frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 19 March 2024
DOI 10.3389/frsen.2024.1370697

OPEN ACCESS

EDITED BY
Nan Gao,
Tsinghua University, China

REVIEWED BY
Yuxi Liu,
Flinders University, Australia
Shusen Jing,
University of California, San Francisco, United States
Chongyang Wang,
Tsinghua University, China

*CORRESPONDENCE
Gang Liu,
43501718@qq.com

RECEIVED 15 January 2024
ACCEPTED 06 March 2024
PUBLISHED 19 March 2024

CITATION
He S, Jin C, Shu L, He X, Wang M and Liu G (2024), A new framework for improving semantic segmentation in aerial imagery. *Front. Remote Sens.* 5:1370697. doi: 10.3389/frsen.2024.1370697

A new framework for improving semantic segmentation in aerial imagery

Shuke He[1], Chen Jin[1], Lisheng Shu[1], Xuzhi He[2], Mingyi Wang[3] and Gang Liu[1]*

[1]Research Department of Aeronautics, Zhejiang Scientific Research Institute of Transport, Hangzhou, China, [2]Department of Biomedical Engineering, Univiersity of California, Davis, CA, United States, [3]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China

High spatial resolution (HSR) remote sensing imagery presents a rich tapestry of foreground-background intricacies, rendering semantic segmentation in aerial contexts a formidable and vital undertaking. At its core, this challenge revolves around two pivotal questions: 1) Mitigating Background Interference and Enhancing Foreground Clarity. 2) Accurate Segmentation in Dense Small Object Cluster. Conventional semantic segmentation methods primarily cater to the segmentation of large-scale objects in natural scenes, yet they often falter when confronted with aerial imagery's characteristic traits such as vast background areas, diminutive foreground objects, and densely clustered targets. In response, we propose a novel semantic segmentation framework tailored to overcome these obstacles. To address the first challenge, we leverage PointFlow modules in tandem with the Foreground-Scene (F-S) module. PointFlow modules act as a barrier against extraneous background information, while the F-S module fosters a symbiotic relationship between the scene and foreground, enhancing clarity. For the second challenge, we adopt a dual-branch structure termed disentangled learning, comprising Foreground Precedence Estimation and Small Object Edge Alignment (SOEA). Our foreground saliency guided loss optimally directs the training process by prioritizing foreground examples and challenging background instances. Extensive experimentation on the iSAID and Vaihingen datasets validates the efficacy of our approach. Not only does our method surpass prevailing generic semantic segmentation techniques, but it also outperforms state-of-the-art remote sensing segmentation methods.

KEYWORDS

deep learning, aerial imagery, remote sensing segmentation, foreground saliency enhancement, small objects semantic segmentation

1 Introduction

Deep neural networks (DNNs) have revolutionized city management, achieving remarkable success (Shao et al., 2021; Shao et al., 2022a; He et al., 2022; Kang et al., 2022; He et al., 2023). However, the complexity intensifies in high spatial resolution (HSR) remote sensing images, featuring diverse geospatial entities like aircraft, vessels, and buildings. Deciphering these entities is crucial for urban monitoring (Volpi and Ferrari, 2015; Kemker et al., 2017; Shao et al., 2022b), posing challenges due to unique imaging mechanisms and scene intricacies. In this context, semantic segmentation in very-high-

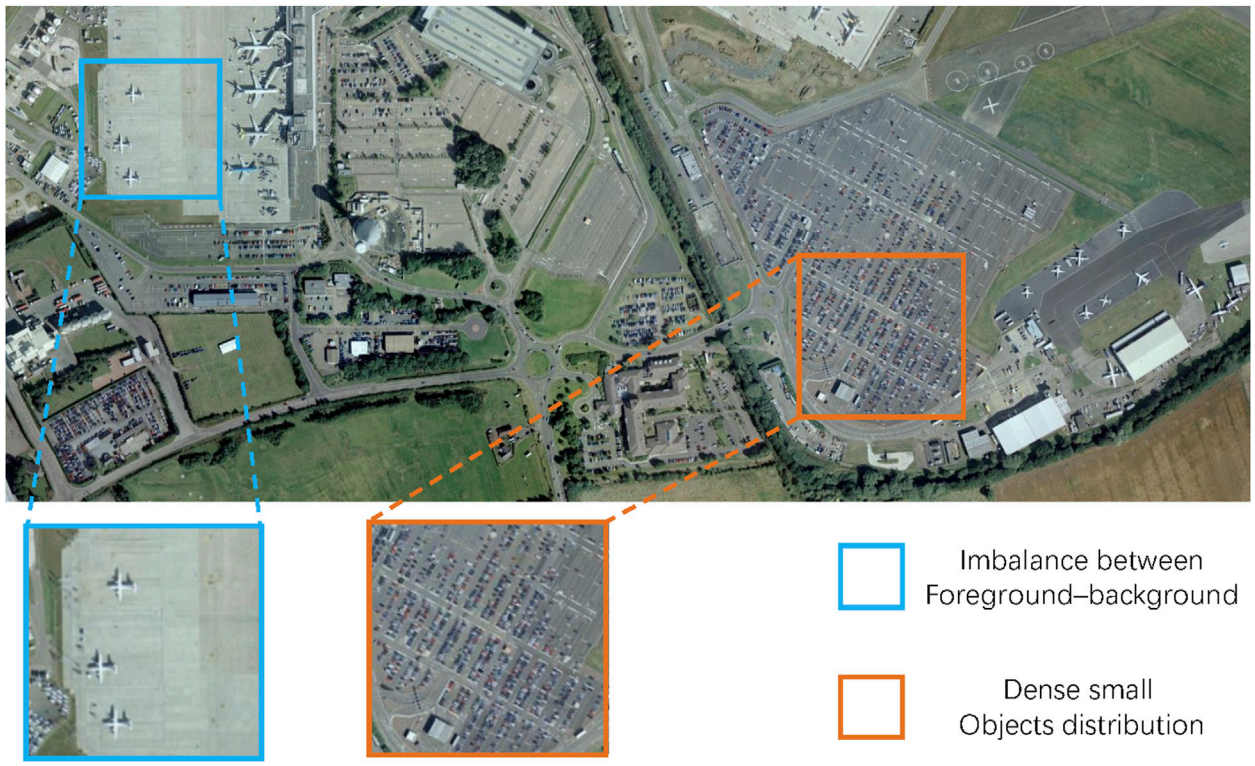

FIGURE 1
Illustration of two pressing challenges when addressing the task of semantic segmentation in HSR imagery. As the blue box shows, there is a clear imbalance in the proportion of aircraft and airports. And in the orange box there are a lot of dense cars.

resolution (VHR) aerial imagery becomes a formidable task, acting as a crucial intermediary step between raw images and vector map layers.

The semantic segmentation process involves extracting foreground objects and predicting pixel-level probabilities within these objects. This intricate task is crucial for understanding the urban environment and its complexities. Compared with the semantic segmentation task in natural scenes, semantic segmentation of geographical objects is more challenging in VHR aereial imagery, surpassing conventional large-scale variations (Zheng et al., 2020; Hou et al., 2022), as shown in Figure 1. In iSAID dataset imagery (Zamir et al., 2019), foreground objects like vehicles occupy a mere 10 m^2 within images spanning several square kilometers (Ma et al., 2022). Furthermore, given the relatively elevated altitude at which HSR remote sensing images are captured, these images often harbor densely clustered small objects, further exacerbating the intricacies of image segmentation (Li et al., 2021; Ma et al., 2022; Niu et al., 2022). Consequently, when addressing the task of semantic segmentation in HSR imagery, we confront several pressing challenges:

(1) *Mitigating Background Interference and Enhancing Foreground Clarity:* HSR imagery often exhibits an extreme imbalance between foreground and background elements, necessitating an emphasis on accentuating foreground features. This challenge involves mitigating background interference and refining foreground saliency modeling.

(2) *Accurate Segmentation in Dense Small Object Clusters:* In scenarios with densely distributed small objects, the primary challenge is precision in pinpointing object clusters and delineating intricate contours. This requires a meticulous approach to edge segmentation and object localization within crowded contexts.

