	59.41	7.84	21.58	18.54	22.27	9.84
BEV (IoU@0.5)	82.84	32.91	50.86	38.79	53.56	35.81
BEV (IoU@0.3)	88.87	49.88	66.79	54.72	67.69	50.04
Heading	89.34	54.25	65.71	55.82	70.24	50.97
2D (IoU@0.7)	89.57	61.91	70.02	61.43	74.41	57.66

Comparative experiments under identical network architectures and parameters reveal a notable performance disparity: BEV detection accuracy substantially exceeds that of full 3D detection. This stems from the differences in task complexity - BEV detection eliminates the

challenging height estimation requirement inherent to 3D detection, instead focusing exclusively on 2D spatial coordinate prediction within the ground plane.

C. Ablation experiments

As demonstrated in Table 4, a comparative analysis between GSOP-DLA and baseline models reveals that GSOP-DLA achieves robust performance with 78.34% 3D detection accuracy at IoU@0.5, surpassing DLA34 by 13.49%. In feature fusion design, the enhanced DLAup-SE+, incorporating channel attention mechanisms, shows improvements over DLAup, confirming its effectiveness in 3D feature selection.

Table 4. Ablation Results (%).

	3D(IoU@0.5)	BEV(IoU@0.5)	Heading
DLA34	64.85	69.14	80.15
DLA60	65.22	68.41	77.83
ResNet-50	52.32	58.85	72.76
GSOP-DLA	78.34	81.22	88.91
DLAup	67.43	73.57	85.32
DLAup-SE+	76.63	81.46	89.72
Original settings	15.98	23.31	41.30
+2D loss	60.53	67.22	78.33
+Depth loss(3D)	63.33	72.25	81.28
+Size and offset loss(3D)	68.94	74.36	82.13
+Heading loss(3D)	75.27	79.32	87.76
Original settings	65.94	73.36	82.11
+Training optimization	73.11	80.32	89.76

Systematic validation reveals the critical role of the "2D+3D" joint loss function in monocular ship 3D detection performance. The improved Smooth L1 loss, combined with a dynamic error weighting mechanism, improves the stability of 2D/3D bounding box regression. The dynamic weight function employs nonlinear mapping to amplify gradient contributions from high-error samples, rapidly boosting 3D localization accuracy (IoU@0.5), while gradient smoothing accelerates early-stage convergence. For depth estimation, probabilistic modeling with Laplacian uncertainty loss effectively mitigates monocular scale ambiguity, reducing the relative depth error of distant ships from 36.9% to 12.8%. The weighted smooth loss, integrated with ship-class priors, decreases BEV localization error by 7.14% and size estimation error by 16.3%. Additionally, the hybrid classification-regression framework for heading

prediction, incorporating label smoothing and uncertainty weighting, reduces close-range heading error by 55.6%.

The experiment evaluated multiple training strategies for optimizing the MonoStrong algorithm. Results demonstrate that a progressive warmup strategy—linearly increasing the learning rate from 1e-5 to 1e-4 over the first 10 epochs combined with parameter freezing—effectively mitigates parameter oscillation in deep neural network training. Replacing traditional step decay with adaptive cosine annealing accelerated convergence speed by 23.2% on the training set. These optimizations enabled MonoStrong to achieve 73.11% 3D detection accuracy (IoU@0.5) on the full test set, validating their effectiveness for detection model training.

D. Algorithm Comparison for Ship Image Detection

Figure 6 demonstrates the comprehensive capability of the MonoStrong framework in monocular 3D ship detection across varying operational ranges, showcasing its adaptive performance characteristics under diverse maritime conditions.

