FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery

Ailong Ma, *Member, IEEE*, Junjue Wang, *Student Member, IEEE*, Yanfei Zhong, *Senior Member, IEEE*, and Zhuo Zheng, *Graduate Student Member, IEEE*

***Abstract*— The small object semantic segmentation task is aimed at automatically extracting key objects from high-resolution remote sensing (HRS) imagery. Compared with the large-scale coverage areas for remote sensing imagery, the key objects, such as cars and ships, in HRS imagery often contain only a few pixels. In this article, to tackle this problem, the foreground activation (FA)-driven small object semantic segmentation (FactSeg) framework is proposed from perspectives of structure and optimization. *In the structure design*, FA object representation is proposed to enhance the awareness of the weak features in small objects. The FA object representation framework is made up of a dual-branch decoder and collaborative probability (CP) loss. In the dual-branch decoder, the FA branch is designed to activate the small object features (activation) and suppress the large-scale background, and the semantic refinement (SR) branch is designed to further distinguish small objects (refinement). The CP loss is proposed to effectively combine the activation and refinement outputs of the decoder under the CP hypothesis. During the collaboration, the weak features of the small objects are enhanced with the activation output, and the refined output can be viewed as the refinement of the binary outputs. *In the optimization stage*, small object mining (SOM)-based network optimization is applied to automatically select effective samples and refine the direction of the optimization while addressing the imbalanced sample problem between the small objects and the large-scale background. The experimental results obtained with two benchmark HRS imagery segmentation datasets demonstrate that the proposed framework outperforms the state-of-the-art semantic segmentation methods and achieves a good tradeoff between accuracy and efficiency. Code will be available at: http://rsidea.whu.edu.cn/FactSeg.htm**

***Index Terms*— Deep learning, high-resolution remote sensing (HRS) imagery, semantic segmentation, small objects.**

I. Introduction

LARGE amounts of high-resolution remote sensing (HRS) images are now being acquired from both airborne and spaceborne platforms, providing base data for mapping and observation. Compared with low-resolution images, HRS images contain more details, which means that small objects are visible. However, this also brings challenges as the key objects can be very small, and they often account for a very small proportion of all the pixels in large-scale regions [1], [2]. In the past decades, extensive efforts have been made in the development of various feature representation approaches for HRS imagery mapping and updating [3].

Semantic segmentation, which is also called pixel-level classification, is an important intermediate step between raw images and a vector map layer. Small object semantic segmentation is a special case of semantic segmentation that only focuses on key object recognition [1], [2], [4]. These key objects are also called foreground objects, compared with the background, which is not of interest. The standard small object semantic segmentation problem is to first extract the foreground objects from the background and then predict the conditional probabilities within the foreground objects at the pixel level.

Most of the previous studies have viewed small object semantic segmentation as a traditional semantic segmentation problem. However, the conventional semantic segmentation methods rely purely upon handcrafted spectral and spatial features [5]–[7], with which it is not easy to fit a complex distribution. Deep learning has now become the state of the art for a whole range of image analysis tasks, and many classic convolutional neural networks (CNNs) have been successfully applied in HRS imagery semantic segmentation [8]–[10]. Compared with the traditional handcrafted features, deep learning methods are data-driven methods, in which the representative features are learned end-to-end, hierarchically. Although the performance of HRS imagery segmentation has been improved with the help of deep learning methods, most of the previous works have directly employed advanced CNNs with minor modifications and are poor in recognizing small objects. The long-distance observation brings unique characteristics to HRS imagery semantic segmentation. Compared with natural images, HRS imagery often covers large-scale regions, while the foreground objects often contain only a few pixels. For example, each image in the Instance Segmentation in Aerial Images (iSAID) dataset [11] covers several square kilometers, while key objects, such as vehicles, only cover about 10 m^2. In Fig. 1, the HRS imagery is taken from the iSAID dataset, where the background makes up 98.66% of the

Manuscript received January 10, 2021; revised May 14, 2021; accepted July 1, 2021. Date of publication July 27, 2021; date of current version January 17, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0504202, in part by the National Natural Science Foundation of China under Grant 41771385 and Grant 41801267, and in part by the China Postdoctoral Science Foundation under Grant 2017M622522. *(Corresponding authors: Junjue Wang; Yanfei Zhong.)*