The general semantic segmentation methods (Pinheiro and Collobert, 2014; Long et al., 2015; Chen et al., 2018a) mainly focus on multi-scale modeling. However, for aerial imagery, these semantic segmentation methods ignore the problem of foreground-background imbalance and small object aggregation. Recently, in order to solve the problem of foreground-background imbalance in aerial imagery, some researches (Zheng et al., 2020; Li et al., 2021) improve the segmentation performance by enhancing foreground significance. However, these methods do not take into account that there are a large number of densely distributed small objects in the foreground, so the foreground target may be lost and the boundary tends to be fuzzy.

In this paper, our enhanced semantic segmentation framework addresses both foreground-background imbalance and dense small object challenges. In our pursuit of reducing background interference and enhancing the clarity of foreground objects, we draw inspiration from the innovative work of PointFlow (Li et al., 2021) and FarSeg (Zheng et al., 2020). To achieve this, we introduce two essential components into our framework: the PointFlow Modules (PFMs) and the Foreground-Scene (F-S) Module. PointFlow modules serve as a bulwark against the undue influx

of background information, preserving the integrity of foreground modeling. Meanwhile, the F-S module orchestrates the cultivation of a symbiotic relationship between the scene and foreground, bolstering the prominence of foreground features.

The PFMs play a pivotal role in curtailing the intrusion of background information during foreground modeling. We strategically incorporate PFMs into the feature pyramid network (FPN) to carefully select representative points between adjacent feature pyramid levels. This approach is a departure from previous techniques like simple fusion or dense affinity propagation (Fu et al., 2019) applied to each point, as seen in non-local modules (Wang et al., 2018). The outcome is twofold: a noticeable reduction in background noise and a substantial improvement in segmentation efficiency. Meanwhile, the Foreground-Scene (F-S) Module capitalizes on the symbiotic connection between the scene and foreground. By assimilating foreground-related context, we boost the features of foreground objects. This modeling of the relationship between foreground and geographic spatial scenes is used to enhance the input feature map. This, in turn, widens the gap between foreground and background features, ultimately enhancing the distinctiveness of foreground features.

In our pursuit of accurate segmentation within densely populated small object clusters, we emphasize the explicit modeling of foreground and boundary objects as indispensable in aerial imagery semantic segmentation. Here, our inspiration comes from disentangled learning methodologies (Higgins et al., 2017; Niu et al., 2022). We adopt a two-branch structure, dedicating one branch to the aggregation positioning of small objects through Foreground Precedence Estimation (FPE) and the other to dynamically correcting the edge contours of small objects via end-to-end training, which we term Small Object Edge Alignment (SOEA). This dual-branch disentangled learning approach effectively mitigates issues such as the loss of fine-grained information and the production of blurry boundary contours, common drawbacks in joint learning schemes.

The challenge of foreground-background imbalance frequently results in background samples dominating the training process, causing early saturation and hindering model optimization. Drawing inspiration from approaches like Xu et al. (2023); Zheng et al. (2020), we recognize that in the latter stages of training, further refinement of intricate background features, like distinctive texture characteristics, is essential to prevent overfitting and enhance the segmentation model's performance. To address this issue, we propose the introduction of a foreground saliency guided loss. This loss function serves to suppress the undue influence of numerous easy background examples, effectively mitigating the foreground-background imbalance challenge. We note that a shorter conference version of this paper appeared in CBASE 2023 (Jin et al., 2023). Our initial conference paper did not provide experimental validation that the foreground and scene are related, nor does it perform ablation studies on the SOEA module. In this manuscript we conducted more experiments to verify our method. Some parts of the paper have also been reasonably rearranged. In summary, our study contributes significantly to the field of aerial image semantic segmentation through the following key advancements:

- Comprehensive Analysis and Framework Proposal: We conduct a thorough analysis of challenges in aerial images for semantic segmentation, leading to the introduction of a novel framework based on feature pyramids. This framework adaptively addresses issues of foreground-background imbalance and the complexities of dense small object distribution.
- Innovative Module Integration: To overcome foreground modeling challenges, we integrate PointFlow Modules (PFMs) and a foreground-scene module. PFMs enhance foreground saliency by mitigating background interference, while disentangled learning ensures accurate segmentation of densely distributed small objects, avoiding risks associated with joint learning.
- Foreground Saliency Guided Loss: We introduce a novel foreground saliency guided loss during training to prevent segmentation errors resulting from model overfitting to simplistic background features, ensuring robust performance.
- Empirical Validation: Extensive experiments on widely recognized datasets provide solid empirical evidence supporting the effectiveness of our proposed segmentation framework.

2 Related work

2.1 General semantic segmentation

The field of semantic segmentation has witnessed significant evolution, primarily driven by advancements in deep learning techniques. Traditional methods heavily reliant on handcrafted features and rule-based approaches have faced limitations in performance and generalization. The advent of deep learning, particularly convolutional neural networks (CNNs), has revolutionized semantic segmentation. Early explorations employed CNNs for patch-wise classification, a structured feature representation approach (Gupta et al., 2014; Pinheiro and Collobert, 2014). However, this method posed challenges such as information loss and redundant computations in overlapping areas between patches. To overcome these limitations, the fully convolutional network (FCN) (Long et al., 2015) was introduced, replacing fully connected layers with convolutional layers to preserve spatial information. While FCN marked a significant advancement, subsequent CNN-based methods have continued to push the boundaries of semantic segmentation. Deeplab v1 (Chen et al., 2015), for instance, leveraged atrous convolution to expand the CNN's receptive field and enhance spatial context awareness. Approaches like ASPP (Chen et al., 2018a) and PPM (Zhao et al., 2017a) further extended these ideas by utilizing atrous convolutions with different rates and generating pyramidal feature maps through pyramid pooling. Nevertheless, these methods often struggled to capture fine details, particularly object boundaries. To address this limitation, architectures like U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2017) introduced the "encoder-decoder" network paradigm, effectively refining output details. NLNet (Wang et al., 2019) took a different approach by incorporating non-local operators or losses to capture global context in input images. RefineNet

(Lin et al., 2017a), on the other hand, presented a multi-path refinement network designed to progressively recover spatial details, resulting in improved accuracy and enhanced object boundaries. The above method does have a good performance in the natural scene, but the effect is not satisfactory if directly applied to aerial imagery. These methods ignore the complex background information of aerial imagery and the challenges of dense object distribution (as mentioned in Section 1).

2.2 Semantic segmentation in remote sensing

The realm of semantic segmentation in remote sensing boasts a diverse array of noteworthy contributions, each tailored to specific applications. The applications span across Land Use/Land Cover (LULC) classification (Onim et al., 2020), building extraction (He et al., 2021), road delineation (Bastani et al., 2018; Dickenson and Gueguen, 2018; Liang et al., 2019), and more, all of which hinge on the accurate delineation of objects within remote sensing imagery. In this landscape, deep learning has emerged as a focal point, revolutionizing computer vision and pattern recognition (Chen et al., 2016). One notable endeavor by Wang et al. (Wang et al., 2021) introduces a hierarchical neural network search framework that automatically crafts architectures for remote sensing recognition. While these semantic segmentation CNNs often build upon or adapt advanced CNN architectures, they frequently neglect the nuances within small objects prevalent in High Spatial Resolution (HSR) imagery. To address this limitation, relation networks Mou et al. (2019) have made significant strides in semantic segmentation by modeling spatial relationships among pixels and interactions between objects. This approach bolsters segmentation accuracy, consistency, and overall coherence. More recently, FarSeg (Zheng et al., 2020) has harnessed spatial relationship modeling to discern foreground-background boundaries accurately, particularly addressing the foreground-background imbalance issues prevalent in remote sensing imagery. Concurrently, PFNet (Li et al., 2021) has made strides in advancing the comprehension and propagation of semantic information within the context of semantic segmentation. FactSeg, a groundbreaking contribution by Ma et al. (Ma et al., 2022), pioneers a foreground activation-driven approach to small object semantic segmentation. In a quest to streamline architectural complexity, Xie et al. (Xie et al., 2021) introduce SegFormer, a transformer-based model that maintains a lightweight design and employs a simplified segmentation head to yield final results. However, these methods usually only get better results by improving the model of foreground significance, ignoring the dense distribution of small objects in the foreground target, and their segmentation methods are easy to make the foreground small objects lost and the boundary tends to be blurred. Different from these methods, our method can better segment dense small objects in fine detail while improving foreground saliency.