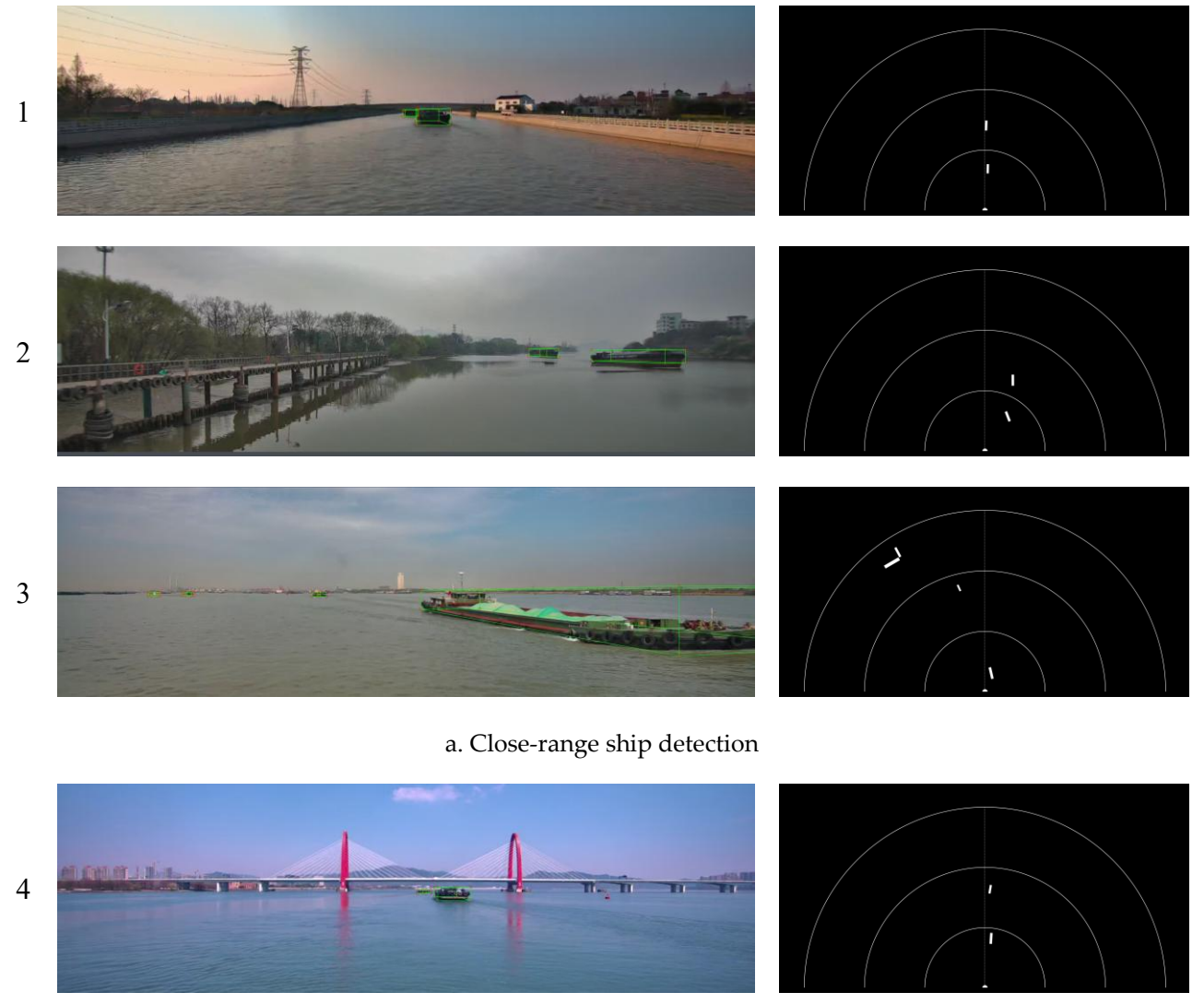
Long-range Detection (500-1000m): For distant targets, MonoStrong achieves an impressive 89% BEV accuracy through its advanced semantic feature extraction mechanism and multi-scale feature fusion architecture. The framework's ability to preserve critical texture details via shallow convolutional layers proves particularly crucial at extended ranges, where other methods often struggle with feature degradation and information loss. This performance demonstrates the model's capacity to maintain discriminative features even when ships appear as small objects with limited pixel representation.

Medium-range Detection (200-500m): Within the intermediate detection range, MonoStrong's cross-scale feature alignment strategy significantly enhances contour recognition capabilities, resulting in a robust 82% 3D localization accuracy at IoU@0.5. This range represents a critical operational zone where precise object boundary delineation becomes essential for accurate attitude estimation and dimensional analysis. The cross-scale alignment mechanism effectively bridges the gap between coarse semantic understanding and fine-grained spatial localization, enabling the model to capture both global ship structure and local geometric details.

