Waterway-BEV: Generate Bird's Eye View Layouts of a Waterway From a First-Person View Camera Using Cross-View Transformers

Feng Ma[1, 3], Xin Jiang[2], Chen Chen[4*], Jie Sun[5], Xin-ping Yan[1, 3], Jin Wang[6]

[1]State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan, China

[2]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China

[3]Intelligent Transportation System Center, Wuhan University of Technology, Wuhan, China

[4]School of Computer Science and Technology, Wuhan Institute of Technology, P. R. China

[5]Nanjing Smart Water Transportation Technology Co., Ltd, Nanjing, China

[6]Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, UK

Abstract: In the domain of autonomous ship navigation, the construction of bird's-eye view (BEV) layouts for waterways has obvious significance. A helmsman can generate the BEV layout of the waterway using his/ her eyes only. To simulate this intelligence, a novel neural network-based algorithm named SECross is proposed, which enables reconstructing a local map formed by the waterway layout and ship occupancies in the bird's-eye view given a first person view monocular image only. SECross employs an efficient SEResNeXt encoder to extract features from first person view (FPV) monocular images, capturing deep semantic information related to waterways and ships. Due to the variations in information across different perspectives, SECross incorporates a Cross-View Transformation Module, which takes the constraint of cycle consistency between views into account and makes full use of their correlation to strengthen the view transformation and scene understanding. To fully leverage the feature output of the SEResNeXt encoder, SECross employs a decoder based on a dedicated lightweight network. This decoder is responsible for decoding the enhanced bird's-eye view (BEV) feature maps and generating the BEV layout. By employing the Focal Loss as the loss function for model optimization, SECross takes into account the quantity and classification difficulty of ship samples during the training process, thereby improving the generation performance and convergence speed. The experiments demonstrated that SECross achieved notable performance metrics, with mIOU and mAP rates reaching 97.8% and 98.2%, respectively, in waterway bird's-eye view layout generation. SECross outperformed other state-of-the-art (SOTA) algorithms in generating BEV layouts of waterways. In particular, during specialized scenarios such as crossroads of waterways and tasks involving small target ships, SECross consistently generated satisfactory bird's-eye view layouts, demonstrating robustness and applicability.

Keywords: bird's-eye view, semantic segmentation, cross-view transformer, loss function

1.INTRODUCTION

In the realm of autonomous ship navigation, the perception of the surrounding environment is deemed a pivotal task. It carries significant importance in ensuring navigational safety, mitigating ship accident rates, and devising optimal routes for ship navigation. The Bird's Eye View (BEV) layout finds extensive application in diverse fields within the domain of autonomous ship navigation, including waterway recognition, navigation planning, obstacle detection, as well as data collection and analysis. This layout facilitates a panoramic view of the ship's frontal 180 degrees from a top-down perspective, particularly beneficial in complex environments. Consequently, it covers an extensive imaging area and ensures high accuracy in capturing the surroundings. Compared to First Person View (FPV), the Bird's Eye View (BEV) layout can intuitively present spatial relationships between our ship, other ships, and waterways. Therefore, handling BEV data is more straightforward than FPV, resulting in significantly improved efficiency in acquiring maritime navigation information.

In general, the image information conveyed by the BEV layout primarily consists of two-dimensional image data comprising a series of points. These points represent the positions of other ships and the waterways. Consequently, the process of generating the BEV layout primarily relies on obtaining the relative positional information of other ships and the waterways. While many methods traditionally rely on precise devices such as LiDAR or GNSS devices to accurately obtain the positional information of other ships, a recently proposed approach in the field of autonomous driving is solely based on deep learning. These novel approaches enable the generation of BEV layouts exclusively from a monocular FPV color image, eliminating the need for additional devices.

Traditional BEV layout generation methods often require the integration of multiple cameras and sensors, including LiDAR [1]. These methods exhibit high accuracy in scenarios with relaxed real-time requirements. However, the real-time generation of precise BEV layouts becomes challenging when dealing with fast-moving objects, primarily due to the substantial computational resources needed. In recent years, deep learning technologies have rapidly advanced in areas such as object detection and instance segmentation, especially with the widespread application and continuous refinement of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) across various tasks. In complex scenarios, such as 3D object detection, traffic sign detection, and obstacle detection, CNN-based object detection methods have shown outstanding performance. Deep learning-based monocular BEV generation methods, compared to

traditional approaches, can achieve more precise target detection and layout generation.

In the navigation environment of inland waterways, monocular BEV layout research using similar methods must face a series of challenges. Firstly, ships encountered in inland waterways generally appear at relatively long distances from our ship, resulting in their representation as very small targets in the FPV images. Secondly, due to the constrained width of inland waterways, occlusion of ships may occur during dense navigation or when ships cross waterways. Lastly, inland waterways typically follow the meandering course of rivers, and ships navigating through these waterways need to traverse channels with significant curvature. Additionally, there may be many crossroads in inland waterways with various geometrical shapes. Therefore, the existing state-of-the-art (SOTA) BEV algorithms might not have satisfactory performances in producing accurate BEV layouts of waterways.

In response to the characteristics of inland waterways, this research proposes a BEV layout generation algorithm based on a customized generative network called SECross. Compared to previous research, the method presented in this work differs significantly in several aspects.

[1] The feature network has been enhanced to extract crucial ship features by employing more efficient convolutional modules.

[2] A transformation module with cross-view attention mechanism is utilized to integrate feature information from different perspectives, further enhancing the features of the output, and enabling the transformation of features from FPV to BEV.

[3] The loss calculation process of the algorithm has been optimized to improve the precision of positioning targets related to ships. Consequently, this improvement contributes to enhancing the algorithm's segmentation effectiveness in scenarios with dense ship presence.

[4] To evaluate the effectiveness of various algorithms for generating BEV layouts in various environmental conditions, a high-quality dataset named MonoWaterwayGen, consisting of 18,000 images with virtual waterways and 15,000 images with real waterways, has been constructed.

The remaining sections of this paper are organized as follows. In Section 2, a brief review of relevant research on waterway environment perception in different scenarios will be presented. Section 3 proposes a waterway BEV layout generation algorithm based on GAN. Section 4 conducts a comparative analysis of various algorithms for BEV layout generation in different scenarios. Finally, Section 5 provides a summary of the main contributions of the proposed method and outlines future developments.

2.REFERERNCES

2.1. Perception Methods in Waterways

Currently, methods for waterway environment perception have been extensively explored and developed in both natural scenes and infrared images. In the early stages, traditional methods primarily relied on techniques such as cluster analysis [2] and maritime radar for ship and waterway recognition. However, these methods exhibited certain limitations, showing inadequate adaptability and possibly facing challenges in achieving satisfactory results across different scenarios. Meanwhile, the rapid development of deep learning and neural network technologies has brought significant progress to tasks such as semantic segmentation, demonstrating outstanding performance in various recognition scenarios. However, in waterway perception, very few BEV related research layout generation methods based on deep learning can be found, since no data set has been published yet.

In recent years, significant advancements have been made in object detection, instance segmentation and semantic segmentation using CNNs. In 2014, the R-CNN algorithm demonstrated a significant advantage on the PASCAL VOC dataset, gradually establishing the dominance of deep learning-based algorithms in the field of object detection. As is widely known that single-stage algorithms represented by the YOLO [3] series have achieved a satisfactory balance between computational speed and detection accuracy, allowing for end-to-end training and widespread adoption in various recognition tasks. Meanwhile, in tasks such as semantic segmentation, two-stage algorithms like Mask R-CNN [4] typically exhibit higher accuracy. The progress of semantic segmentation algorithms is also advancing rapidly, Kirillov *et al.* [5] proposed a semantic segmentation model that attempts to segment anything and has achieved expected performance in most scenarios.

In the research on ship detection and waterway recognition in inland waterway scenarios, several scholars have achieved abundant research results by integration deep learning methods in the realm of object detection. In the field of ship detection, Yang *et al.* [6] proposed a propagation detection and tracking algorithm based on the K-Means clustering algorithm and Soft Non-Maximum Suppression (Soft-NMS), aiming to enhance the algorithm's accuracy and robustness. To address the instance segmentation problem of ships, Zhang *et al.* [7] proposed a comprehensive ship segmentation network based on a SqueezeNet discriminator and a DeepLabv3+ extractor to suppress interference information in images and accurately segment ships. In the task of waterway segmentation, Yin *et al.* [8] introduced a shoreline detection method based on the ResNet backbone

network and Canny edge detection, aiming for accurate segmentation of inland waterways. Furthermore, attention mechanisms and various types of feature pyramid structures have been widely applied in algorithm design [9-11]. In summary, the use of deep neural networks can effectively extract features related to ships and waterways.

However, the existing perception technologies mainly focus on studying information related to ships and waterways in FPV, lacking the acquisition of relevant information for BEV layouts of ships or waterways. Although deep learning-based object detection and instance segmentation techniques have been enhanced for applicability in the maritime domain and have achieved commendable results, they often lack representation of three-dimensional spatial information. As a result, it is difficult for them to directly provide precise position, size, or geometrical information of other ships in a three-dimensional space. On the contrary, an accurate BEV layout of the circumstance can provide a comprehensive perspective, enabling ships to perceive their surroundings more fully, including the relative positional information of other ships and obstacles. In fact, deep learning-based methods, such as S-CNN, can extract shorelines from a ship's FPV images, which can also serve as clues to identify navigable areas. However, since the comprehensive outlook has been discarded after these shoreline extractions, any occlusion on shorelines might result in detection failure, leading to fatal errors in the identification of navigable areas. The information contained in the BEV layout can be directly extracted from FPV images as human does, which is intuitive and comprehensive. The key challenge lies in simulating such intelligence, a task that has been partially accomplished in the context of driverless cars.