3 Methods

3.1 Overall Framework

Semantic segmentation in aerial imagery presents a multifaceted challenge, particularly when it comes to discerning foreground and boundary objects. To overcome these challenges, we introduce an advanced semantic segmentation framework, illustrated in Figure 2. This framework encompasses a suite of critical components, including a Feature Pyramid Network (FPN) enhanced with Point Flow Modules (PFMs), a Foreground-Scene (F-S) Module, disentangled learning, and optimization guided by foreground saliency.

As we mentioned in Section 1, aerial imagery semantic segmentation needs to address two key challenges. For the problem of foreground background imbalance, PointFlow modules serve as a bulwark against the undue influx of background information, preserving the integrity of foreground modeling. Meanwhile, the F-S module orchestrates the cultivation of a symbiotic relationship between the scene and foreground, bolstering the prominence of foreground features. Specifically, we embed PFMs in FPN to selectively select feature points in different feature layers of FPN. Meanwhile, the F-S Module delves into the intricate interplay between the scene and foreground, endowing foreground features with enhanced discriminatory capabilities. Moreover, considering that there are a large number of dense small objects in the extracted foreground that need to be finely segmented, we incorporate disentangled learning into our framework. This strategic approach heightens the significance of intermediate features, imbuing the network with stronger discriminative abilities while effectively aligning boundary features. Finally, in order to solve the problem of overfitting background features in the training process caused by the wide coverage of background, we propose the introduction of a foreground saliency guided loss. This loss function serves to suppress the undue influence of numerous easy background examples, effectively mitigating the foreground-background imbalance challenge. This comprehensive framework equips us to navigate the complexities of semantic segmentation in aerial imagery, paving the way for more accurate and robust results.

3.2 FPN with PFMs

A prominent challenge in semantic segmentation arises from the semantic disparities within the Feature Pyramid Network (FPN), primarily concerning foreground objects. This disparity manifests as a gap between high-resolution features, which offer limited semantic information, and low-resolution features, endowed with more profound semantic content. Notably, addressing this gap proves crucial for tiny objects that necessitate richer semantic information, even within high-resolution layers. However, prior methods, exemplified by the fusion of the entire feature set (Zhang et al., 2020), inadvertently emphasize background objects like roads, exacerbating the well-documented imbalance issue in aerial imagery. To redress this imbalance and quell background noise, we introduce a modified version of the Feature Pyramid Network (FPN) (Li et al., 2021).

Our strategy involves the integration of PointFlow Modules (PFMs) into the FPN architecture to facilitate targeted semantic point propagation between adjacent features. Unlike conventional methods (Wang et al., 2018), which apply uniform fusion or dense affinity propagation across all points, PointFlow takes a distinctive approach. It selectively identifies a subset of representative points between adjacent feature pyramid levels, subsequently computing

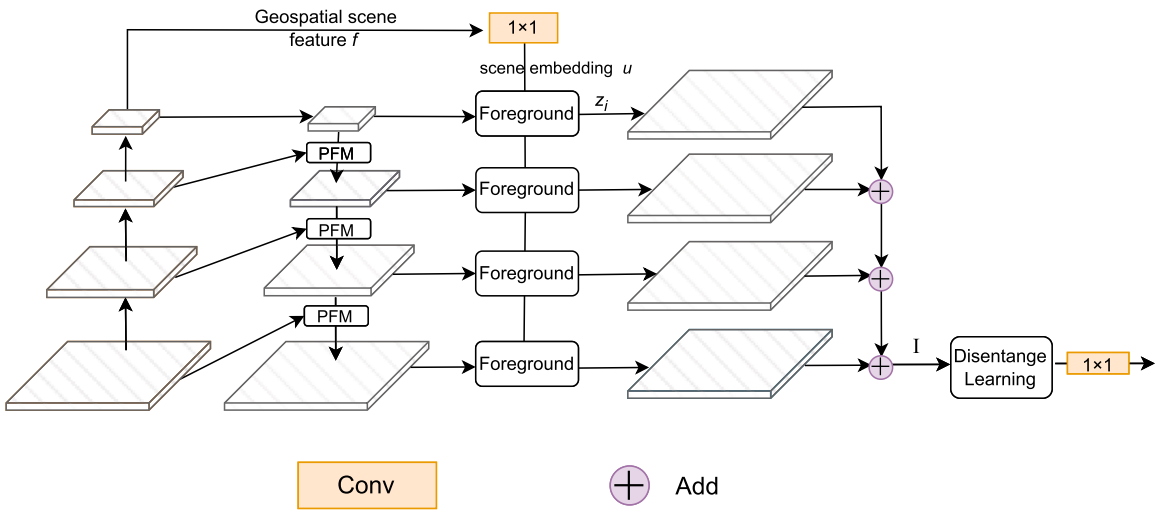

FIGURE 2
Overall framework of our proposed framework. The PFMs (see Section 3.2) are embedded in each stage of the FPN. The foreground represents the foreground-scene module (see Section 3.3). The disentangled learning is shown in Section 3.4.

point-wise affinities among these selected points. Ultimately, high-resolution, low-semantic points benefit from their low-resolution, high-semantic counterparts, guided by the estimated affinity map. Through the incorporation of PFMs into the FPN's feature pyramids, we concurrently surmount the semantic challenges and foreground-background imbalance.

As depicted in Figure 2, our segmentation framework is structured around a dual-path architecture, comprising a bottom-up encoder and a top-down decoder. The encoder, serving as the foundation of the network, generates multiple feature pyramid outputs. Conversely, the decoder utilizes an enhanced FPN, bolstered by the integration of PFMs, replacing the conventional bilinear upsampling employed in the top-down FPN pathway. Parallel to the original FPN, this top-down pathway, complemented by lateral connections, generates pyramidal feature maps v_i) with uniform channel numbers. With this top-down connection and horizontal connection, feature maps can get more detailed information from shallow layers and more semantic information from deeper layers. In addition to v_i, to better aggregate global information, we have also introduced a auxiliary branch for creating a geospatial scene feature represented as f, which can help model the relationship between the foreground and the scene. To obtain f, we use global average pooling as the aggregation function. The geospatial scene feature assumes a pivotal role in modeling the intricate relationship between the scene and the foreground, a concept elaborated further in Section 3.3.

3.3 Foreground-scene module

HSR remote sensing imagery introduces a significantly more intricate background compared to conventional imagery. This heightened complexity results in a greater intraclass variance within the background, consequently giving rise to the issue of false alarms in semantic segmentation. To mitigate this problem, we introduce a F-S Module, inspired by the work of Zheng et al. (2020). This module is designed to capture the close relationship between the scene and foreground, with a specific focus on enhancing the foreground while reducing the impact of irrelevant background information. The main concept behind the F-S Module is to clearly represent how foreground objects interact with the surrounding scene. It uses hidden spatial information to create meaningful connections between individual foreground objects and their surroundings. Once these connections are established, they are used to enhance the input feature maps, making it easier to distinguish between foreground and background features. This enhancement helps improve the accuracy of identifying foreground objects and reduces false alarms.

Our main idea is shown in Figure 3. We first model the relationship between foreground and scene, and then use geospatial scene feature $f$to associate foreground and related context information. Finally, these relationships are used to enhance the input feature map to improve the foreground salience. As shown in Figure 3, the pyramidal feature maps v_i) undergo a compression process aimed at achieving uniform channel depth. This is achieved through 1 × 1 convolution layers, which are subsequently subjected to batch normalization and ReLU activation. Simultaneously, a scene embedding vector u) is computed using a 1 × 1 convolutional layer with an output channel size of d_u, applied to the geospatial scene feature f). Notably, this scene embedding vector u) remains consistent across all pyramids, as latent geospatial scene semantics inherently exhibit scale-invariance. Consequently, the relation maps (r_i) are naturally derived via pointwise inner product computations, ensuring a streamlined and computationally efficient process.