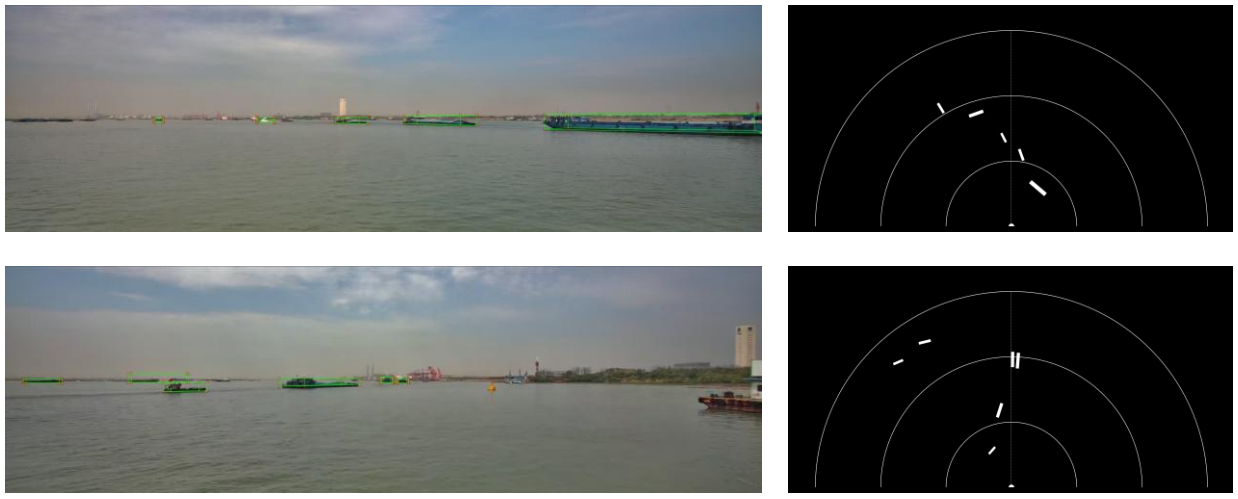
Close-range Detection (0-200m): For near-field targets, MonoStrong maintains satisfied accuracy in both distance measurement and heading prediction, demonstrating its versatility across the entire operational range. At close range, where ships occupy larger portions of the image frame, the model leverages its enhanced feature representation to provide precise

spatial coordinates and orientation estimates, which are crucial for navigation safety and collision avoidance.

Figure 7 demonstrates the detection capabilities across various challenging samples. MonoStrong successfully identifies small-scale targets—those with fewer than 30 longitudinal pixels or beyond 900 meters. Furthermore, MonoStrong maintains its high accuracy even when ships are partially occluded or truncated. It's built-in feature selection mechanism enables robust 3D reconstruction even with partial visibility, avoiding localization errors from missing local features. However, in extreme occlusion cases with severely insufficient pixel-level features, detection failures may still occur, indicating a need for further optimization.



a. Close-range ship detection



b. Medium-range ship detection

c. Long-range ship detection

Figure 6. Ship 3D detection under multi-source images. White semicircle marks a detection range of 300 meters.