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Hubei Provincial Engineering Research Center of Natural Resources Remote Sensing Monitoring, Wuhan University, Wuhan 430079, China (e-mail: maailong007@whu.edu.cn; kingdrone@whu.edu.cn; zhongyanfei@whu.edu.cn; zhengzhuo@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3097148

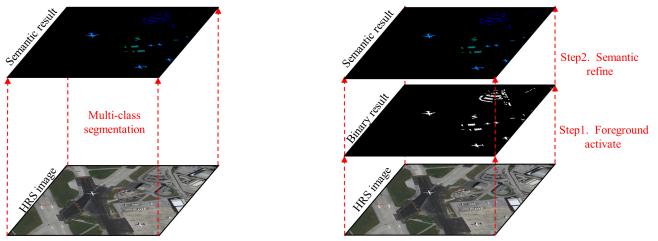


Fig. 1. Comparison between (a) traditional multiclass semantic segmentation workflow and (b) proposed FactSeg workflow. The example is of large-scale and densely annotated instance segmentation using an aerial image dataset (iSAID).

pixels and the key objects amount to only 1.34% of the pixels. The small objects and large-scale background result in two specific difficulties in the HRS segmentation task, as follows.

1) *Weak Features:* The foreground objects often contain only a few pixels, leaving minimal appearance cues to exploit [12]. The traditional CNNs employ uniform sliding window sampling, without emphasis, and it is, thus, difficult for CNNs to extract enough distinct features for small object semantic segmentation.
2) *Imbalanced Samples:* The number of background samples of no interest significantly exceeds the number of foreground samples. Large-scale background regions contain a lot of homogeneous samples, which are easy to distinguish. In contrast, the rare foreground regions contain scarce and complex samples, which are hard to classify. This imbalanced sample problem can mislead the direction of the optimization during training, making the CNN focus on easy samples rather than hard samples [1].

As shown in Fig. 1(a), the traditional HRS imagery object semantic segmentation task is viewed as a traditional multiclass semantic segmentation task with an end-to-end fully CNN, which is an approach that does not consider the above difficulties properly. In this article, the FactSeg framework is proposed for HRS imagery, based on a novel workflow, as shown in Fig. 1(b). This framework addresses the small object semantic segmentation task from the new perspectives of representation and optimization. The foreground activation (FA) object representation framework is made up of a dual-branch decoder and CP loss. This effectively suppresses background distractions while also enhancing the small object features. In the model optimization stage, the small object mining (SOM) strategy is designed for automatically selecting effective samples to address the imbalanced sample problem. This approach provides an effective option for the HRS small object semantic segmentation task. The contributions of this article can be summarized as follows.

1) *Foreground Activation Object Representation:* Differing from the approach taken in the previous works, the HRS imagery semantic segmentation task is decomposed into two subtasks: 1) binary-class activation, i.e., class-agnostic small object activation and 2) multiclass refinement, i.e., multiclass feature refinement under the guidance of the binary-class activation. This FA strategy decomposes the segmentation task, with binary-class segmentation inserted as an intermediate step, which is realized with the following two modules.
   a) *Dual-Branch Decoder:* In order to obtain the class-agnostic activation and refinement semantic features, an effective dual-branch decoder is proposed, which is made up of an FA branch and a semantic refinement (SR) branch. The FA branch is designed for activating the small objects (activation) while also suppressing the large-scale background, and the SR branch is designed for fine-grained recognition of these object features (refinement). Each branch consists of a variant of a feature pyramid network (FPN), with the aim being to obtain the activation and refinement features.
   b) *Collaborative Probability (CP) Loss:* Because the dual-branch decoder has both activation and refinement semantic outputs, the CP loss is proposed for effective fusion of the outputs. It is assumed that the outputs of the FA branch generate the binary-class probabilities, which can be used to activate the small objects at the pixel level. The activation outputs are used for class-agnostic small object guidance, and the refinement outputs are used for refining the activation results. Moreover, the CP hypothesis is proposed for support, and the CP loss function is also introduced. Based on the FA strategy, the CP loss function can fuse the two branches of outputs at the probability level, thereby improving the optimization efficiency.