2.2. Methods for Generating BEV Layouts

Currently, there is a notable dearth of research on the BEV layout generation for ship navigation. Nevertheless, in recent years, the field of autonomous driving for cars has experienced rapid development. The BEV layout for cars shares similarities with that for ships, as both entail roads, other traffic participants, and obstacles. Hence, we can draw inspiration from technical methods in relevant research within the autonomous driving domain to design algorithms for generating BEV layouts from FPV images for ships.

In the research on generating BEV layouts for automobiles, the mainstream methods have historically involved the use of multiple cameras or LiDAR sensors. The BEV images are generated through projecting the collected images and point cloud data onto the BEV space and then integrating them. Cai *et al.* [12] designed a LiDAR-Guided View Transformer (LGVT) and proposed a Temporal Deformable Alignment (TDA) module, effectively obtaining camera

representations in the BEV space and aggregating BEV features from multiple historical frames. Liu *et al.* [13] proposed an efficient and generic multi-task multi-sensor fusion framework called BEVFusion, which effectively preserves semantic information in the BEV space, enhancing accuracy and reducing computational costs. Liang *et al.* [14] proposed a surprisingly simple yet novel fusion framework, dubbed BEVFusion, whose camera stream does not depend on the input of LiDAR data, ensuring stability in exceptional situations. With recent success of the transformer [15], its ability of explicitly modeling pairwise interactions for elements in a sequence has been leveraged in many vision tasks. Li *et al.* [16] designed a spatial cross-attention mechanism and a temporal self-attention mechanism to aggregate spatial information and fuse historical BEV information, achieving comparable performance without relying on LiDAR.

As technology advances, research on generating BEV layouts solely based on monocular FPV images without relying on LiDAR is becoming increasingly popular. While multi-camera setups and LiDAR systems can provide richer depth information, aiding in more accurate capture of the positions and distances of objects at different ranges, monocular cameras offer the advantages of lower cost, simpler deployment, and higher real-time performance. Therefore, research on generating BEV layouts based on monocular FPV images is of significant importance. Lu *et al.* [17] proposed a Variational Autoencoder (VAE) model that can predict road layouts from given images but cannot infer layouts that are unobstructed by obstacles. Kaustubh *et al.* [18] introduced a unified model called MonoLayout based on GAN to address the tasks of road layout and vehicle distribution estimation from monocular images, and the model employs adversarial feature learning to attempt to complete obscured parts of images. Zhou *et al.* [19] proposed the Cross-view Transformer (CVT) network to perform viewpoint transformation. The network utilizes BEV queries and employs cross-attention to query image features. Additionally, position embeddings calculated from camera parameters are added to the image features to provide better priors. Yang *et al.* [20] proposed a network called Pyva that utilizes a cross attention module to enhance viewpoint transformation and scene understanding by leveraging the correlation between FPV and BEV perspectives.

Based on the above analysis，it can be concluded that the key to generate BEV layouts from FPV images for waterways lies in designing efficient generative networks using deep learning techniques. Furthermore, compared to traditional generation methods, deep learning approaches offer significant advantages. Given the characteristics of large curvatures in inland waterways and the small size of ships, this research proposes a BEV layout generation algorithm from FPV images for waterways based on a Generative Adversarial Network (GAN) framework, incorporating

convolutional neural networks and cross-attention mechanisms. Through comprehensive integration of the feature information derived from both FPV and BEV, this algorithm can achieve accurate localization and precise segmentation of ships and waterways.

3. A PROPOSED METHOD

3.1. Network Overview

The overall framework of the BEV layout generation model from single monocular FPV images proposed in this research is illustrated in Figure 1. The model comprises several components, with the generator network and the Cross-View Transformation Module being the core ones. The generator network is a type of encoder-decoder architecture, wherein the encoder adopts SEResNeXt as the backbone network. The basic idea is to add Squeeze-and-Excitation (SE) modules to ResNet [21] and introduce grouped convolution [22] to dynamically recalibrate the feature channels in the network, thus completing feature extraction. This network can capture important features effectively without introducing extra parameters, which helps in understanding the semantic information of small-scale ships in FPV images of waterways. The Cross-View Transformation Module employs a convolutional computation method based on the Transformer structure. By connecting the features of FPV and BEV images, it achieves the fusion of features between cross-views to enhance the extracted features. The BEV feature decoder is used to decode BEV features, thereby generating the output results. At the final step, in the loss function calculation part, the Focal Loss function [23] is used to optimize the calculation process. This method improves the imbalance in the quantity and recognition difficulty between ship and waterway samples, allowing the model to learn more generalized feature representations.

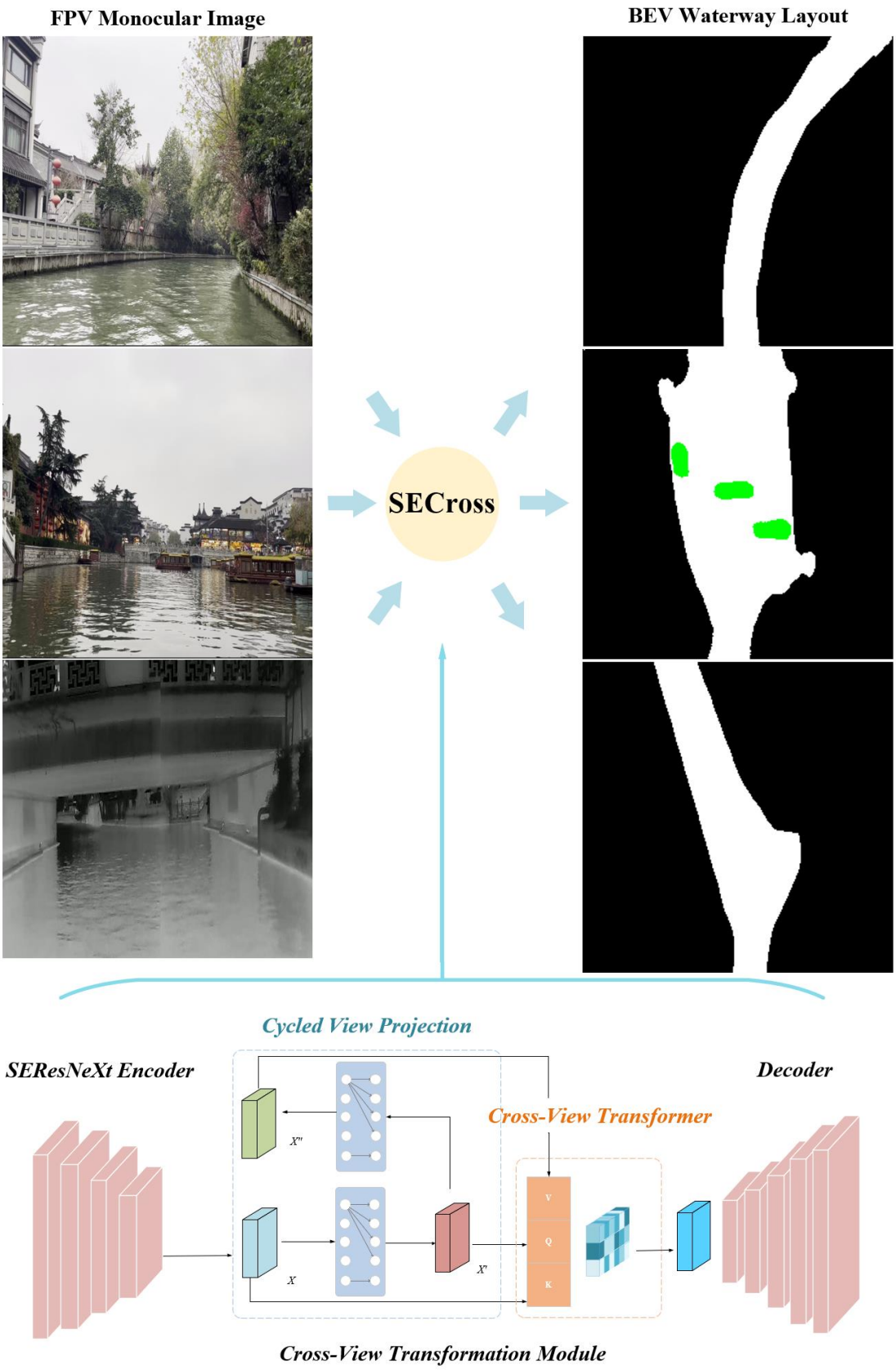


Figure 1. Overall structure of the proposed SECross.

3.2. SEResNeXt Encoder

As mentioned previously, waterway FPV images often exhibit some specific data

characteristics, such as low resolutions of targets, junctions of shorelines and the horizon line. The characteristics result in significant differences between waterway FPV images and traditional natural images. Classic feature networks like ResNet50, ResNet101, VGG16, *etc.*, are typically designed for general-purpose natural image datasets and may not adapt well to the distinctive image characteristics of waterway FPV images, leading to suboptimal feature extraction performance. Additionally, since ships in waterway FPV images tend to have smaller scales, feature networks should possess the ability to perceive small objects. Based on these analyses, the feature extraction network adopted in this research is illustrated in Figure 2. This network is an efficient adaptive neural network that demonstrates satisfactory performances in extracting features related to waterways and ships.