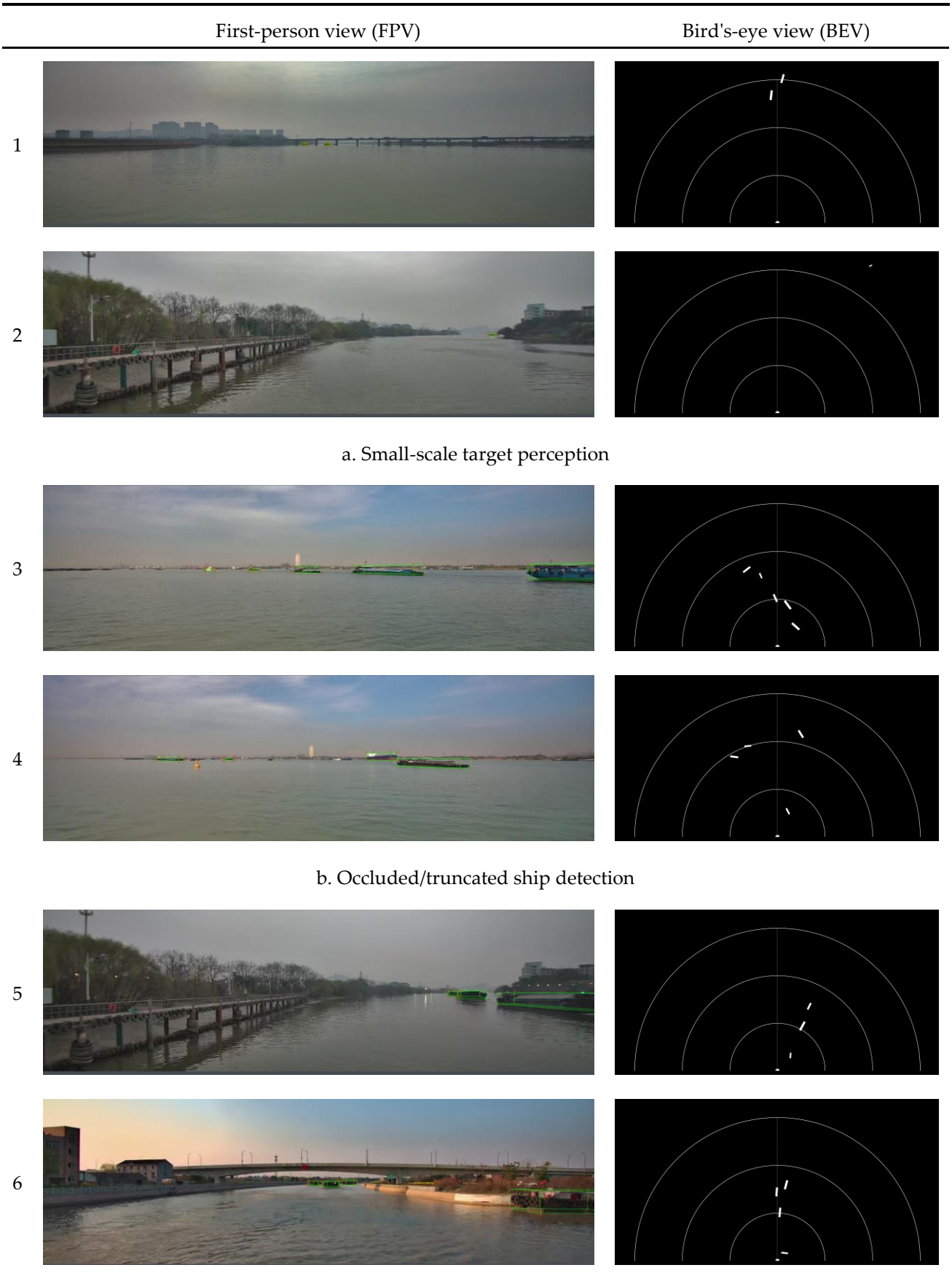


Figure 7. Ship detection under difficult samples

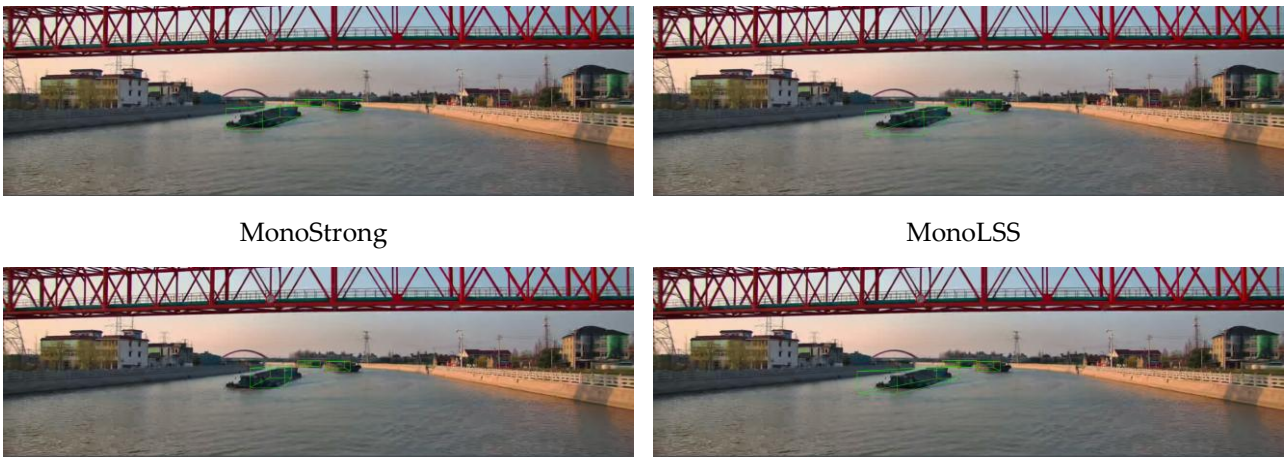
Figures 8 and 9 present the ship detection of MonoStrong, MonoLSS, MonoCon, and MonoDLE algorithms under small-scale ships, crossing ships, and dense ship clusters. In small-scale ship detecting, MonoStrong delivers more accurate results than alternative algorithms. MonoDLE and MonoCon exhibit varying degrees of missed detections or false positives, along with limited positioning precision. While MonoLSS detects targets, its confidence scores and localization accuracy remain inferior to MonoStrong. Thus, it conforms MonoStrong's ability to focus on key features of small-scale ships, enhancing long-range perception.

For crossing ship scenarios, due to large heading variations and high pixel-level similarity between ships, some targets are difficult to identify accurately. However, MonoStrong effectively suppresses feature interference caused by pixel similarity, maintaining high prediction precision. In dense ship environments, all methods suffer from missed detections when identifying distant small targets, yet MonoStrong maintains more stable performance. Overall, MonoStrong demonstrates clear advantages in ship detection under challenging conditions.