   The FA object representation framework is based on the coarse-to-fine approach and is, thus, similar to YOLO [13]. However, as YOLO is an object detection framework and FactSeg is a semantic segmentation framework, the two methods differ in both the task and implementation. Compared with the previous traditional semantic segmentation workflows, the FA object representation framework not only enhances the ability to extract key features but also relieves the burden on the SR branch by reducing the distractions from the large-scale background.
2) *Small Object Mining-Based Network Optimization:* The small objects in HRS imagery often contain only a few pixels compared with the large-scale background. The imbalanced samples can, thus, mislead the optimization direction of the CNN, which is especially the case for the previous deep learning-based methods in the HRS imagery small object semantic segmentation task. In order to resolve this problem, SOM-based network optimization is adopted in the proposed framework. During the training, the pixel-level samples are sorted by their losses. The samples with high losses are considered to be effective and can be selected for backward propagation. The experiments undertaken in this study showed that the most effective samples lie in the small objects. This SOM strategy refines the direction of the model optimization.

The proposed FactSeg method was implemented on two large-scale HRS imagery semantic segmentation datasets. Both comparative experiments and ablation experiments were conducted under the same experimental settings. FactSeg and the reference networks were tested with regard to both accuracy and efficiency. The experimental results indicate that FactSeg outperforms the state-of-the-art semantic segmentation methods.

The rest of this article is organized as follows. Section II introduces the related work on semantic segmentation and imbalanced samples. Section III describes the general framework and key components of FactSeg. In Section IV, we describe the experiments designed to thoroughly evaluate the effectiveness of each proposed module, and we compare FactSeg with the other reference semantic segmentation CNNs. In Section V, the potential of FactSeg in HRS imagery applications is demonstrated using some typical large-scale complex scenes. Finally, Section VI provides our conclusions and future research directions.

II. Related Work

A. Semantic Segmentation in Remote Sensing

The important semantic segmentation works are introduced in the following. Extensive efforts have gone into the semantic segmentation of satellite and aerial images in the past few decades. The conventional methods are generally aimed at designing more robust features. Typical input features include spectral and spatial features, such as raw pixel intensities, object indices, and different statistics or filter responses that describe the local image texture [14]–[17]. For example, Wang and Ming [18] integrated spectral and shape features into the road extraction process; Huang *et al.* [19] proposed a postprocessing framework that describes the characteristics of buildings by simultaneously considering the spectral, geometrical, and contextual information, which alleviates the number of false alarms, to a great extent. An alternative approach to improving the segmentation performance is choosing classifiers that include efficient feature selection (e.g., boosting, decision tree, and random forest classifiers). After the redundant features are computed, these classifiers automatically select the optimal subset to reduce the relevant information [20]. The conditional random field (CRF) model is an undirected graphical method that can model global contextual information. Wang *et al.* [6] used Gabor texture features within a CRF framework to realize urban forest cover mapping. To make full use of the spatial–contextual information and topological information, Huang *et al.* [7] proposed an object-based CRF model for road extraction. However, these conventional methods with handcrafted features often have difficulty in describing complex patterns and distributions. Deep learning is now a hot topic in computer vision and pattern recognition, where the representative and semantic features are learned hierarchically and automatically [21]. The fully convolutional network (FCN) and its extensions have been widely used in remote sensing to address semantic segmentation tasks [22]–[24]. In the meantime, large-scale pixel-level labeled HRS imagery datasets [25] have been developed, promoting the development of deep learning in the remote sensing domain. Based on a large-scale dataset, Wang *et al.* [9] proposed a hierarchical neural network search framework to automatically design remote sensing recognition architectures. These semantic segmentation CNNs are, however, directly employed or modified from advanced CNN architectures, without considering the weak features within the small objects in HRS imagery. As a result, these methods are not suitable for the small object semantic segmentation task in HRS imagery.

Compared with natural images, the use of HRS images results in smaller object recognition problems. For example, Wang *et al.* [26] proposed a synthetic aperture radar dataset for the ship detection task, which is a good example of small object recognition under the scenario of a complex background. For video satellite processing, LaLonde *et al.* [27] proposed ClusterNet for detecting small objects in large scenes by exploiting the spatiotemporal information. The robust detection of small objects has also been widely studied with infrared images [28], [29].