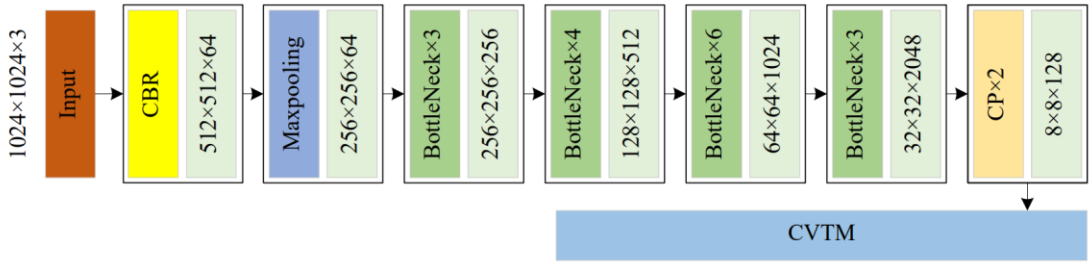


Figure 2. Overall architecture of the encoder network.

3.2.1 The group convolution method based on the inception structure

In FPV images of inland waterways, image sequences often contain ship regions of various sizes and shapes, with a large proportion of ships being relatively small in scale. Hence, it is crucial to capture key ship features with high sensitivity. Additionally, the feature network should possess noise resistance and robustness. To address these requirements, valid convolutional computation methods can be employed to enhance network performance. Furthermore, convolutions with stronger adaptability can enhance network performance and improve recognition accuracy. This research introduces a simple and fast group convolution computation method, resulting in a more efficient feature extraction network design.

With the advancement of CNNs, the depth of networks continues to increase. However, excessively deep layers often contain redundant computations, leading to ineffective increases in model parameters and computational consumption. Moreover, a significant amount of ineffective convolutional computations can impact the extraction of crucial features, especially in ship detection under FPV images where targets are small in scale and feature information is limited. Christian *et al.* [24] proposed an inception convolutional structure that effectively reduces computational costs by utilizing multiple convolution operations of different sizes within a single layer to extract features

at different scales. Therefore, increasing the network width contributes to the network learning rich feature representations without significantly increasing the quantity of network parameters. This research replaces the convolution in the ResNet network's BottleNeck module with group convolution, thereby obtaining the ResNeXt network. This improvement aims to enhance the model's ability to capture critical features of ships and waterways while reducing the occurrence of overfitting.

Figure 3 compares the BottleNeck structures of ResNet and ResNeXt networks. For an input with 256 channels, ResNeXt compresses it into 32 groups of data with 4 channels each using 1×1 convolutions. After convolution operations, it then expands back to 256 channels using 1×1 convolutions. Finally, the 32 groups of data are added together element-wise to form the output with 256 channels. ResNeXt replaces the original three-layer convolutional Bottleneck in ResNet with a parallel stack of the same topological structure Bottlenecks, thereby improving the model's accuracy without significantly increasing the quantity of parameters. Therefore, using the ResNeXt network allows for more accurate and efficient extraction of image features.

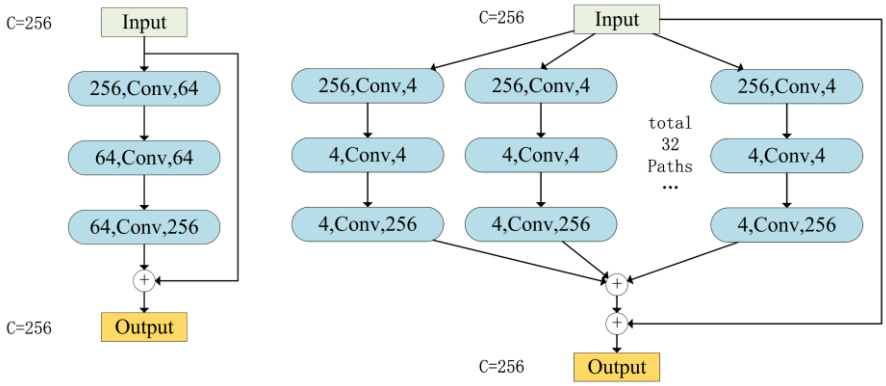


Figure 3. Comparison of Bottleneck structures in ResNet and ResNeXt.

3.2.2 Channel Attention Model Based on SE

For the feature extraction network, it is crucial to fully utilize the contextual information around ships in the FPV images to enhance the accuracy of ship segmentation. To fulfill this requirement, attention mechanisms can be introduced or convolutional structures with global contextual awareness can be utilized. By capturing the correlations between ships and their surrounding environments, the network can more effectively comprehend the shape, contour, and semantic information of ships, thereby enhancing the effectiveness of ship segmentation.

In response to the aforementioned requirements, the SEResNeXt network used in this research incorporates SE modules connected after the residual layers in the ResNeXt network. These modules adjust the weights for each channel, enhancing attention towards important channels, thus

improving the expressive power of the model. The basic structure of the SE module is depicted in Figure 4. The Squeeze operation utilizes global average pooling to extract features for each channel, obtaining the importance coefficient for that channel. In the Excitation operation, the feature dimension is first reduced to $\frac{C}{r}$ through a fully connected layer. Subsequently, after passing through a ReLu activation, it is then up-sampled back to the original dimension through another fully connected layer. Opting for this approach offers more advantages compared to directly employing one fully connected layer. Firstly, it can have more nonlinearity, which can better fit the complex correlations between channels. Secondly, it greatly reduces the quantity of parameters and computational load. Then, the SEResNeXt network obtain the weight for each channel through the Sigmoid operation, and finally apply the normalized weights to the original feature maps through the Scale operation. The introduction of the SE attention mechanism enables the feature network to better capture the complex correlations between channels, thereby enhancing the representation of crucial information about channels and ships and improving the model's recognition performance.

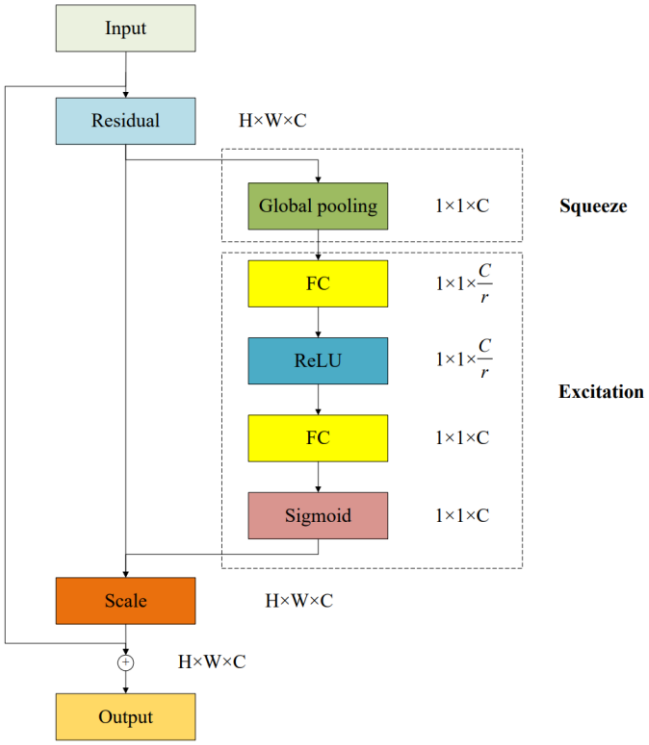


Figure 4. SE module

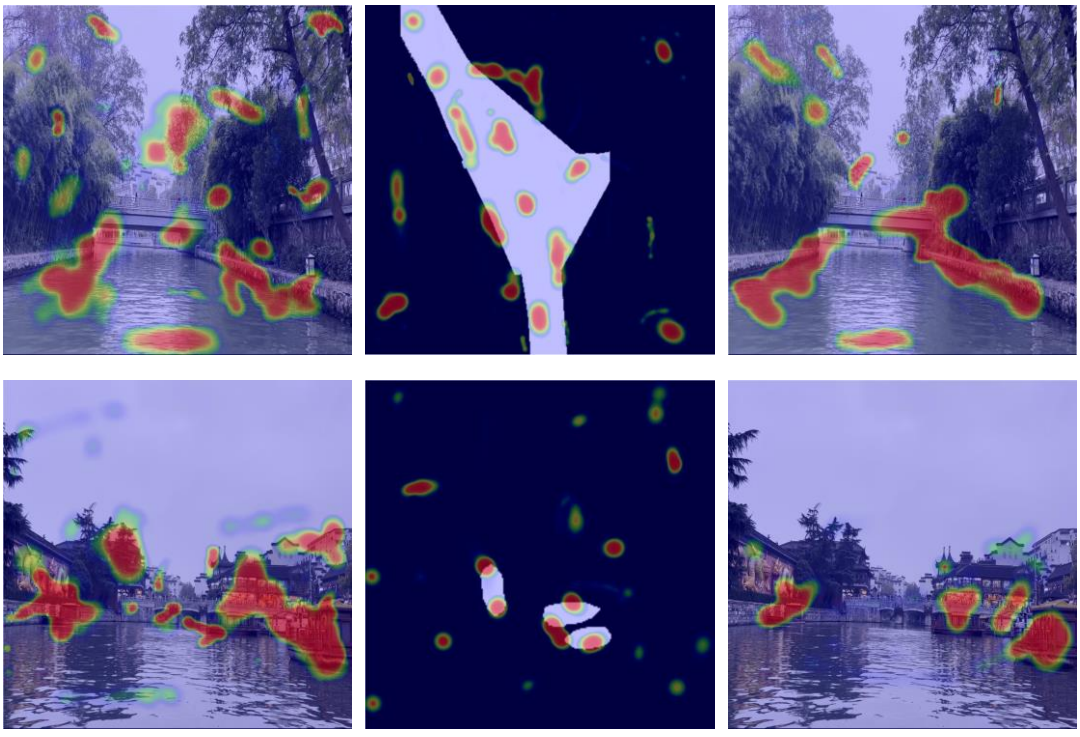
3.3. Cross-View Transformation Module

Due to the significant disparity between FPV and BEV images in waterways, there is a considerable loss of image content during the perspective projection process. Consequently, traditional perspective projection techniques result in flawed outcomes. In order to enhance view

correlation while leveraging the capabilities of deep networks, a Cross-View Transformation Module is introduce into the generator of the GAN framework. This module strengthens the extracted visual features for projecting FPV onto BEV, which consists of two parts: cycled view projection and cross-view transformer. They can be considered as a 'guessing modules' that make reasonable conjectures based on surrounding semantic information.