MonoStrong MonoLSS

MonoCon MonoDLE

a. Small-scale targets



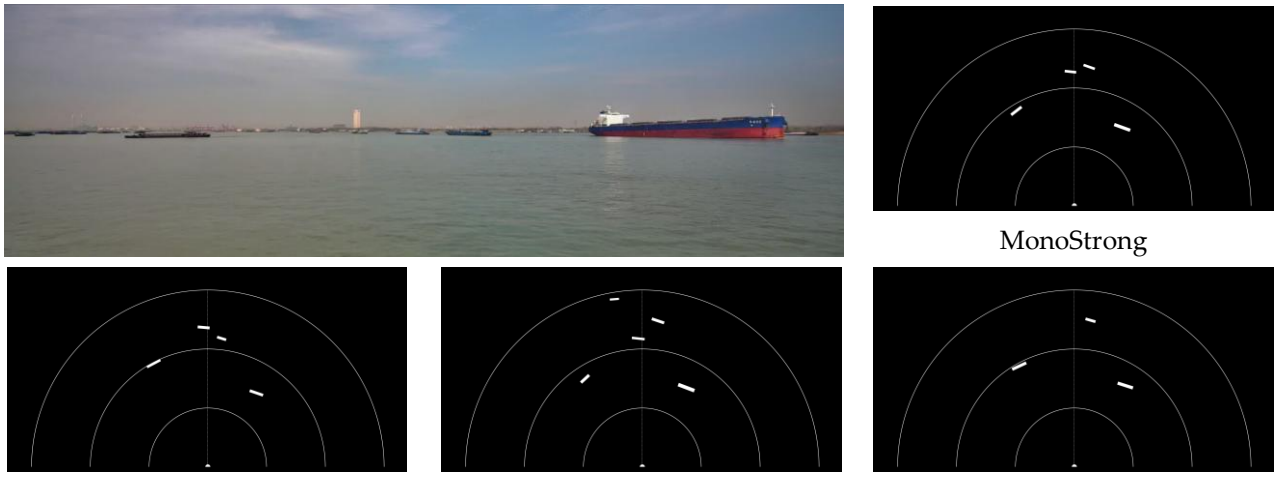
MonoStrong MonoLSS

MonoCon MonoDLE

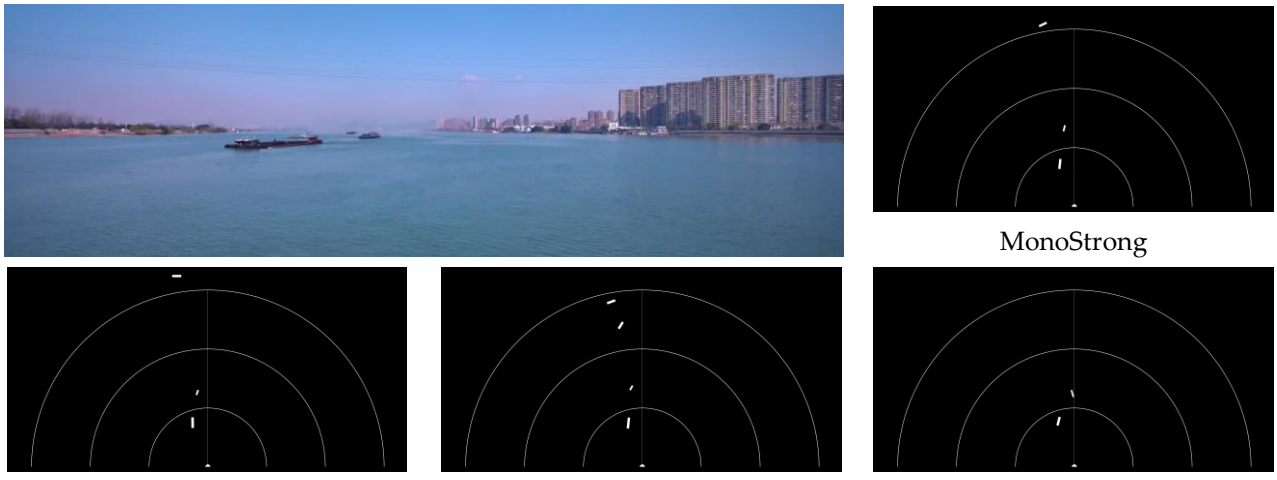
b. Navigational crossing

c. Dense ship traffic

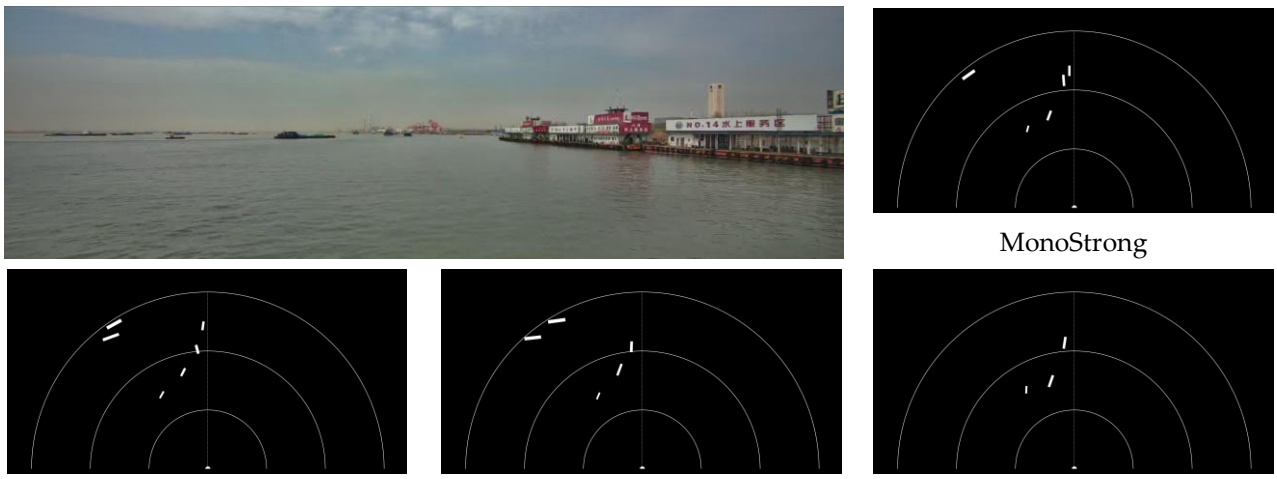
Figure 8. Performance Comparison of Ship Detection - FPV



a. Small-scale targets



b. Navigational crossing



c. Dense ship traffic

Figure 9. Performance comparison of ship detection - BEV

E. Validation

A validation is implemented to test the adaptability of the proposed method in general tasks. Figure 10 demonstrates MonoStrong's ability in object detecting based on sample images from the standard KITTI dataset. In open road scenes (Fig. 10a), MonoStrong achieves high depth estimation accuracy for distant vehicles. Experiments show that MonoStrong delivers reliable 3D bounding box detection when the target's minimum bounding box width exceeds 25 pixels, effectively mitigating scale ambiguity-induced localization drift in monocular vision. For closely parked vehicles (Fig. 10b), MonoStrong precisely separates 3D bounding boxes in overlapping regions. It maintains strong robustness against partial occlusion without false detections, ensuring stable performance. Under extreme lighting conditions (Fig. 10c), the algorithm exhibits low positional error variance, demonstrating effective suppression of overexposure noise in backlit scenarios.

a. Open road scenario

b. Dense parking scenario

c. Adverse lighting scenario

Figure 10. Adaptability validation of MonoStrong

5. Conclusions

The paper proposes a monocular 3D object detection algorithm, MonoStrong, tested in inland waterway scenarios, achieving high-precision perception of ship spatial information. A key innovation is the custom-designed target 3D feature extraction network, enhanced with channel attention mechanisms that adaptively strengthen geometric feature representation. It addresses scale sensitivity caused by large variations in ship aspect ratios and vertical scale compression. Additionally, a multi-scale feature fusion network is proposed, which integrates cross-level convolutional features and attention mechanisms, to enhance detection robustness under occlusion and truncation. The method further improves the association between 2D and 3D bounding boxes through new optimized loss functions, reducing uncertainties in distance and heading estimation. The proposed method leads to more stable heading predictions for nearby ships and improved localization accuracy for distant targets affected by sparse pixel features. Future work could be conducted to further enrich the

KITTI3D-Ship dataset by integrating real and synthetic ship samples under extreme conditions, covering more complex viewpoints and scenarios.

Acknowledgements

The authors gratefully acknowledge support from the National Key Research and Development Program of China under grant (2023YFB4302303) and Henan Technology Innovation Fund of China (241111243000). This research was also funded by a European Research Council project under the European Union's Horizon 2020 research and innovation program (TRUST CoG 2019 864724).