For HRS images, Dong *et al.* [30] addressed the small object detection task with the Sig-NMS module, to decrease the possibility of missing small targets. A unified and self-reinforced network was also proposed by Pang *et al.* [31] to inhibit false positives in tiny object detection. Small object semantic segmentation is also now being studied. For example, Hamaguchi *et al.* [2] effectively adopted dilated convolutions to keep the high spatial resolution details for small objects. Kampffmeyer *et al.* [1] utilized a support vector machine to refine the uncertain pixels of small objects, thereby improving the segmentation accuracy. Zheng *et al.* [32] designed a foreground-aware relation network and modeled the relationship between the foreground and the geospatial scene with a 1-D scene embedding vector, thereby improving the discrimination of foreground features. However, this method only associates the foreground with the scene implicitly via nonlinear transformation. Dong *et al.* [33] proposed DenseU-Net for small object semantic segmentation in urban remote sensing images, where DenseU-Net connects the CNN features through cascade operations and fuses the detail features in the shallow layers, as well as the abstract semantic features in the deep layers. The FactSeg method proposed in this article utilizes a dual-branch decoder with direct supervision of the CP loss to explicitly formulate the FA object representation.

B. Imbalanced Samples

The methods of addressing the imbalanced sample problem in CNNs fall into four main approaches: 1) oversampling; 2) undersampling; 3) two-phase training; and 4) thresholding [34]. In the segmentation task, the effective approaches to prevent the difficulties encountered with imbalanced samples are mainly based on the design of the loss function. For example, Rajpurkar *et al.* [35] employed inverse weighted cross-entropy (CE) loss to balance the proportion of each category; Zhou *et al.* [36] applied dice loss in road extraction by increasing the weights of the key road regions. Nevertheless, the dice loss can only be used in binary segmentation tasks. The online hard example mining (OHEM) strategy has

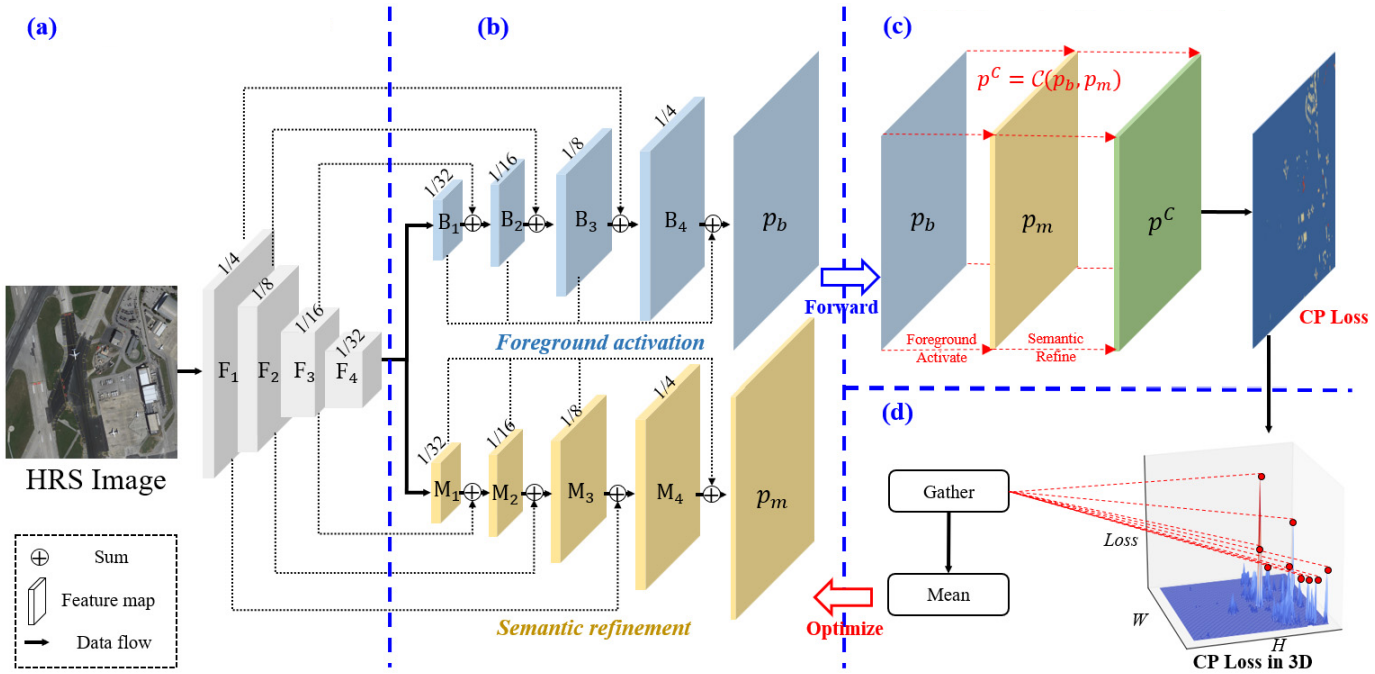


Fig. 2. Overview of the proposed FactSeg framework: (a) deep residual encoder for deep semantic feature extraction, (b) dual-branch decoder, (c) CP loss, and (d) small object mining. The CP loss is represented in three dimensions for a simple demonstration.