The formation of the new feature map (z_i) is subsequently carried out using the following equation:

$$z_i = \frac{1}{1+\exp(-r_i)} \cdot \psi_{\theta_i}(v_i), \tag{1}$$

where $\psi_{\theta_i}(\cdot)$ denotes the encoder, which has specific settings called parameters (θ_i). The encoder introduces an additional non-linear component to counteract feature degradation, as the weighting

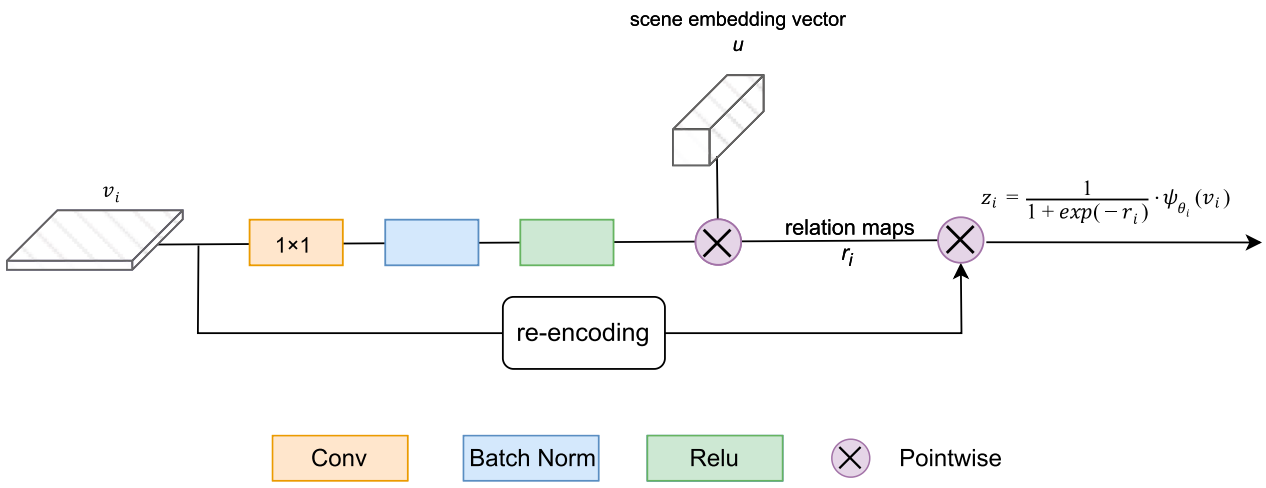

FIGURE 3
Foreground-scene module.

TABLE 1 The results of F-S module in the different pyramid levels on iSAID dataset. OS stands for output stride.

Scene	Prediction	Relation (OS = 4)	Relation (OS = 8)	Relation (OS = 16)
airport	0.75	0.72	0.69	0.66
harbor	0.73	0.71	0.67	0.64
parking-lot	0.76	0.73	0.71	0.67

operation is inherently linear. Consequently, we use an efficient encoder design, comprising a 1 × 1 convolutional layer followed by batch normalization and ReLU activation, optimizing both parameter efficiency and computational speed. The relation maps (r_i) serve as weighting factors, employing a normalized relation map that utilizes a sigmoid gate function (Hu et al., 2018). We tested the effectiveness of the F-S module in the different pyramid levels on iSAID dataset, as shown in Table 1. From left to right: original scene, segmentation prediction results, and images with F-S relation heatmaps in the different pyramid level. It can be seen from the experimental results that the spatial scene and the foreground information are related.

3.4 Disentangled learning

In contemporary semantic segmentation methodologies, the prevailing trend involves joint learning approaches that, while effective, may inadvertently neglect potential ambiguities inherent to coupled features. Recent advancements in the field (Yin et al., 2020; Yuan et al., 2020; Niu et al., 2022) have illuminated the benefits of disentangled designs for the explicit extraction of features tailored to various tasks. Inspired by the work of Niu et al. (2022), we adopt a disentangled learning method to explicitly represent foreground objects and align the edge features of small objects. This disentangled approach encompasses two pivotal components: Foreground Precedence Estimation and Small Object Edge Alignment.

3.4.1 Foreground precedence estimation

Considering the serious foreground and background imbalance in aerial imagery, we hope to model more important foreground objects. As depicted in the upper portion of Figure 4, this process begins with obtaining a compact feature using a specialized type of convolution called depthwise separable convolution. This feature is then made larger and combined with the original feature. Then, the foreground mask $M_{fg} \in \mathbb{R}^{1\times H\times W}$ is generated through the 1 × 1 convolution. These priors capture both foreground and background contexts, expressed as:

$$B = \delta\left(M_{fg} \cdot I \| \left(1 - M_{fg}\right) \cdot I \| I\right), \tag{2}$$

Here, $I \in \mathbb{R}^{512\times H\times W}$ represents the combined feature obtained from the FPN, and ‖ represents the concatenation operation. The aggregated feature $B \in \mathbb{R}^{C/2\times H\times W}$ is subsequently utilized for propagation. The transformation function $\delta(\cdot)$ is achieved through a 3 × 3 convolution followed by batch normalization and ReLU activation.

3.4.2 Small object edge alignment

As visualized in the lower section of Figure 4, dense prediction tasks often struggle with boundary pixels due to object characteristics and image resolution limitations. These boundary pixels correspond to high-frequency regions within the image, marked by rapid feature changes. Drawing inspiration from Niu et al. (2022) and Yuan et al. (2020), we adopt a small object edge

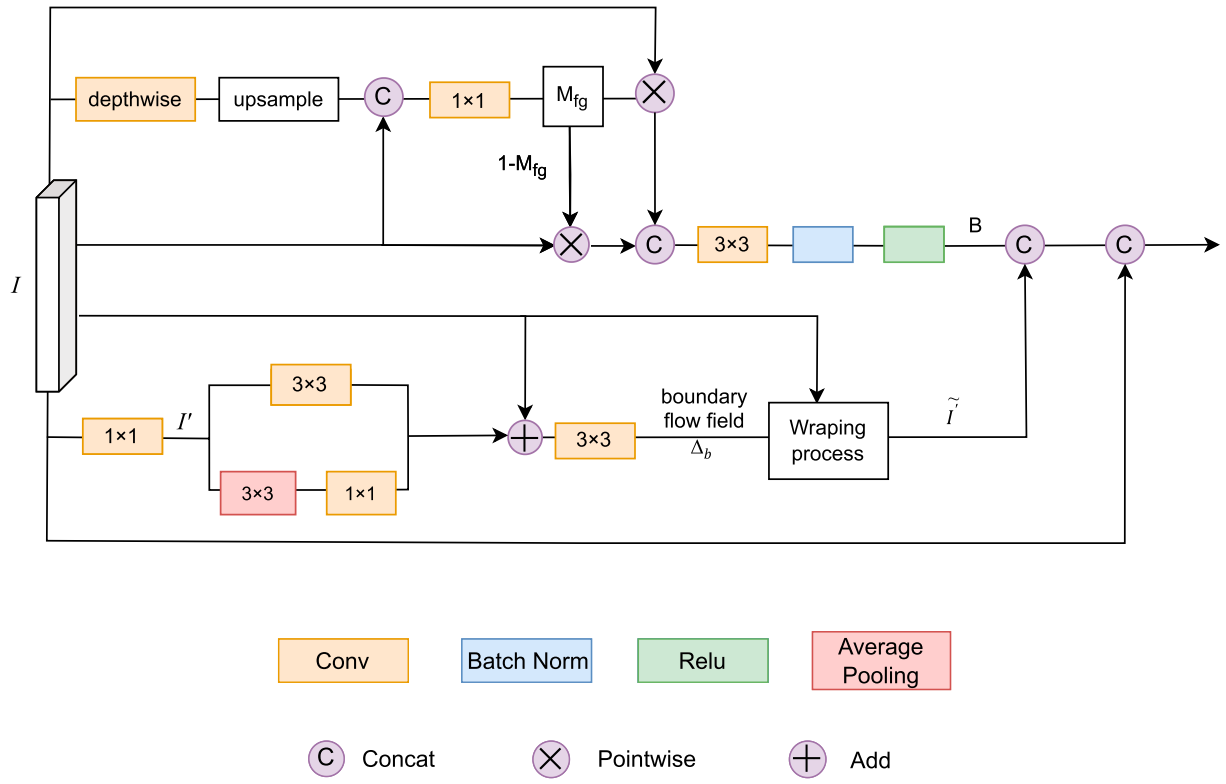

FIGURE 4
Overview of disentangled learning module. The top half: Foreground Precedence Estimation. The bottom half: Small Object Edge Alignment.

alignment (SOEA) approach to indirectly supervise the alignment of edge features.