3.3.1. Cycled View Projection (CVP)

Since the features of FPV are not spatially aligned with the ones of BEV due to their large gap, it is crucial to effectively utilize data from different perspectives for understanding spatial information. Following the practice of [25], the MLP structure is deployed consisting of two fully connected layers to project the features of FPV to BEV, which can overtake the standard information flow of stacking convolution layers. As shown in Figure. 1, X and X' represent the feature maps before and after view projection, respectively. Hence, the holistic view projection can be achieved by: $X' = F_{\mathrm{MLP}}(X)$, where X refers to the features extracted from the SEResNeXt encoder.



X X' X''

Figure 5. Visualization of the features at FPV and BEV

However, such a simple view projection structure cannot guarantee the information of FPV to be effectively delivered. Zhu *et al.* [26] proposed a cycled mapping approach, which involves mapping images from the target domain back to the source domain, thereby enhancing the transfer of information between domains. Therefore, this research introduces a cycled self-supervised

scheme, which involves reprojecting BEV features back to FPV to reinforce view projection. As shown in Figure 1, an MLP is utilized with the same structure to project X' back to X'', i.e., $X'' = F'_{MLP}(X')$. To guarantee the domain consistencies between X' andX'', a cycle loss, i.e., L_{cycle}, is incorporated as expressed below:

$$L_{\text{cycle}} = \|X - X''\|_1 \tag{1}$$

The cycled structure strengthens the connection between FPV and BEV views. X'' retains the most relevant information regarding view projection when the discrepancy between X and X'' cannot be further reduced, since X'' is projected back from X'. Hence, X and X' refer to the features before and after view projection. X'' contains the most relevant features in FPV for view projection.

3.3.2. Cross-View Transformer (CVT)

Compared to road scenarios in the field of autonomous driving, waterway scenarios present more complex traffic conditions. For instance, waterways are generally broader than roads, with ships and boats having irregular heading directions compared to the fixed direction of cars. Additionally, while cars primarily need to be concerned about nearby traffic participants, ships must also consider distant and small ships. These complex conditions not only demand powerful feature extraction networks but also require further enhancement of the extracted features. To tackle this challenge, the research employs an efficient Cross-View Transformer network, leveraging the feature information preserved by the cyclic view projector from different perspectives.

Specifically, the proposed method is based on the attention mechanism of CVT to enhance the key features of waterways and ships in FPV images of waterways. This mechanism fully utilizes the feature information under multiple perspectives and enhances BEV features by learning a cross-attention matrix. In specific implementation, a correlation matrix is obtained by calculating the inner product of the features X and X' before and after projection. Next, the information most related to view projection is obtained from the feature selection part as X''. Finally, this most relevant feature is merged with the projected feature X' to form the final output result.

For waterway images, CVT effectively enhances the features X' by correlating the features before view projection (i.e., X) with the features after view projection (i.e., X'). With the help of the context information of view projection, the network is capable of enhancing the features of the view projection, thus obtaining additional information about the ships and the surrounding environment and background of the waterways. The specific network structure is shown in Figure 6, where CVT consists of two parts: the cross-view association part, which connects FPV and BEV

features to obtain an attention map W to enhance X'; and the feature selection part, which is responsible for extracting the most important information from X''.

Particularly, X, X', and X'' serve as the key $K(K \equiv X)$, the query $Q(Q \equiv X')$, and the value $V(V \equiv X'')$ of CVT. The dimensions of X, X', and X'' are set as the same. X' and X are both flattened into patches, and each patch is denoted as $\mathbf{x}_i' \in X'(i \in [1, \dots, hw])$ and $x_j \in X(j \in [1, \dots, hw])$, where hw refers the width of X times its height. In the cross-view association part, the correlation between each x_i' in X' and each x_j in X is measured by the normalized inner-product:

$$r_{ij} = \langle \frac{x_i'}{\|x_i'\|}, \frac{x_j}{\|x_j\|} \rangle \quad (2)$$

which results in the relevance matrix R。With the relevance matrix R, two vectors $W(W = \{w_i\}, \forall i \in [1, \dots, hw])$ and H $(H = \{h_i\}, \forall i \in [1, \dots, hw])$ are created:

$$w_i = \max_i r_{ij}, \forall r_{ij} \in R \quad (3)$$

$$h_i = arg \max_i r_{ij}, \forall r_{ij} \in R \quad (4)$$

each element of W implies the degree of correlation between each patch of X' and all the patches of X, which can serve as an attention map. Each element of H indicates the index of the most relevant patch in X with regard to each patch of X' .

In the feature selection part, both X and X'' are FPV features, except that X contains complete information while X'' only includes information relevant to view projection. Assuming the correlation between X and X' is similar to the correlation between X'' and X', the relevance between X and X' (i.e., R) can be used to extract the most important information from X''. This method adopts a feature selection scheme FS, which generates a new feature map T, $T(T = \{t_i\}, \forall i \in [1, \dots, hw])$ by retrieving the most relevant features in X'':

$$t_i = FS(X'', h_i), \forall h_i \in H \quad (5)$$

where FS retrieves the feature vector t_i from the h_i-th position of X''.

Hence, T stores the most relevant information of X'' for each patch of X'. It can be reshaped as the same dimension as X' and concatenated with X'. Then, the concatenated features will be weighted by the attention map W and finally aggregated with X' via a residual structure.

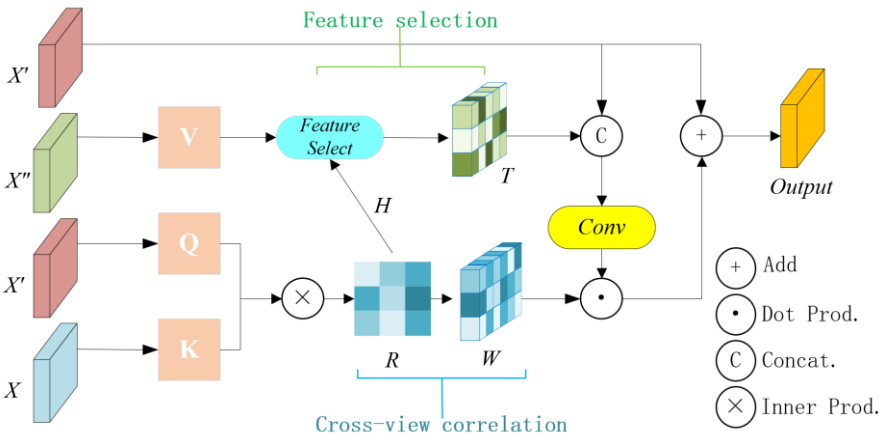


Figure 6. Illustration of Cross-view Transformer

To sum up, the process can be formally expressed as below:

$$Output = X' + Conv\big(Concat(X', T)\big) \odot W \quad (6)$$

where $\odot$ denotes the element-wise multiplication and $Conv$ refers to a convolutional layer with 3×3 kernel size. $Output$ is the final output of CVT and will then be passed to the decoder network to produce the segmentation mask of BEV.

3.4. Feature decoder

In the SEResNeXt encoder structure, the spatial resolution of the FPV images of the waterways is reduced from 1024×1024×3 to 8×8×128 to facilitate the extraction of abstract feature representations. Additionally, the Cross-View Transformation Module only enhances the extracted features without altering the spatial resolution. Therefore, a decoder network is needed to decode the FPV features to obtain the final BEV layout.

Considering the characteristics of the network structure described above, two feature decoders with identical structures were constructed to decode the enhanced features of the waterway and ships separately to generate the BEV layout of the inland waterway. The decoder structure is illustrated in Figure 7. The input is a feature map of size 8×8×128. It undergoes the following operations. The first convolutional layer doubles the number of input channels and increases the resolution twice. Then, it passes through 4 deconvolutional blocks that double the spatial resolution and halve the number of channels each time. Finally, the last convolutional block adjusts the number of channels to 2, resulting in an output feature map size of 256×256×2.

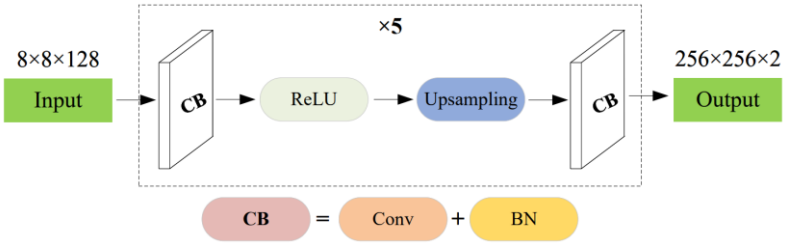


Figure 7. Decoder network structure

The training of the model is defined as minimizing the weighted sum of three loss functions:

$$L = L_G + \alpha L_{cycle} + \beta L_{trans} \quad (10)$$

where L_G represents the generator loss, serving as the primary objective of the network to minimize the gap between the generated layout and the ground truth layout. Since the features enhanced by the Cross-View Transformation Module already possess satisfactory detection performance, there is no need to train a discriminator to normalize the generated layouts. Hence, the discriminator module is removed, reducing the training parameters, and eliminating the need to compute the discriminator loss. Instead, the training of the Cross-View Transformation module is emphasized by adding a cycle loss L_{cycle} to train the cycle view projection module. Additionally, L_{trans} is defined to minimize the gap between the generated layout without enhancement by the Cross-View Transformation Module and the ground truth layout to train the Cross-View Transformation Module. α and β are the balance weights for the cycle loss and the transformation loss, respectively, set to 0.001 and 1.