been adopted to dynamically select hard samples for effective optimization. The SOM strategy is a modification of the OHEM strategy, which addresses the imbalanced samples by selecting the hard examples in the object detection task [37]. Yu *et al.* [38] also extended OHEM to a more general method for real-time detectors. This technology has been successfully applied in the remote sensing-based animal detection task [39]. Many object detection methods now use OHEM to select effective object proposal regions of interest (ROIs). For example, Wu *et al.* [40] first proposed the online bootstrapping (Online Bs.) of hard training pixels for semantic segmentation, inspired by the OHEM strategy. The Online Bs. method utilizes a prediction probability threshold to filter the hard samples automatically in each minibatch. However, it cannot fix the number of batch samples during training. In the proposed approach, inspired by these works, the SOM strategy is introduced into the HRS segmentation task. Differing from threshold filtering [40], the SOM strategy guarantees a fixed number of training samples in each minibatch, and the hard samples are selected with the sorted losses. A stable number of training samples is more conducive to model optimization in the HRS segmentation scenario where the foreground and background samples are extremely imbalanced.

III. Foreground Activation-Driven Small Object Semantic Segmentation Framework

An overview of the proposed FactSeg framework is shown in Fig. 2. The proposed framework is made up of four main parts:

A. Deep Residual Encoder

ResNet [41] has a powerful feature extraction capability due to its elaborate residual modules. Many studies have demonstrated the good performance of a pretrained ResNet backbone in segmentation tasks [42]–[44]. In the proposed framework, ResNet is utilized, without fully connected layers, as an encoder. As shown in Fig. 2, given an input image, the feature map is successively decreased in a bottom-up pathway [45]. In deep residual encoders, there are often many residual blocks producing output maps of the same size grouped by the network stages. The last stage is chosen here since the deepest layer of each stage should have the strongest features. The outputs of these last features in each stage are denoted as $\{F_1, F_2, F_3, F_4\}$, and they have strides of $\{4, 8, 16, 32\}$ pixels with respect to the input image. ResNet50 was chosen in this study for its accuracy and efficiency. Moreover, the FactSeg framework is indeed flexible as other backbones can be used. If faced with VGG, the multiscale features in different stages can also be extracted.

B. Dual-Branch Decoder

The dual-branch decoder includes an FA branch and an SR branch, as shown in Fig. 2, with each branch consisting of a variant of an FPN.

In order to enhance the multiscale feature fusion ability of the model and leverage the ResNet encoder's pyramidal feature hierarchically, an FPN is adopted in the proposed method. The FPN was first proposed for the object detection task by Liu *et al.* [45], who aimed to capture strong semantics at all scales. This excellent multiscale feature processing module has also been successfully applied in instance segmentation [46] and panoptic segmentation [47]. Similar to the original FPN, the FA branch utilizes a top-down pathway and skip connections, yielding the pyramidal features $\{B_1, B_2, B_3, B_4\}$. Specifically, the deepest-layer features F_4 in the encoder are chosen as the inputs. The channel dimension of F_4 is first

reduced (256 by default) with a 1×1 convolution followed by a 3×3 convolution for refining the information, obtaining the B_1 features. The B_1 features are then upsampled by a factor of two for a higher resolution. The transformed version of the 1/16 resolution features F_3 from the encoder is then gathered to sum with the upsampled features, followed by a 3×3 convolution, obtaining B_2. Generally speaking, the deepest-layer features F_4 are progressively upsampled while being elementwise added with transformed versions of the higher resolution features from the encoder. Finally, the FA features $\{B_1, B_2, B_3, B_4\}$ in the FA branch are acquired. The same structure is used in the SR branch, yielding the fine stage features $\{M_1, M_2, M_3, M_4\}$. The procedures can be formulated as follows:

$$B_{i+1} = \text{Upsample}_{\times 2}(\Gamma(B_i)) + \zeta(F_{4-i}), \quad i = 1, 2, 3 \tag{1}$$

$$M_{i+1} = \text{Upsample}_{\times 2}(\Gamma(M_i)) + \zeta(F_{4-i}), \quad i = 1, 2, 3 \tag{2}$$

where ζ denotes the skip connection, utilizing high spatial resolution details from the shallow layers, and Γ denotes the transformed process of the features in the decoder.