Starting with the feature $I \in \mathbb{R}^{C\times H\times W}$ from the FPN, we use a 1 × 1 convolution to reduce the channel dimension, yielding $I' \in \mathbb{R}^{C/2\times H\times W}$. Subsequently, the convolution operation and pooling operation are employed to gather contextual information. This information significantly contributes to guiding the subsequent small object edge alignment. This process results in the creation of a boundary flow field denoted as $\triangle_b \in \mathbb{R}^{H\times W\times 2}$, defined as:

$$\triangle_b = \eta[I' \| \varphi(conv(I' + avgPool(I')))], \tag{3}$$

Here, $\varphi(\cdot)$ signifies bilinear interpolation, $conv(\cdot)$ denotes the 3 × 3 strided dilated convolution, and $avgPool(\cdot)$ is constructed via applying successive 3 × 3 pooling with a certain step size followed by a 1 × 1 convolution. $\eta(\cdot)$ represents a 3 × 3 convolutional layer.

Once $\triangle_b$ is established, each edge representation j within the feature I' is transformed into a new value k using the learned flow field. This process is defined as follows:

$$\tilde{I}' = \sum_{j\in S_b} W\left(I'_j; \triangle_{b,j\to k}\right), \tag{4}$$

Where $\tilde{I}'$ represents the warped feature, and S_b encompasses the group of pixels related to the boundary. $W(x; w)$ represents the outcome of input x using w. The warped feature $\tilde{I}'$, upon concatenation with the input, undergoes further processing via a convolutional layer, ultimately producing the final output.

3.5 Foreground saliency guided loss

The significant imbalance between foreground and background samples within HSR imagery presents a formidable challenge for segmentation task. This imbalance often results in the domination of background infromation during the training process. It is crucial to recognize that the optimization of the network is profoundly influenced only by the challenging portions of both foreground and background samples. Therefore, effectively leveraging these challenging samples becomes paramount. Drawing inspiration from Zheng et al. (2020); Xu et al. (2023); Lin et al. (2017b), we introduce the foreground saliency guided loss to guide the model's attention towards foreground and challenging examples, thus achieving a balance in optimization while enhancing foreground saliency.

To derive weights that accurately reflect the difficulty level of examples and tailor the pixel-wise loss distribution, we use $(1-p)^{\gamma}$ as the weight for estimating challenging samples. In this formulation, $p \in [0, 1]$ represents the predicted probability, while γ acts as the focusing factor. Specifically, higher weights are assigned to more challenging examples. To modify the distribution of loss without changing the total sum and prevent the issue of gradients disappearing, we introduce a normalization constant, Z, which leads to the following expression:

$$\sum \mathcal{L}(p_i, y_i) = \frac{1}{Z}\sum (1-p_i)^{\gamma}\mathcal{L}(p_i, y_i), \tag{5}$$

Here, $\mathcal{L}(p_i, y_i)$ signifies the cross-entropy (CE) loss for the ith pixel, which is computed using p_i and the ground truth y_i.

In our pursuit of dynamically adjusting the model's discrimination, a pivotal aspect of estimating challenging examples, we introduce a dynamic weighting strategy based on Cosine annealing (Loshchilov and Hutter, 2017). This strategy is formulated as follows:

$$\mathcal{L}'(p_i, y_i) = \left[\frac{1}{Z}(1-p_i)^{\gamma} + \tau(t)\left(1 - \frac{1}{Z}(1-p_i)^{\gamma}\right)\right] \cdot \mathcal{L}(p_i, y_i), \quad (6)$$

Here, $\tau(t)$ represents a cosine annealing function that depends on the current step t, with $\tau(t) \in [0, 1]$ constituting a monotonically decreasing function. This strategy gradually shifts the focus of the loss distribution towards challenging examples as the network's confidence in estimating challenging examples increases with training steps.

4 Experiments

4.1 Experimental setting

4.1.1 Dataset

The iSAID dataset (Zamir et al., 2019) is our primary benchmark for aerial imagery semantic segmentation tasks. With 2,806 high-resolution images ranging from 800 to 13,000 pixels in width, it offers a diverse collection captured by various sensors and platforms. Featuring 655,451 instance annotations across 15 object categories, iSAID stands out as the most extensive dataset for instance segmentation within High Spatial Resolution (HSR) remote sensing imagery. For experimentation, 1,411 images are used for training, and 458 images are reserved for evaluation. The Vaihingen dataset[1] complements our evaluation and contains 33 aerial images of varying sizes, covering an area of 1.38 square kilometers. The dataset categorizes pixels into six distinct land cover classes and includes Digital Surface Models (DSMs) providing crucial height information. Following the division in Mou et al. (2019) and Xu et al. (2023), our utilization divides the dataset into a training set with 11 images and a test set with five images (identified by image IDs 11, 15, 28, 30, and 34).

4.1.2 Implementation detail

As the backbone of our architecture, we leverage ResNet-50/101 (Ferjaoui et al., 2022) models pretrained on the ImageNet dataset (Russakovsky et al., 2015). To adapt these models for our task, we remove the final fully connected layer. All models undergo 16 epochs of training on cropped images. Training employs the stochastic gradient descent (SGD) optimizer with a weight decay of 0.0001 and momentum of 0.9. The implementation of these networks is carried out using the PyTorch deep learning framework, with the added advantage of NVIDIA's automatic mixed-precision training strategy for expedited computations. Augmentation techniques include horizontal and vertical flips, as well as rotations in increments of $90 \times k$ degrees, where k takes values of 1, 2, and 3. In terms of data preprocessing, we crop images to a fixed size of (896, 896) using a sliding window approach with a stride of 512 pixels. For our model, we set the number of channels of FPN to 256, and the dimension of shared manifold in F-S module to 256. For our Foreground Saliency Guided Loss, set annealing_step to 10k and decay_factor to 0.9.

4.1.3 Evaluation metric

The performance of our networks is rigorously evaluated using three commonly accepted metrics: mean F1 score (mF1), mean intersection over union (mIoU), and overall accuracy (OA). OA signifies the proportion of correctly classified pixels in relation to the total pixel count. The mIoU is calculated as:

$$mIoU = \frac{TP}{FP + FN + TP}, \quad (7)$$

where TP represents true positives, FP stands for false positives, and FN denotes false negatives. The F1 score, a harmonic mean of precision P) and recall R), is expressed as:

$$F_1 = \left(1 + \beta^2\right) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}, \beta = 1, \quad (8)$$

$$P = \frac{TP}{TP + FP}, \quad (9)$$

$$R = \frac{TP}{FP + FN}, \quad (10)$$

In these equations, P stands for precision, R represents recall, TP is the count of true positives, FP corresponds to false positives, and FN indicates false negatives. The metrics collectively provide a comprehensive assessment of our segmentation models' performance.

4.2 Comparison on the iSAID dataset

In this section, we comprehensively assess the efficacy of our framework through a comparative analysis against contemporary state-of-the-art methods. We utilize the iSAID dataset as the testing ground for this evaluation, with the results compiled in Table 2. The benchmarked methodologies encompass a diverse selection, including PSPNet (Zhao et al., 2017b), Semantic FPN (Liu et al., 2019), FarSeg (Zheng et al., 2020), FactSeg (Ma et al., 2022), SegFormer (Xie et al., 2021), RSSFormer (Xu et al., 2023), and PFNet (Li et al., 2021). To ensure fairness, we meticulously standardized the settings across all methods. This standardization process ensures uniformity and impartiality in our evaluation. The implementation of these networks is carried out using the PyTorch deep learning framework, with the added advantage of NVIDIA's automatic mixed-precision training strategy for expedited computations.