References

Bläser, N., Magnussen, B.B., Fuentes, G., Lu, H., Reinhardt, L., 2024. MATNEC: AIS data-driven environment-adaptive maritime traffic network construction for realistic route generation. Transportation Research Part C: Emerging Technologies 169, 104853. https://doi.org/10.1016/j.trc.2024.104853

Bochkovskiy, A., Wang, C., Liao, H., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Anush, K., Yu P., Giancarlo B., Beijbom, O., 2020. nuScenes: A multimodal dataset for autonomous driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11618-11628. https://doi.org/10.1109/CVPR42600.2020.01164

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End to End object detection with transformers. In: 16th European Conference on Computer Vision (ECCV). Presented at the 16th European Conference on Computer Vision (ECCV), pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, Y., Huang, S., Liu, S., Yu, B., Jia, J., 2022. DSGN++: Exploiting visual-spatial relation for stereo-based 3d detectors. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(4), 1–14. https://doi.org/10.1109/tpami.2022.3197236

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q., 2019. CenterNet: Keypoint triplets for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6568-6577. https://doi.org/10.1109/ICCV.2019.00667

Fan, L., Wang, F., Wang, N., Zhang, Z., 2024. FSD V2: Improving fully sparse 3d object detection with virtual voxels. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(2), 1279–1292. https://doi.org/10.1109/tpami.2024.3502456

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645. https://doi.org/10.1109/tpami.2009.167

G Girshick, R., 2015. Fast R CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). Presented at the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, https://doi.org/10.1109/ICCV.2015.169

Gao, Z., Xie, J., Wang, Q., Li, P., 2019. Global second-order pooling convolutional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3019-3028. https://doi.org/10.1109/CVPR.2019.00314

Ge Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354-3361, https://doi.org/10.1109/CVPR.2012.6248074

He, J., Chen, Y., Wang, N., Zhang, Z., 2023. 3D video object detection with learnable object-centric global optimization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5106-5115. https://doi.org/10.1109/CVPR52729.2023.00494

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask R CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., 2019. Searching for mobilenetv3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314-1324. https://doi.org/10.1109/ICCV.2019.00140

Kelvin, L., Che, Y., Hui, Q., Zhuan, J., Yunli, L., 2010. Human detection using histogram of oriented gradients and human body ratio estimation. In: 2010 3rd International Conference on Computer Science and Information Technology. Presented at the 2010 3rd International Conference on Computer Science and Information Technology, pp. 18-22. https://doi.org/10.1109/ICCSIT.2010.5564984

Li, H., Siu, J., Yang, Z., Liu, J., Liu R.W., Liang, M., Li, Y., 2022. Unsupervised hierarchical methodology of maritime traffic pattern extraction for knowledge discovery. Transportation Research Part C: Emerging Technologies 143, 103856–103856. https://doi.org/10.1016/j.trc.2022.103856

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z., 2023. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. In: 37th AAAI Conference on Artificial Intelligence. Presented at the 37th AAAI Conference on Artificial Intelligence 37(2), pp. 1477-1485. https://doi.org/10.1609/aaai.v37i2.25233

Li, Z., Jia, J., Shi, Y., 2024. MonoLSS: Learnable sample selection for monocular 3d detection. In: 2024 International Conference on 3D Vision (3DV). Presented at the 2024 International Conference on 3D Vision (3DV), pp. 1125-1135. https://doi.org/10.1109/3DV62453.2024.00088

Liu, L., Zhang, M., Liu, C., Yan, R., Lang, X., Wang, H., 2025. Shipping map: An innovative method in grid generation of global maritime network for automatic vessel route planning using AIS data. Transportation Research Part C: Emerging Technologies 171, 105015–105015. https://doi.org/10.1016/j.trc.2025.105015

Liu, X., Xue, N., Wu, T., 2022. Learning auxiliary monocular contexts helps monocular 3d object detection. In: 36th AAAI Conference on Artificial Intelligence. Presented at the 36th AAAI Conference on Artificial Intelligence 36(2), pp. 1810-1818. https://doi.org/10.1609/aaai.v36i2.20074

Liu, Z., Wu, Z., Tóth, R., 2020. SMOKE: Single-stage monocular 3d object detection via keypoint estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4289-4298. https://doi.org/10.1109/CVPRW50498.2020.00506

Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., Shen, H., Ren, J., Han, S., Ding, E., Wen, S., 2020. PP-YOLO: An effective and efficient implementation of object detector. arXiv preprint arXiv:2007.12099. https://arxiv.org/abs/2007.12099

Ma, F., Kang, Z., Chen, C., Sun, J., Xu, X.B., Wang, J., 2024. Identifying ships from radar blips like humans using a customized neural network. IEEE Transactions on Intelligent Transportation Systems 25(7), 7187–7205. https://doi.org/10.1109/tits.2023.3347761

Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., 2021. Delving into localization errors for monocular 3d object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4719-4728. https://doi.org/10.1109/CVPR46437.2021.00469

Meng, J.E., Zhang, Y., Chen, J., Gao, W., 2023. Ship detection with deep learning: A survey. Artificial Intelligence Review 56(10), 11825–11865. https://doi.org/10.1007/s10462-023-10455-x