After obtaining features from the deep residual encoder, a simple fusion module is designed for multiscale feature fusion. As illustrated in Fig. 2, for the FA branch, the deepest-level features B_1 are applied with three times upsampling stages to yield features at 1/4 scale, where each upsampling stage includes a 3×3 convolution, group normalization, a rectified linear unit (ReLU) activation, and $2\times$ bilinear upsampling [47]. Meanwhile, the other features B_2, B_3, and B_4 are, respectively, applied with fewer upsampling stages at 1/4 scale. All the resulting feature maps are then elementwise summed, followed with a 1×1 convolution, $4\times$ bilinear upsampling, and a softmax classifier for the per-pixel class labels at the original image resolution. The FA and SR branches have symmetric structures, as shown in Fig. 2.

Based on the FPN's powerful multiscale feature capture and detail retention capabilities, we adopt symmetric structures in a very concise and efficient form. There are three main differences between the proposed dual-branch decoder and the original FPN.

1) *Different Tasks and Implementations:* The FPN is designed for the object detection task, generating multiscale anchors, while the proposed decoder is aimed at fusing multiscale features at a pixelwise level for semantic segmentation. Compared with the original FPN, the proposed decoder utilizes symmetric FPNs followed by the corresponding multiscale fusion modules, which is a dual-branch structure.
2) *Different Motivations:* The dual-branch decoder is designed for decomposing the features. The FA branch extracts features that distinguish the foreground objects from the background. The SR branch extracts features that focus on multiclass classification. When combined with the CP loss, the learnable weights in the dual-branch decoder are trained based on different emphases.

Differing from the traditional semantic segmentation networks, the proposed model has an additional FA branch to activate the class-agnostic foreground objects. By decomposing the segmentation task, this FA design not only enhances the capability of the foreground object and background separation but also reduces the burden on the SR branch, allowing it to pay more attention to small object feature interpretation.

TABLE I
DEFINITIONS OF THE OUTPUTS FROM THE DUAL-BRANCH DECODER

	Background	Foreground
FA branch	p_b	$1 - p_b$
SR branch	p_{m_0}	$p_{m_{i>0}}$

C. Collaborative Probability Loss

In order to better combine the outputs of the dual-branch decoder, CP loss is proposed. The CP loss can effectively fuse the two output branches in the same probability framework, which enables the model to fully utilize the information reasonably and effectively during the training and inference. In the traditional multiclass semantic segmentation, the CE loss is used, as follows:

$$\text{CE}(p, y) = -\sum_{i=0}^{N-1} y_i log(p_i) \tag{3}$$

where $y_i \in \{0, 1\}$ specifies the ground-truth class and $p_i \in [0, 1]$ is the model's estimated probability for class i with label $y_i = 1$. N denotes the total class number. For the traditional model, p_i is obtained from the last output normalized with the softmax function. However, there are two kinds of outputs from the dual-branch decoder. In previous research [46], [48], multitask loss was used to make the branches function differently. By optimizing toward different goals, each branch focuses on decoding different outputs. Differing from the multitask loss design, the CP framework is proposed to explicitly model the dual-branch outputs at the probability level so that they can, respectively, focus on binary-class and multiclass classification.

Inspired by human perception, in which the image is first scanned and the key objects are then focused on [49], the FA branch is utilized to activate class-agnostic foreground objects, and the SR branch focuses on the foreground object classification. In the implementation, the binary probability from the FA branch activates pixels that may form objects. With the guidance of the FA branch output, the SR branch pays more attention to the foreground object classification.

Specifically, the two outputs of the dual-branch decoder are fused in a pixelwise manner at the probability level. As is shown in Table I, we denote the output probability of the FA branch as p_b and that of the SR branch as p_{m_i} (where $0 \leq i < N$). p_b represents the probability of the background, and $1 - p_b$ represents the probability of the foreground.

In order to avoid conflicts, the outputs of the two branches are combined under the CP assumption.

1) The distributions of the two branch outputs are independent. Each branch processes the shared outputs in the deep residual encoder using nonlinear representation.