In the calculation of the network loss function based on the generative adversarial network framework, cross entropy loss is commonly used to compute the losses for both the generator and the discriminator. However, in the task of generating BEV layouts for waterways, the ship targets are typically small, resulting in less distinctive features. Additionally, in real-world scenarios, there is a significant disparity in the number of ship samples compared to waterway samples. To address these challenges, this research adopts the method proposed by Tariq *et al.* [27]: when computing L_G and L_{trans}, Focal Loss is used instead of Cross Entropy Loss. Specifically, the formula for Focal Loss is defined as follows:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma} log(p_t) \quad (11)$$

where the term α represents the category weight factor, used to control the weighting of positive and negative samples. p_t denotes the difficulty level of sample classification, while γ serves as the hyperparameter for Focal Loss. When $\gamma = 0$, it reverts to the standard cross-entropy loss. With $\gamma > 0$, the loss imposes greater weight on poorly predicted labels, directing the model's attention

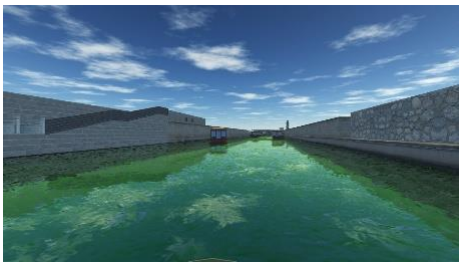
towards challenging samples. Adopting the Focal Loss function highlights incorrectly classified pixels, often associated with rare categories, thereby enhancing the performance of these minority classes. Consequently, in the task of generating BEV layouts for inland waterway navigation, employing the Focal Loss function proves advantageous for the generation of ship BEV layouts.

4. A CASE STUDY

4.1. Dataset

This research had preprocessed data from both virtual and real-world scenarios, creating a high-quality waterway image dataset named MonoWaterwayGen, comprising 33,000 images. As shown in Figure 8, the MonoWaterwayGen dataset encompasses diverse background environments, including narrow waterways, broad waterways, waterways without left or right bank in virtual scenes, and visible light and infrared images of Qinhuai River and Yangtze River in real scenes. Additionally, the dataset also takes into account factors such as the ship heading, and angle variations. Besides, it places a particular emphasis on augmenting the quantity of images depicting ships of small-scale and miniature sizes. Considering the potential for misidentification in scenarios with waterway bifurcations and dense ship presence, MonoWaterwayGen augments the image count in these scenarios to enhance algorithm training and experimental outcomes.

In the MonoWaterwayGen dataset, synchronized FPV and BEV images are captured using Unity3D, drones and ship-borne cameras, preprocessed to meet the dataset format requirements for the proposed SECross algorithm. Moreover, all the images are divided into training, validation, and testing sets in an 8:1:1 ratio. In particular, various SOTA algorithms were trained on the training and validation sets, and evaluation metrics and algorithm performance were assessed on the testing set.



A narrow waterway

A waterway without left bank

A waterway without right bank a broad waterway

(a) Examples of virtual waterway images (Unity3D)

Visible light waterway images of the Qinhuai River

Visible light waterway images of the Yangtze River

(b) Examples of visible light waterway images

Infrared waterway images of the Qinhuai River

Infrared waterway images of the Yangtze River

(c) Examples of infrared waterway images

Figure 8. MonoWaterwayGen Dataset

4.2. Experimental environment and training optimization methods

The experimental setup relied on the Ubuntu 20.04 and utilized the NVIDIA A40 graphics card with an effective memory size of 48GB. The CUDA version used was 11.3.0, PyTorch version was 1.10.0, and the Python environment was 3.7. The performances of different SOTA algorithms in the BEV layout generation were compared, including Monolayout, Pyva and SECross. All algorithms

were evaluated using the same dataset of MonoWaterwayGen. During algorithm training, the input image size was set to 1024×1024 pixels, the BEV layout image size was set to 256×256 pixels, the batch size was set to 16, and the training was conducted for 300 epochs with an initial learning rate of 10^{-4}.

This research optimized the training process of the proposed algorithm by improving aspects such as learning rate decay, loss calculation, and data augmentation. Normally, traditional neural network-based algorithms usually employ a series of fixed learning rates for training, which may result in significant learning rate decay at different training stages, leading to unstable changes in model momentum and negatively affecting the algorithm's training effectiveness. To address this issue, the Adam optimizer was introduced to optimize the learning rate, which adaptively adjusts the learning rate without the need for manual tuning, based on the magnitude of parameter gradients. Furthermore, data augmentation methods, including random mirror flipping and color augmentation, were employed to increase the diversity of training data and improve the model's robustness and generalization abilities.

4.3. Comparisons and Discussions

The metrics of mean Intersection Over Union (mIOU) and mean Average Precision (mAP) were selected to measure the performance of BEV generation in waterway scenarios. Firstly, comparative experiments were designed to assess the actual performance of various algorithms across different evaluation metrics, thereby validating the effectiveness of the SECross. Moreover, ablation experiments were conducted to analyze the individual improvements in SECross and verify the specific effects of different methods. Finally, the adaptability of the SECross in waterway images was examined through the BEV generation of different waterway scenarios.

4.3.1. Experimental analysis of different algorithms

On the constructed MonoWaterwayGen dataset, a comparison was conducted between several commonly used standard algorithms and the proposed SECross. All the algorithms were trained using the same hyperparameters and tested on the same dataset. Moreover, each algorithm underwent cross-testing, and the mIOU and mAP values were averaged over three experimental trials. As shown in Table 1, for the BEV layout generation on inland waterways, our proposed method achieved high mIOU and mAP on the testing images. In comparison to other algorithms, our method showcased promising experimental outcomes, credited to SECross' enhanced recognition capability for small-sized ships, densely navigated ships, and intricate sharp bends and

high-curvature branches within inland waterway images.

Based on the analysis of the virtual waterway BEV generation, it is evident that in scenarios involving narrow inland waterways, SECross only shows a modest improvement of up to 3% compared to other algorithms. The reason lies in the fact that such narrow inland waterways are quite like roads, where the waterways are narrow, relatively smooth, and straight. Additionally, the ships' visual sizes are large and uniform. Since MonoLayout and Pyva were designed for roads, the performance gap of SECross compared to these algorithms is modest and foreseeable. However, the data on waterway and ship detection from scenarios with waterways without left or right banks shows that SECross improves by around 10% compared to MonoLayout. The analysis suggests that the lackluster performance of MonoLayout in recognizing complex scenarios with narrow waterways and small ships stems from its failure to enhance extracted waterway features before decoding to generate the BEV layout. Additionally, SECross outperforms the Pyva algorithm by approximately 3% in ship detection. Moreover, in tasks involving ship detection in virtual wide waterway scenarios, it leads by around 5%. This is attributed to the presence of more small-sized ships in wide waterway scenarios. The SEResNeXt encoder utilized by SECross exhibits stronger capability in extracting features from small-sized targets, thus providing an advantage in the detection of small-sized ships.

To validate the performances of SECross in generating BEV layouts of waterways in real world, a comparative experiment was conducted using visible light and infrared images of the Qinhuai River and the Yangtze River. The results indicate that, regardless of whether it is the Qinhuai River or the Yangtze River, the BEV generation effect of the three algorithms on the infrared images is better than that on the visible light images. This is because the shoreline features of the waterways are more prominent in the infrared images, which is advantageous for the algorithms to segment the waterways. Furthermore, despite the relatively small quantity of ships in the real scene of the Qinhuai River, SECross still has a 1% advantage over other algorithms, which suggests that the network structure of SECross also outperforms other algorithms in generating BEV layouts of waterways in real scenes.

Table 1. Specific experimental results

Scenarios	Algorithms	Waterway		Ship	
		mIOU（%）	mAP（%）	mIOU（%）	mAP（%）
Virtual narrow waterway	MonoLayout	95.39	96.59	82.54	88.79
	Pyva	97.72	98.26	83.77	91.20
	Ours	98.03	98.78	84.35	95.17

Virtual waterway without left bank	MonoLayout	85.83	89.36	59.09	75.59
	Pyva	95.92	96.65	74.25	85.63
	Ours	97.89	98.26	77.42	88.74
Virtual waterway without right bank	MonoLayout	83.43	88.95	58.85	67.30
	Pyva	97.55	98.90	67.00	78.73
	Ours	99.56	99.79	70.95	82.46
Virtual narrow waterway	MonoLayout	-	-	54.58	65.31
	Pyva	-	-	65.77	79.20
	Ours	-	-	71.35	84.53
Visible light waterway of the Qinhuai River	MonoLayout	64.01	72.98	88.71	90.19
	Pyva	71.07	78.96	88.78	90.55
	Ours	74.01	83.25	89.39	91.97
Infrared waterway of the Qinhuai River	MonoLayout	68.87	77.23	88.17	88.58
	Pyva	76.60	83.45	89.26	91.22
	Ours	77.31	85.62	89.39	92.17
Visible light waterway of the Yangtze River	MonoLayout	73.23	84.95	-	-
	Pyva	90.84	91.18	-	-
	Ours	95.29	96.85	-	-
Infrared waterway of the Yangtze River	MonoLayout	75.85	87.51	-	-
	Pyva	91.57	93.28	-	-
	Ours	96.06	97.02	-	-

4.3.2. Ablation Experiments

To further validate the practical performance of each improvement method in SECross, a comprehensive decomposition analysis was conducted based on the MonoWaterwayGen dataset to analyze their impact on BEV layout generation. The main experimental process involved step-by-step application of various improvement methods on MonoLayout, followed by testing their respective performance metrics. The specific results of the ablation experiments for SECross are presented in Table 2.