Our thorough evaluation on the iSAID validation set, presented in Table 2, emphatically underscores the remarkable performance of our proposed framework in the realm of geospatial object segmentation. Notably, both Semantic FPN (Liu et al., 2019) and PSPNet (Zhao et al., 2017b) exhibit relatively lackluster results, primarily

1 http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

TABLE 2 Comparative Results on the iSAID dataset with State-of-the-Art Methods. The abbreviations for each category in the iSAID dataset respectively represent plane, baseball diamond, bridge, ground track field, small vehicle, ship, tennis court, basketball court, storage tank, soccer ball field, roundabout, harbor, swimming pool, and helicopter. The metric used is mIoU (mean Intersection over Union), where the highest-performing results are highlighted in bold.

Method	IoU per category (%)															mIoU
	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	
PSPNet	83.8	76.8	31.2	53.4	47.3	58.2	63.9	86.5	56.7	68.3	67.4	52.6	53	44	33.3	60.9
Semantic FPN	81.2	71.5	33.8	52.2	45.4	60.1	63.5	87.1	57.8	61.5	60.2	59	51.5	46.6	31.2	60.1
FarSeg	82	77.7	36.7	56.7	46.3	60.6	65.4	86.4	62.1	61.8	72.5	71.4	53.9	51.2	35.8	63.7
FactSeg	84.1	78.3	36.3	54.6	49.5	62.6	68.3	88.9	64.8	56.8	73.5	69.4	55.7	51.4	42.7	64.7
SegFormer	83.4	79.3	35.9	53.8	48.8	61.7	68.1	87.9	63.7	55.7	73.1	68.7	55.1	50.9	42.4	62.8
RSSFormer	82.9	78.1	34.8	53.4	47.9	61.5	67.4	87.5	63.1	54.9	72.8	68.2	54.6	50.2	41.8	65.8
PFNet	84	79.8	36.9	54.2	49.3	62.4	68.3	88.5	64.5	56.4	73.3	69.4	55.3	51.4	42.5	66.9
Ours	**84.2**	**80**	**37.1**	**56.9**	**49.7**	**62.8**	**68.5**	**89.2**	**65.4**	**68.7**	**74.1**	**72.6**	**56.3**	**52.8**	**43.1**	**67.3**

TABLE 3 Quantitative comparison results (%) on the Vaihigen dataset. The abbreviations for each category in the Vaihigen dataset respectively represent impervious surfaces, buildings, low vegetation, trees, and cars. OA is overall accuracy (%). mF1 is mean F1 score (%). The best results are indicated in bold.

Method	OA	Imp. Surf	Building	Low veg	Tree	Car	mF1
PSPNet	88.0	90.2	93.8	80.9	86.7	81.7	86.7
DeepLabV3+	89.05	89.98	93.91	80.66	89.41	83.39	87.47
SegFormer	89.05	89.98	93.70	80.67	89.41	81.39	87.03
RSSFormer	90.84	93.71	96.86	81.31	91.77	89.20	90.57
FactSeg	90.4	92.8	96.7	80.9	91.4	88.7	90.1
PFNet	91.2	93.9	96.91	82.3	92.5	90.2	92.3
Ours	**91.7**	**94.12**	**97.01**	**82.9**	**92.7**	**90.67**	**92.83**

TABLE 4 The mIoU results of different methods on the vaihingen dataset. The best results are indicated in bold.

Method	mIoU
PSPNet	0.62
DeepLabV3+	0.65
FactSeg	0.59
SegFormer	0.68
Ours	**0.72**

attributed to their limited prowess in foreground modeling. The performance of FarSeg (Zheng et al., 2020) and SegFormer (Xie et al., 2021) also falls short of expectations, likely due to their constraints in effectively handling the intricate task of segmenting small objects densely scattered across the imagery.

In stark contrast, our innovative framework outshines these baselines, achieving an impressive mIoU (mean Intersection over Union) score of 67.3%. This accomplishment, in comparison to existing object semantic segmentation techniques, marks a significant leap in performance. Our approach not only excels in foreground modeling but also excels in the meticulous segmentation of small objects, further validating its suitability for tackling the formidable challenge of small object semantic segmentation in HSR imagery.

4.3 Comparison on the Vaihingen dataset

In this section, we present an evaluation aimed at affirming the superior performance of our method. To achieve this, we conduct a comprehensive comparative analysis using the Vaihingen dataset as our testing ground. Our chosen benchmarks encompass several state-of-the-art segmentation networks, including PSPNet (Zhao et al., 2017b), DeepLabV3+ (Chen et al., 2018b), SegFormer (Xie et al., 2021), RSSFormer (Xu et al., 2023), FactSeg (Ma et al., 2022), and PFNet (Li et al., 2021). To ensure a fair and meaningful comparison, we meticulously adhere to the experimental settings detailed in Section 4.1.

The results of these comparative experiments are meticulously summarized in Table 3. It is of paramount importance to note that our proposed framework consistently outperforms the competing

TABLE 5 Ablation study for each component of the proposed method on iSAID dataset. PFMs: PointFlow Modules, Disentangled learning includes Foreground Precedence Estimation and Small Object Edge Alignment, and loss: foreground saliency guided loss. The best results are indicated in bold.

PFMs	Foreground-scene relation module	Disentangled learning	Loss	Method	mIoU (%)
				Baseline	67.3
✓				Ours	70.1
✓	✓				71.04
✓	✓	✓			72.9
✓	✓	✓	✓		**76.8**

TABLE 6 Ablation study for the loss functions in the iSAID dataset. CE is the cross-entropy loss. FL means the focal loss. FL + CE indicates that the total loss function is a simple addition of focal loss and cross-entropy loss. Our loss is the foreground saliency guided loss. The best results are indicated in bold.

Variants	F1 (%)	mIoU (%)
ours with CE	76.9	77.3
ours with FL	77.8	78.1
ours with FL + CE	78.5	77.64
ours with our loss	**80.4**	**79.05**

methods across various performance metrics, including mean F1 score, overall accuracy, and category-specific accuracy. In addition, we also compared the mIoU indicators of different methods, as shown in Table 4.

Notably, our method showcases remarkable improvements over its counterparts. When contrasted with PSPNet (Zhao et al., 2017b) and SegFormer (Xie et al., 2021), our framework exhibits substantial increments in mean F1 score, boasting an impressive 6.13% and 5.8% improvement, respectively. These noteworthy enhancements underscore the efficacy of the innovative modules seamlessly integrated into our framework.

Furthermore, in comparison to state-of-the-art segmentation networks specifically tailored for aerial imagery, our proposed framework emerges as the undisputed champion, boasting the highest overall accuracy and mean F1 score. Of particular significance is our framework's exceptional proficiency in identifying scattered vehicles, a testament to its unparalleled capability to capture small objects effectively. This distinctive feature sets it apart from the competition, establishing our framework as the preferred choice for scenarios demanding exceptional small object segmentation prowess.

4.4 Ablation study

In this section, we delve into a series of meticulous ablation experiments designed to dissect the individual contributions of various components within our proposed network. These components encompass the PointFlow Modules (PFMs), the Foreground-Scene module, disentangled learning, and the foreground saliency-guided loss. To establish a performance baseline for these experiments, we use the vanilla ResNet-50 + Feature Pyramid Network (FPN). The primary evaluation metric utilized throughout this section is mIoU, and the assessments are conducted on the iSAID validation set.