Meng, H., Li, C., Chen, G., Chen, L., Knoll, A., 2024. Efficient 3d object detection based on pseudo-lidar representation. IEEE Transactions on Intelligent Vehicles 9(1), 1953–1964. https://doi.org/10.1109/tiv.2023.3319985

Misra, D., 2019. Mish: A self-regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681. https://doi.org/10.48550/arxiv.1908.08681

Philion, J., Fidler, S., 2020. Lift, Splat, Shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: 16th European Conference on Computer Vision (ECCV). Presented at the 16th European Conference on Computer Vision (ECCV), pp. 194-210. https://doi.org/10.1007/978-3-030-58568-6_12

Pon, A.D., Ku, J., Li, C., Waslander, S.L., 2020. Object-centric stereo matching for 3d object detection. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). Presented at the 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 8383-8389, https://doi.org/10.1109/ICRA40945.2020.9196660

Qian, R., Garg, D., Wang, Y., You, Y., Belongie, S., Hariharan, B., Mark, C., Kilian, Q.W., Chao, W.L., 2020. End-to-end pseudo-lidar for image-based 3d object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5880-5889. https://doi.org/10.1109/CVPR42600.2020.00592

Qin, Z., Wang, J., Lu, Y., 2021. MonoGRNet: A general framework for monocular 3d object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(9), 5170-5184. https://doi.org/10.1109/tpami.2021.3074363

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525. https://doi.org/10.1109/CVPR.2017.690

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767

Sigirci, I.O., Bilgin, G., 2022. Spectral-spatial classification of hyperspectral images using bert-based methods with hyperslic segment embeddings. IEEE Access 10, 79152–79164. https://doi.org/10.1109/access.2022.3194650

Sohan, M., Sai Ram, T., Rami Reddy, C.V., 2024. A Review on YOLOv8 and Its Advancements. In: 2023 International Conference on Data Intelligence and Cognitive Informatics (ICDICI). Presented at the 2023 International Conference on Data Intelligence and Cognitive Informatics (ICDICI), pp. 529–545. https://doi.org/10.1007/978-981-99-7962-2_39

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D., 2020. Scalability in perception for autonomous driving: Waymo open dataset. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2443-2451. https://doi.org/10.1109/CVPR42600.2020.00252

Tao, R., Han, W., Qiu, Z., Xu, C. Z., Shen, J., 2023. Weakly supervised monocular 3d object detection using multi-view projection and direction consistency. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17482-17492. https://doi.org/10.1109/CVPR52729.2023.01677

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Wang, T., Zhu, X., Pang, J., Lin, D., 2021. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Presented at the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 913-922, https://doi.org/10.1109/ICCVW54120.2021.00107

Xin, X., Liu, K., Yu, Y., Yang, Z., 2025. Developing robust traffic navigation scenarios for autonomous ship testing: an integrated approach to scenario extraction, characterization, and sampling in complex waters. Transportation Research Part C: Emerging Technologies 178, 105246–105246. https://doi.org/10.1016/j.trc.2025.105246

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987-5995. https://doi.org/10.1109/CVPR.2017.634

Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., Zhang, X., Chen, P., 2023. BEVHeight: A robust framework for vision-based roadside 3d object detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21611-21620. https://doi.org/10.1109/CVPR52729.2023.02070

Yu, Y., Yang, X., Li, Q., Zhou, Y., Zhang, G., Da, F., Yan, J., 2023. H2RBox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.04403

Zhang, Y., Lu, J., Zhou, J., 2021. Objects are different: Flexible monocular 3d object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3288-3297. https://doi.org/10.1109/CVPR46437.2021.00330

Zhang, R., Qiu, H., Wang, T., Guo, Z., Cui, Z., Qiao, Y., 2023. MonoDETR: Depth-guided Transformer for Monocular 3D Object Detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9121-9132. https://doi.org/10.1109/ICCV51070.2023.00840

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J., 2024. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965-16974. https://doi.org/10.1109/CVPR52733.2024.01605

Zhou, Y., Zhu, H., Liu, Q., Chang, S., Guo, M., 2023. MonoATT: Online monocular 3d object detection with adaptive token transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17493-17503. https://doi.org/10.1109/CVPR52729.2023.01678

Zou, J., Tian, K., Zhu, Z., Ye, Y., Wang, X., 2024. DiffBEV: Conditional diffusion model for bird's eye view perception. In: 38th AAAI Conference on Artificial Intelligence. Presented at the 38th AAAI Conference on Artificial Intelligence 38(7), pp. 7846-7854. https://doi.org/10.1609/aaai.v38i7.28620