(1) Analysis of the SEResNeXt network. By replacing the encoder network with SEResNeXt, experimental results reveal that the improved feature extraction network increased the accuracy of waterway recognition by 2.1% and ship detection by 3.3%. Additionally, there was an improvement in the intersection over union metric. This indicates that the algorithm's capability in extracting features of small-scale ships has been enhanced, reducing misidentifications and decreasing the probability of missing targets.

(2) Analysis of Cross-View Transformation Module. In this research, a Cross-View

Transformation Module was added between the SEResNeXt encoder and the feature decoder, allowing the algorithm to consider feature differences between different perspectives and focus more on features most relevant to the perspective projection. This enables the model to perform better when facing challenges such as intersections and curves in complex scenes. Based on the experimental results, the Cross-View Transformation Module strengthens the correlation between FPV and BEV features to enhance significant waterway features, significantly improving the evaluation metrics of inland waterway BEV layout generation algorithms.

(3) Analysis of the Focal Loss function. Experimental results show that the Focal Loss function improved the accuracy of waterway BEV layout generation by approximately 1%. The analysis indicates that a loss function that takes into account sample quantity and classification difficulty yields relatively better results in BEV layout generation tasks under inland waterway images. Furthermore, through multiple experiments, it has been observed that the Focal Loss function could accelerate the convergence speed of the algorithm, leading to relatively rapid and stable convergence of the loss values based on the training and validation sets.

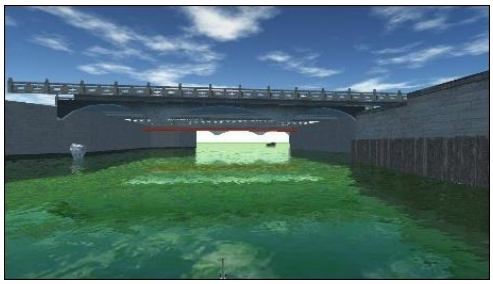
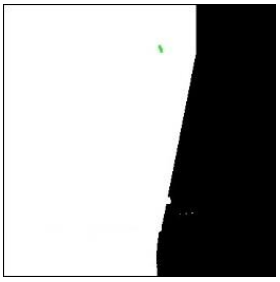
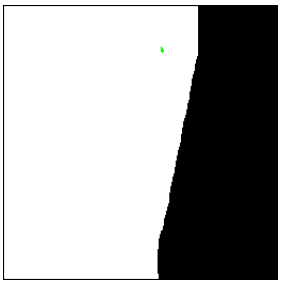
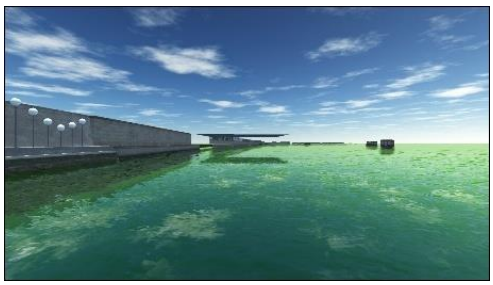
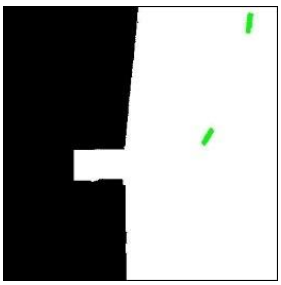
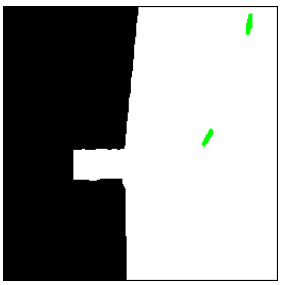
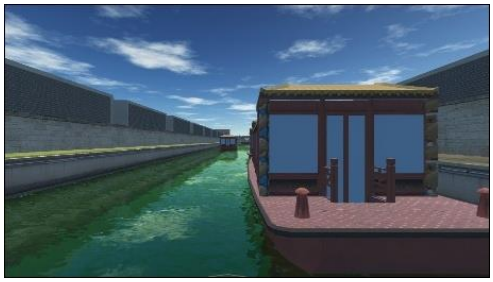
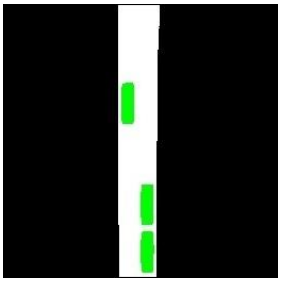
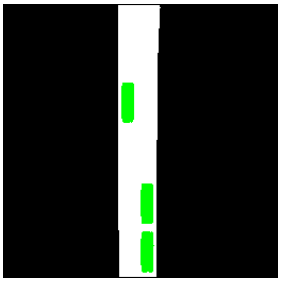
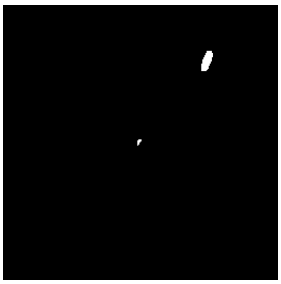
Table 2. Ablation experiments of SECross

Methods		Model 1	Model 2	Model 3	Model 4
MonoLayout		★	★	★	★
SEResNeXt			★	★	★
CVTM				★	★
Focal Loss					★
Waterway	mIOU	0.858	0.871	0.969	0.978
	mAP	0.893	0.914	0.977	0.982
Ship	mIOU	0.590	0.625	0.762	0.774
	mAP	0.755	0.788	0.876	0.887

4.3.3. Comparisons in different scenarios

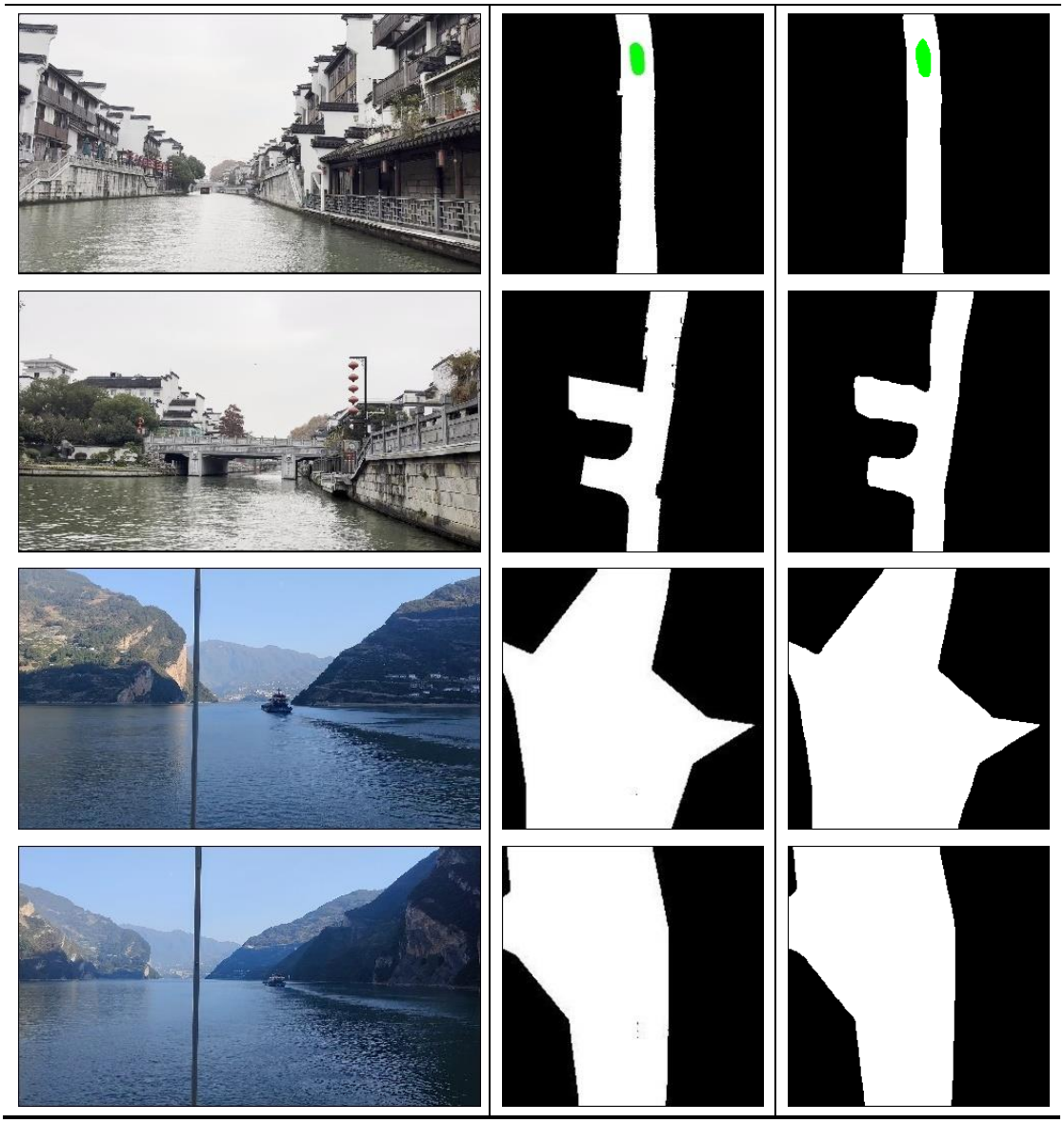
Figure 9 illustrates the BEV layout generation results of SECross for navigational waterways in different scenes, primarily evaluating its experimental performance on various sizes of ships and diverse shapes of waterways in different maritime environments. From Figure 9, it can be observed that SECross accurately generates BEV layouts for virtual, visible lights, and infrared navigational waterway images, indicating that the proposed algorithm exhibits strong generalization and can be applied to various scenarios. In Figure 9 (a) third row, (b) second row, and (c) first row, SECross demonstrates satisfactory recognition performance for cross-waterways and curves in different scenarios, highlighting its strong capability to extract features from waterways with diverse shapes. Additionally, SECross can generate cross-waterways occluded by bridges, indicating that the algorithm's generative network has strong completion capabilities. Furthermore, SECross exhibits high identification accuracy for ships in waterways. The last three rows of Figure 9 (a), the first row of (b), and the second row of (c) demonstrate that SECross does not exhibit omissions or misidentification issues for small-scale or even miniature ships in various scenarios. This