1) Ablation Study on Overall Framework: Table 5 presents a comprehensive overview of the relative improvements achieved by each proposed module in comparison to the baseline. The baseline, characterized by the vanilla ResNet-50 + FPN, exhibits a suboptimal performance, yielding an mIoU score of merely 67.3%. This underscores its inherent limitations in effectively addressing the intricate task of small object semantic segmentation within HSR imagery. With the introduction of PFMs into the FPN architecture, we observe a notable 2.8% increase in mIoU, highlighting the substantial contribution of these modules. Subsequently, the incorporation of the foreground-scene module elevates the mIoU to 71.04%, surpassing the baseline by a considerable margin. Further enhancements in performance are achieved with the introduction of disentangled learning, resulting in an mIoU of 72.9%. Ultimately, the amalgamation of all three components, coupled with our specialized loss function, culminates in the highest mIoU of 76.8%. These experimental findings underscore the pivotal role played by PFMs, the foreground-scene module in enhancing foreground saliency, and the significance of disentangled learning in the context of object segmentation, particularly within imbalanced scenes.
2) Ablation Study on Loss: To ascertain the efficacy of our foreground saliency-guided loss, we conduct a thorough investigation comparing it with different loss functions. Table 6 showcases the results of this ablation study. Specifically, we compare our foreground saliency-guided loss with cross-entropy (CE) loss and focal loss (FL). Additionally, we explore various combinations, including our loss with CE, our loss with FL, and our loss with FL + CE. Notably, the weights for the latter combination are identical. The findings reveal that the exclusive utilization of either cross-entropy or focal loss results in lower F1 scores and mIoU. Although the combination of cross-entropy and focal loss leads to an improvement in segmentation performance, it still lags behind our foreground saliency guided loss. This divergence arises primarily because HSR remote sensing images often contain a significant proportion of background. The adoption of a standard loss function

TABLE 7 Ablation study of boundary flow field and feature maps in our SOEA.

	Original input	Learned boundary flow field	Feature before SOEA	Feature after SOEA
Color Fidelity	0.83	0.89	0.91	0.92
Feature Extraction Efficiency	0.76	0.82	0.84	0.88
Edge Alignment Accuracy	0.71	0.75	0.8	0.85

can lead to overfitting on simpler background examples during the later stages of training, thus impeding the final efficacy of the segmentation. In contrast, our foreground saliency guided loss introduces an adjustment factor that enhances learning on challenging background examples in the later stages of training, effectively elevating the segmentation network's overall performance. Consequently, our approach achieves a higher mIoU of 79.05%, highlighting its effectiveness in addressing the unique challenges presented by HSR imagery.

3) Ablation study on SOEA: We evaluated the effectiveness of SOEA using three common image processing metrics, as shown in Table 7. Edge Alignment Accuracy is used to measure the accuracy of the algorithm to align image edges. The ideal accuracy is close to 100%. Feature Extraction Efficiency is used to measure the efficiency of the algorithm in extracting image features, including calculation time and accuracy. The higher the efficiency, the closer the value is to 100%. Color Fidelity is a measure of color consistency before and after image processing. If the color stays the same, it is close to 100%. It can be seen from the experimental results that our SOEA can effectively improve the segmentation effect.

4.5 Discussion

4.5.1 Computational Cost

Our framework contains multiple modules, which intuitively require more computing power and time costs than other approaches. However, from the details of the framework design, we can see that most of these modules only apply the basic convolution operations, which we believe do not cost too much computational cost.

4.5.2 Applicability to domains other than semantic segmentation in aerial imagery

We propose a new semantic segmentation framework for aerial imagery. We believe that this framework is not only applicable to HSR remote sensing images, but also can be applied to other long-distance semantic segmentation tasks, such as long-distance semantic segmentation in autonomous driving. This will also be an interesting research direction in the future.

5 Conclusion

This article delves into two critical challenges within the realm of aerial imagery: 1) Mitigating Background Noise and Enhancing Foreground Saliency. 2) Accurate Segmentation in Dense Small Object Distributions. To surmount these challenges, we introduce an upgraded semantic segmentation framework founded on feature pyramids. This framework comprises three core components: PointFlow modules, the foreground-scene module, disentangled learning, and the foreground saliency guided loss. Specifically, we integrate PointFlow modules into the Feature Pyramid Network (FPN) architecture. These modules are designed to select representative points between adjacent feature pyramid levels, replacing the conventional methods of simple fusion or dense affinity. The foreground-scene (F-S) module plays a pivotal role in associating context relevant to the foreground, thereby enhancing the saliency of foreground features. Our disentangled learning explicitly models foreground objects and aligns edge features, contributing to more precise segmentation outcomes. Furthermore, we train our network using the foreground saliency guided loss, ensuring a balanced approach between foreground and background. The comprehensive set of experimental results, spanning the iSAID and Vaihingen datasets, demonstrates the efficacy of our proposed framework. Looking ahead, we anticipate broader applications in various domains, envisioning adaptation for in-orbit satellite challenges (Zhong et al., 2020).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

SH: Conceptualization, Methodology, Writing–original draft, Writing–review and editing. CJ: Methodology, Writing–original draft, Writing–review and editing. LS: Writing–original draft. XH: Writing–review and editing. MW: Writing–review and editing. GL: Conceptualization, Funding acquisition, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Zhejiang 'JIANBING' R&D Project (No. 2022C01055) and Zhejiang Provincial Department of Transport Technology Project (No. 2024011).

Acknowledgments

This work is an expanded version of our previous conference papers "A semantic segmentation framework for small objects segmentation in remote sensing images." In Proceedings - 2023 International Conference on Cloud Computing, Big Data Applications and Software Engineering, CBASE 2023. (to be published).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi:10.1109/TPAMI.2016.2644615

Bastani, F., He, S., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., et al. (2018). "Roadtracer: automatic extraction of road networks from aerial images," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (Computer Vision Foundation/ IEEE Computer Society), 4720–4728. doi:10.1109/CVPR.2018.00496

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2015). "Semantic image segmentation with deep convolutional nets and fully connected crfs," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015. Editors Y. Bengio and Y LeCun. Conference Track Proceedings.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi:10.1109/TPAMI.2017.2699184

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII. Editors V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Springer), 833–851. 11211 of Lecture Notes in Computer Science. doi:10.1007/978-3-030-01234-2_49

Chen, Y., Jiang, H., Li, C., Jia, X., and Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote. Sens.* 54, 6232–6251. doi:10.1109/TGRS.2016.2584107

Dickenson, M., and Gueguen, L. (2018). "Rotated rectangles for symbolized building footprint extraction," in 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018 (Computer Vision Foundation/ IEEE Computer Society), 225–228. doi:10.1109/CVPRW.2018.00039

Ferjaoui, R., Cherni, M. A., Abidi, F., and Zidi, A. (2022). "Deep residual learning based on resnet50 for COVID-19 recognition in lung CT images," in 8th International Conference on Control, Decision and Information Technologies, CoDIT 2022, Istanbul, Turkey, May 17-20, 2022 (IEEE), 407–412. doi:10.1109/CoDIT55151.2022.9804094

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). "Dual attention network for scene segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (Computer Vision Foundation/ IEEE), 3146–3154. doi:10.1109/CVPR.2019.00326

Gupta, S., Girshick, R. B., Arbeláez, P. A., and Malik, J. (2014). "Learning rich features from RGB-D images for object detection and segmentation," in Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII. Editors D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Springer), 345–360. 8695 of Lecture Notes in Computer Science. doi:10.1007/978-3-319-10584-0_23

He, H., Wang, S., Zhao, Q., Lv, Z., and Sun, D. (2021). "Building extraction based on u-net and conditional random fields," in 2021 6th International Conference on Image, Vision and Computing (ICIVC 2021), Qingdao, China, July 23–25, 2021, 273–277.

He, R., Xiao, X., Kang, Y., Zhao, H., and Shao, W. (2022). "Heterogeneous pointer network for travelling officer problem," in International Joint Conference on Neural Networks, IJCNN 2022, Padua, Italy, July 18-23, 2022 (IEEE), 1–8. doi:10.1109/IJCNN55064.2022.9892069

He, X., Trigila, C., Ariño-Estrada, G., and Roncali, E. (2023). Potential of depth-of-interaction-based detection time correction in cherenkov emitter crystals for tof-pet. *IEEE Trans. Radiat. Plasma Med. Sci.* 7, 233–240. doi:10.1109/trpms.2022.3226950

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., et al. (2017). "beta-vae: learning basic visual concepts with a constrained variational framework," in 5th International Conference on Learning Representations, Toulon, France, April 24-26, 2017 (ICLR 2017). *Conference Track Proceedings* (OpenReview.net).