underscores the algorithm's strong capability in extracting features of small-scale targets, effectively mitigating the impact of scene complexity, structures, obstacles, and other interferences. The first row of Figure 9 (a) indicates that even in scenarios with dense ship traffic and occlusion, SECross can extract features of obscured vessels and accurately generate the BEV layout. This suggests that SECross achieves precise segmentation for dense ship scenarios in navigational waterway images and exhibits good recognition capabilities for complex situations like intersecting traffic. In conclusion, the application of attention mechanisms and Cross-View Transformation Modules in SECross enables it to capture prominent features of vessels in situations with dense small targets.

FPV	Waterway-Truth	SECross
		
		
		
		

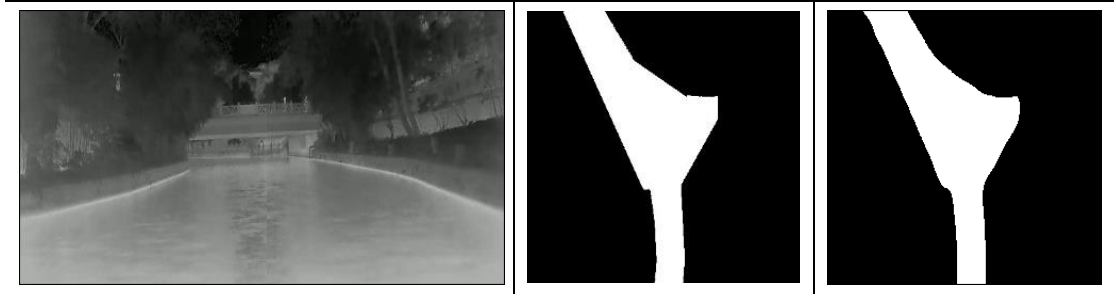
(a) Generation of various virtual waterways

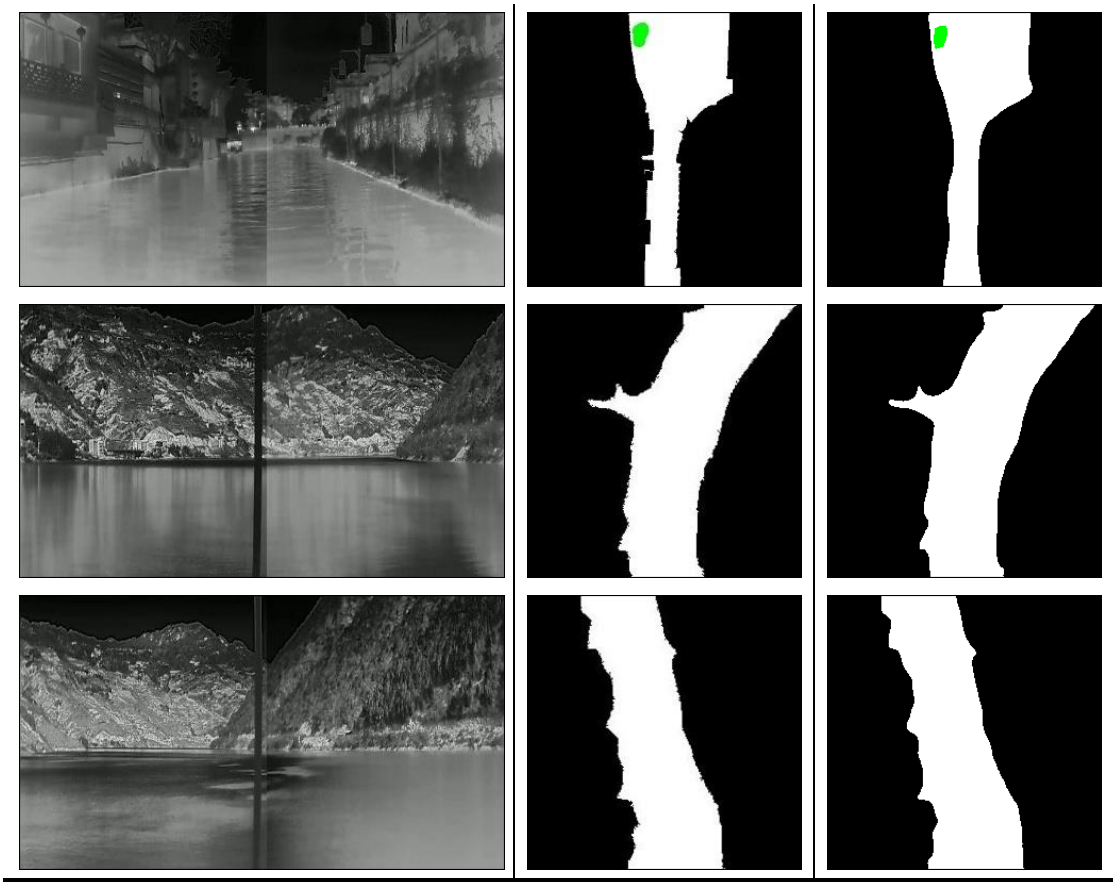
FPV	Waterway-Truth	SECross



(b) Generation of various visible light waterways

FPV	Waterway-Truth	SECross





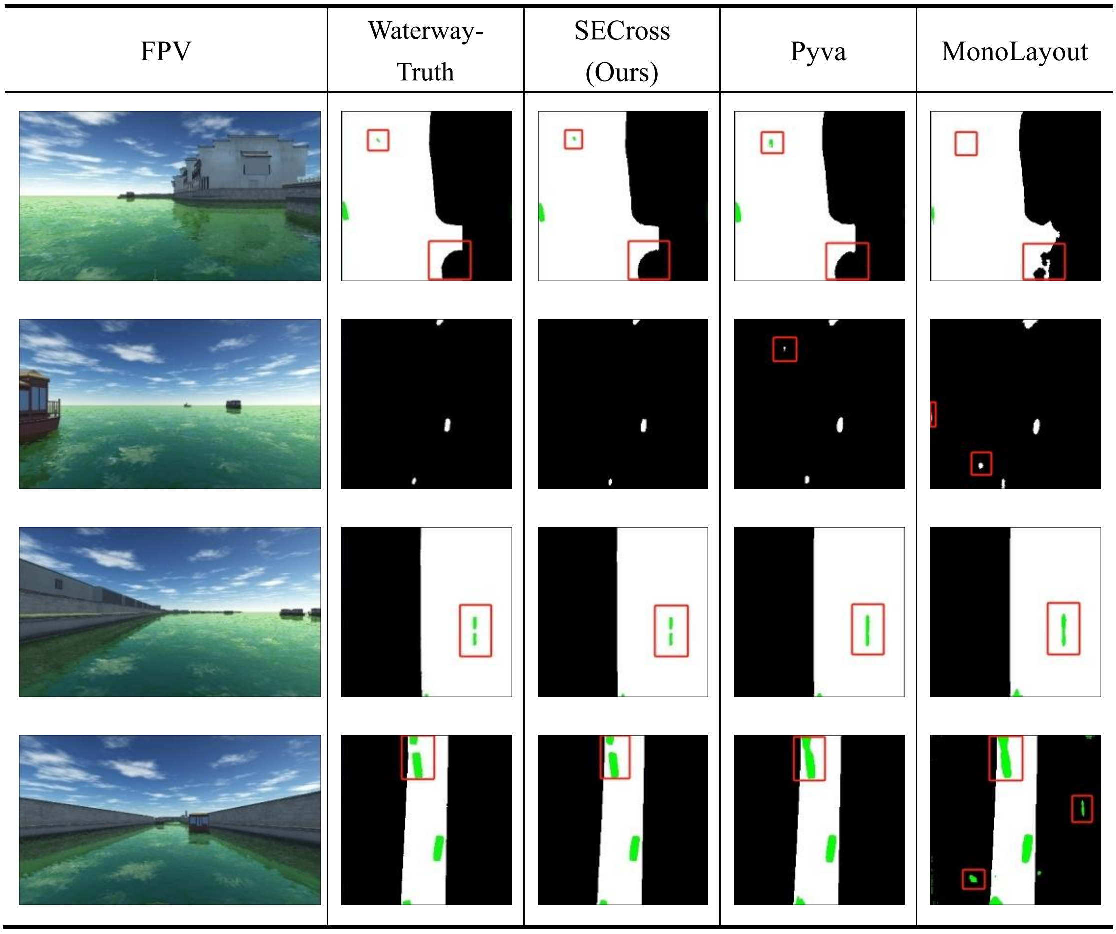
(c) Generation of various infrared waterways