Hou, J., Guo, Z., Wu, Y., Diao, W., and Xu, T. (2022). Bsnet: dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–22. doi:10.1109/TGRS.2022.3176028

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (Computer Vision Foundation/ IEEE Computer Society), 7132–7141. doi:10.1109/CVPR.2018.00745

Jin, C., He, S., Hou, Y., He, X., Wang, M., and Liu, G. (2023). "A semantic segmentation framework for small objects segmentation in remote sensing images," in Proceedings - 2023 International Conference on Cloud Computing, Big Data Applications and Software Engineering (CBASE 2023), Chengdu, China, November 3–5, 2023.

Kang, Y., Rahaman, M. S., Ren, Y., Sanderson, M., White, R. W., and Salim, F. D. (2022). App usage on-the-move: context- and commute-aware next app prediction. *Pervasive Mob. Comput.* 87, 101704. doi:10.1016/j.pmcj.2022.101704

Kemker, R., Salvaggio, C., and Kanan, C. (2017). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogrammetry Remote Sens.* 145, 60–77. doi:10.1016/j.isprsjprs.2018.04.014

Li, X., He, H., Li, X., Li, D., Cheng, G., Shi, J., et al. (2021). "Pointflow: flowing semantics through points for aerial image segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021 (Computer Vision Foundation/ IEEE), 4217–4226. doi:10.1109/CVPR46437.2021.00420

Liang, J., Homayounfar, N., Ma, W., Wang, S., and Urtasun, R. (2019). "Convolutional recurrent network for road boundary extraction," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (Computer Vision Foundation/ IEEE), 9512–9521. doi:10.1109/CVPR.2019.00974

Lin, G., Milan, A., Shen, C., and Reid, I. D. (2017a). "Refinenet: multi-path refinement networks for high-resolution semantic segmentation," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 (IEEE Computer Society), 5168–5177. doi:10.1109/CVPR.2017.549

Lin, T., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017b). "Focal loss for dense object detection," in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017 (IEEE Computer Society), 2999–3007. doi:10.1109/ICCV.2017.324

Liu, H., Peng, C., Yu, C., Wang, J., Liu, X., Yu, G., et al. (2019). "An end-to-end network for panoptic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (Computer Vision Foundation/ IEEE), 6172–6181. doi:10.1109/CVPR.2019.00633

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in IEEE Conference on Computer Vision and Pattern

Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015 (IEEE Computer Society), 3431–3440. doi:10.1109/CVPR.2015.7298965

Loshchilov, I., and Hutter, F. (2017). "SGDR: stochastic gradient descent with warm restarts," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (OpenReview.net).

Ma, A., Wang, J., Zhong, Y., and Zheng, Z. (2022). Factseg: foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–16. doi:10.1109/TGRS.2021.3097148

Mou, L., Hua, Y., and Zhu, X. X. (2019). "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (Computer Vision Foundation/ IEEE), 12416–12425. doi:10.1109/CVPR.2019.01270

Niu, R., Sun, X., Tian, Y., Diao, W., Feng, Y., and Fu, K. (2022). Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–18. doi:10.1109/TGRS.2021.3121471

Onim, M. S. H., Ehtesham, A. R., Anbar, A., Islam, A., and Rahman, A. M. (2020). "Lulc classification by semantic segmentation of satellite images using fastfcn," in 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT 2020), United International University, Bangladesh, November 28–29, 471–475.

Pinheiro, P. H. O., and Collobert, R. (2014). "Recurrent convolutional neural networks for scene labeling," in Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR.org), 82–90. 32 of JMLR Workshop and Conference Proceedings.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Editors N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi (Springer), 234–241. 9351 of Lecture Notes in Computer Science. doi:10.1007/978-3-319-24574-4_28

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi:10.1007/s11263-015-0816-y

Shao, W., Jin, Z., Wang, S., Kang, Y., Xiao, X., Menouar, H., et al. (2022a). "Long-term spatio-temporal forecasting via dynamic multiple-graph attention," in International Joint Conference on Artificial Intelligence.

Shao, W., Prabowo, A., Zhao, S., Koniusz, P., and Salim, F. D. (2022b). Predicting flight delay with spatio-temporal trajectory convolutional network and airport situational awareness map. *Neurocomputing* 472, 280–293. doi:10.1016/j.neucom.2021.04.136

Shao, W., Zhao, S., Zhang, Z., Wang, S., Rahaman, M. S., Song, A., et al. (2021). "FADACS: a few-shot adversarial domain adaptation architecture for context-aware parking availability sensing," in 19th IEEE International Conference on Pervasive Computing and Communications, PerCom 2021, Kassel, Germany, March 22-26, 2021 (IEEE), 1–10. doi:10.1109/PERCOM50583.2021.9439123

Volpi, M., and Ferrari, V. (2015). "Semantic segmentation of urban scenes by learning local class interactions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2015, Boston, MA, USA, June 7-12, 2015 (IEEE Computer Society), 1–9. doi:10.1109/CVPRW.2015.7301377

Wang, C., Bai, X., Zhou, L., and Zhou, J. (2019). "Hyperspectral image classification based on non-local neural networks," in 2019 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2019, Yokohama, Japan, July 28 - August 2, 2019 (IEEE), 584–587. doi:10.1109/IGARSS.2019.8897931

Wang, J., Zhong, Y., Zheng, Z., Ma, A., and Zhang, L. (2021). Rsnet: the search for remote sensing deep neural networks in recognition tasks. *IEEE Trans. Geosci. Remote. Sens.* 59, 2520–2534. doi:10.1109/TGRS.2020.3001401

Wang, X., Girshick, R. B., Gupta, A., and He, K. (2018). "Non-local neural networks," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (Computer Vision Foundation/ IEEE Computer Society), 7794–7803. doi:10.1109/CVPR.2018.00813

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J. M., and Luo, P. (2021). "Segformer: simple and efficient design for semantic segmentation with transformers," in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021. Editors M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, 12077–12090. virtual.

Xu, R., Wang, C., Zhang, J., Xu, S., Meng, W., and Zhang, X. (2023). Rssformer: foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Trans. Image Process.* 32, 1052–1064. doi:10.1109/TIP.2023.3238648

Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., et al. (2020). "Disentangled non-local neural networks," in Proceedings, Part XV Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020. Editors A. Vedaldi, H. Bischof, T. Brox, and J. Frahm (Springer), 191–207. 12360 of Lecture Notes in Computer Science. doi:10.1007/978-3-030-58555-6_12

Yuan, Y., Xie, J., Chen, X., and Wang, J. (2020). "Segfix: model-agnostic boundary refinement for segmentation," in Proceedings, Part XII Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020. Editors A. Vedaldi, H. Bischof, T. Brox, and J. Frahm (Springer), 489–506. 12357 of Lecture Notes in Computer Science. doi:10.1007/978-3-030-58610-2_29

Zamir, S. W., Arora, A., Gupta, A., Khan, S. H., Sun, G., Khan, F. S., et al. (2019). "isaid: a large-scale dataset for instance segmentation in aerial images," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019 (Computer Vision Foundation/ IEEE), 28–37.

Zhang, D., Zhang, H., Tang, J., Wang, M., Hua, X., and Sun, Q. (2020). "Feature pyramid transformer," in Proceedings, Part XXVIII Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020. Editors A. Vedaldi, H. Bischof, T. Brox, and J. Frahm (Springer), 323–339. 12373 of *Lecture Notes in Computer Science*. doi:10.1007/978-3-030-58604-1_20

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017a). "Pyramid scene parsing network," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 (IEEE Computer Society), 6230–6239. doi:10.1109/CVPR.2017.660

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017b). "Pyramid scene parsing network," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 (IEEE Computer Society), 6230–6239. doi:10.1109/CVPR.2017.660

Zheng, Z., Zhong, Y., Wang, J., and Ma, A. (2020). "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020 (Computer Vision Foundation/ IEEE), 4095–4104. doi:10.1109/CVPR42600.2020.00415

Zhong, Y., Li, W., Wang, X., Jin, S., and Zhang, L. (2020). Satellite-ground integrated destriping network: a new perspective for eo-1 hyperion and Chinese hyperspectral satellite datasets. *Remote Sens. Environ.* 237, 111416. doi:10.1016/j.rse.2019.111416