Figure 9. Generation results of SECross under waterway images

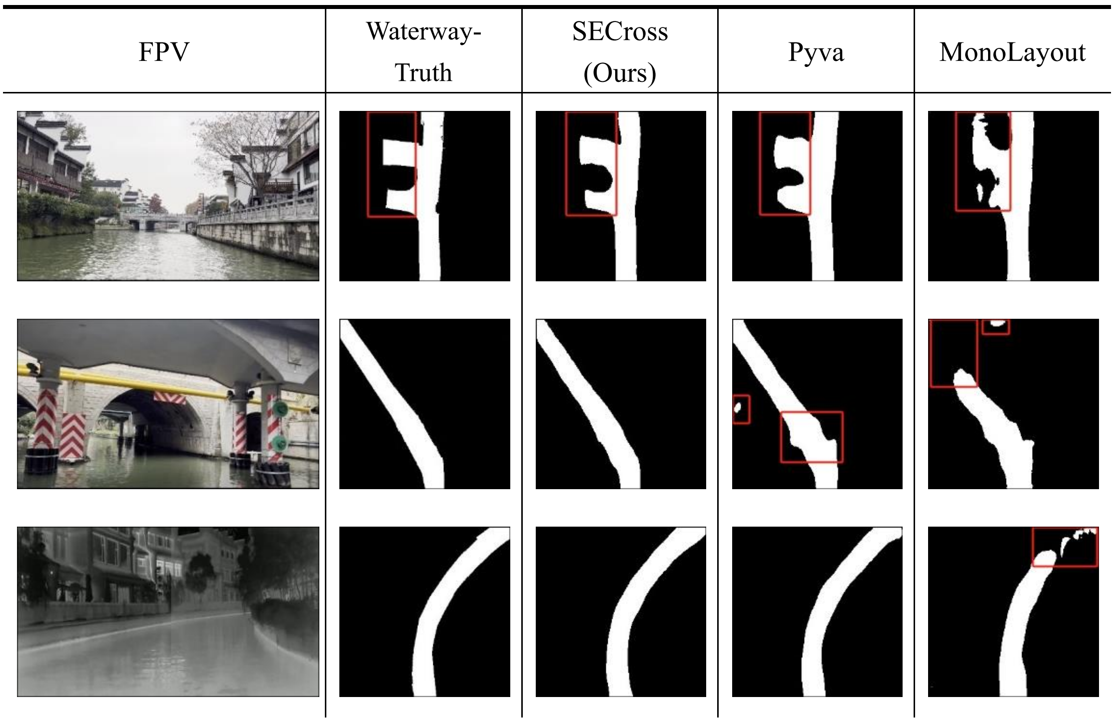
4.3.4. BEV layout generation of special scenarios in inland waterways

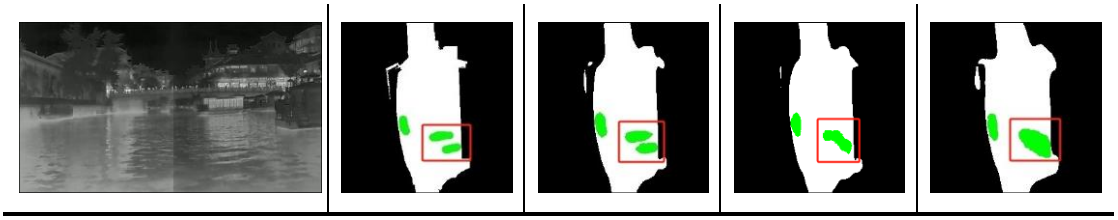
As is well known, there are some specific navigation scenarios in inland waterways, posing numerous challenges for accurate BEV generation in these situations. To assess the BEV generation performance of different algorithms in these scenarios, this research conducted experiments in four settings: sharp bends with high curvature, cross-waterways, dense navigation of ships, and small-sized ships. These scenarios involve numerous small targets, and the identification of cross-waterways particularly tests the model's performance. To enhance the credibility of the conclusions, this experiment compared SECross with MonoLayout and Pyva. As depicted in the first two rows of Figure 10(a), SECross demonstrates a more accurate positioning of small-sized ships compared to the contrastive algorithms, indicating superior detection precision for small targets. The analysis of the first row in Figures 10(a) and (b) leads to the conclusion that MonoLayout exhibits poor cross-waterway recognition capabilities in both real and virtual scenarios, while Pyva can generate cross-waterways, its performance in detailing the shapes of these junctions is not as proficient as SECross. From the last two rows of Figure 10(a) and the final row in Figure 10(b), it is evident that when ships navigate densely, Pyva and MonoLayout tend to identify them as a single entity, whereas SECross accurately segments them into multiple individuals. Finally, examining the second and third rows of Figure 10(b), it becomes apparent that, whether in visible light or infrared scenarios, MonoLayout may miss sharp bends or curves of waterways, while SECross and Pyva exhibit comparable generation capabilities. However, Pyva might introduce errors in some scenario. These results indicate that by constructing a rational network structure, SECross can effectively extract features of waterways and ships from FPV images, significantly enhancing the model's performance

in BEV generation tasks in specific scenarios.



(a) Comparison of special virtual scenarios





(b) Comparison of special real-world scenarios

Figure 10. Comparison of BEV layout Generation results for special scenarios across various algorithms.

5. CONCLUSIONS AND DISCUSSIONS

This paper introduced a BEV layouts generation algorithm named SECross relied on monocular FPV images of waterways. SECross employs a novel feature extraction network called SEResNeXt, integrating group convolution and attention mechanisms to precisely extract key features from inland waterway images. Additionally, it enhances the extracted features by introducing a Cross-View Transformation Module to connect features from different perspectives. In the calculation of the loss function, SECross improves the loss computation of the generator network using the Focal Loss function, thereby generating a more accurate BEV layout for the waterway.

Experimental results have demonstrated that SECross outperforms other available algorithms in BEV layout generation on waterway images. SECross accurately generates BEV layouts for challenging scenarios like curves and bifurcations, achieving mIOU and mAP values of over 95% in various complex scenarios. In the generation of ship BEV, SECross exhibits satisfactory performance with mIOU and mAP values exceeding 85%, especially in scenarios involving micro-sized ships and dense ship traffic. Considering the relatively limited image samples, future research will focus on expanding the MonoWaterwayGen dataset, extending it beyond the Qinhuai River, Yangtze River to coastal areas, and incorporating more real-world ship data. Additionally, the next steps in research will involve introducing more interference factors to enhance the model's robustness.

Acknowledgements

This work is financially supported by the Funds for the National Key R&D Program of China (Grant No. 2021YFB1600400), National Natural Science Foundation of China under Grant Numbers 52171352 and 52201415.

References

[1] Yenıaydin Y, Schmidt K W. Sensor fusion of a camera and 2d lidar for lane detection[C]//2019 27th Signal Processing and Communications Applications Conference (SIU). IEEE, 2019: 1-4.

[2] He W, Ma F, Liu X. A recognition approach of radar blips based on improved fuzzy c means[J]. Eurasia Journal of Mathematics, Science and Technology Education, 2017, 13(8): 6005-6017.

[3] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[4] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[5] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[6] Jie Y, Leonidas L A, Mumtaz F, et al. Ship detection and tracking in inland waterways using improved YOLOv3 and Deep SORT[J]. Symmetry, 2021, 13(2): 308.

[7] Zhang W, He X, Li W, et al. An integrated ship segmentation method based on discriminator and extractor[J]. Image and Vision Computing, 2020, 93: 103824.

[8] Yin Y, Guo Y, Deng L, et al. Improved PSPNet-based water shoreline detection in complex inland river scenarios[J]. Complex & Intelligent Systems, 2023, 9(1): 233-245.

[9] Gao L, Shu G, Wei H. Adversarial unsupervised domain adaptive inland vessel detection method[C]//Second International Conference on Algorithms, Microchips, and Network Applications (AMNA 2023). SPIE, 2023, 12635: 352-358.

[10] Yu N, Fan X, Deng T, et al. Ship Detection in Inland Rivers Based on Multi-Head Self-Attention[C]//2022 7th International Conference on Signal and Image Processing (ICSIP). IEEE, 2022: 295-299.

[11] Zhang W B, Wu C Y, Bao Z S. SA-BiSeNet: Swap attention bilateral segmentation network for real-time inland waterways segmentation[J]. IET Image Processing, 2023, 17(1): 166-177.

[12] Cai H, Zhang Z, Zhou Z, et al. BEVFusion4D: Learning LiDAR-Camera Fusion Under Bird's-Eye-View via Cross-Modality Guidance and Temporal Aggregation[J]. arXiv preprint arXiv:2303.17099, 2023.

[13] Liu Z, Tang H, Amini A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]//2023 IEEE international conference on robotics and automation (ICRA). IEEE, 2023: 2774-2781.

[14] Liang T, Xie H, Yu K, et al. Bevfusion: A simple and robust lidar-camera fusion framework[J]. Advances in Neural Information Processing Systems, 2022, 35: 10421-10434.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural

information processing systems, 2017, 30.

[16] Li Z, Wang W, Li H, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 1-18.

[17] Lu C, van de Molengraft M J G, Dubbelman G. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 445-452.

[18] Mani K, Daga S, Garg S, et al. Monolayout: Amodal scene layout from a single image[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 1689-1697.

[19] Zhou B, Krähenbühl P. Cross-view transformers for real-time map-view semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13760-13769.

[20] Yang W, Li Q, Liu W, et al. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15536-15545.

[21] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[22] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[23] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[24] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[25] Pan B, Sun J, Leung H Y T, et al. Cross-view semantic segmentation for sensing surroundings[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4867-4873.

[26] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[27] Berrada T, Verbeek J, Couprie C, et al. Unlocking Pre-trained Image Backbones for Semantic Image Synthesis[J]. arXiv preprint arXiv:2312.13314, 2023.