2) Only the same prediction outputs (background or foreground) for the two branches can yield the final CP. The contradictory outputs $1 - p_b$ and p_{m_0}, and p_b and $p_{m_{i>0}}$ are not fused.

According to the above assumption, the CP p_i^C is defined as follows:

$$p_i^C = \mathcal{C}(p_b, p_m) = \begin{cases} \frac{1}{Z} p_b p_{m_i}, & i = 0 \\ \frac{1}{Z}(1 - p_b) p_{m_i}, & 0 < i < N \end{cases} \tag{4}$$

where $\mathcal{C}$ denotes the fusion process and Z denotes the normalizing factor, which is defined as follows:

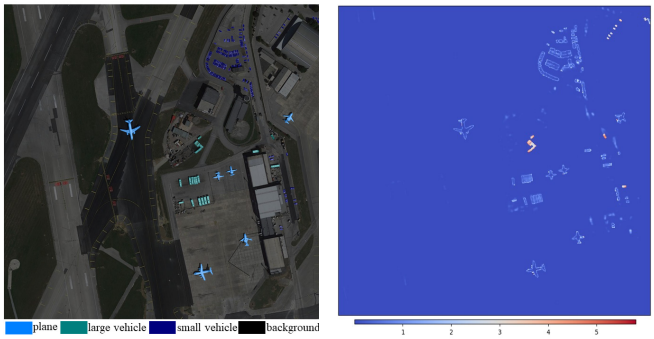
$$Z = p_b p_{m_0} + \sum_{i=1}^{N-1} (1 - p_b) p_{m_i}. \tag{5}$$

This factor makes p_i^C satisfy the constraint that $\sum_{i=0}^{N-1} p_i^C = 1, p_i^C \geq 0$. p_i^C is then used to replace p_i for the final multiclass probability in (3). With the CP loss, the two-step task is converted into one-step end-to-end optimization. In the proposed framework, the CE loss is adopted during the training, and p_i^C denotes the final multiclass probabilities during the inference. The CP loss is defined as follows:

$$\begin{aligned} CP(p_b, p_{m_i}, y) &= \mathrm{CE}\left(p_i^C, y\right) = -\sum_{i=0}^{N-1} y_i \log\left(p_i^C\right) \\ &= -y_0 \log p_b p_{m_0} - \sum_{i=1}^{N-1} y_i \log p_{m_i} \\ &\quad - \sum_{i=1}^{N-1} y_i log(1 - p_b) + \log Z. \end{aligned} \tag{6}$$

Differing from the previous multitask loss design [48], which leads different structures to learn task-specific features, the CP loss design fuses the features from the dual-branch decoder at the probability level. Compared with the multitask loss strategy, the CP loss has two advantages.

1) It can avoid optimization conflicts as the multitask loss optimizes the network by fusing the task-specific losses from the different outputs, ignoring the divergence in the different tasks. However, the unreasonable distribution of the loss scale factors for different tasks can lead to optimization conflicts. Specifically, unreasonable scale factors for binary-class classification and multiclass classification may cause imbalanced optimization in the FA and SR branches. The CP loss fuses the two branch outputs at the probability level and obtains a single loss, avoiding the conflicts in an explainable and effective way.
2) *Unified Probability Framework:* The multitask loss design can only use the SR branch outputs for prediction if there is no postprocessing fusion strategy, but this approach may lose some useful features in the FA branch. However, through the CP assumption, the network optimization and prediction can be performed under the same probability framework. Specifically, the two outputs are effectively fused [referring to (4) and (5)] during the prediction.



(a) (b)

Fig. 3. Visualization of the losses during model training. (a) An example from the iSAID dataset. (b) visualization of the losses in the heat map.

D. Small Object Mining-Based Network Optimization

The proposed SOM-based network optimization strategy is a modification of the OHEM strategy [37] in the object detection task. The small objects are considered to contain most of the "hard samples," which are the samples that are poorly predicted by the model. The hard samples can be estimated with the training loss, which is the metric between the predicted value and the target supervised learning. The loss is positively related to the difficulty of the sample. More specifically, the loss is higher when the sample is harder.

It can be observed that the large-scale background regions are usually continuously distributed. These regions have similar spectra and high homogeneity and are, thus, easy to classify. However, the small objects are often discretely distributed, have complex features, and are, thus, difficult to classify. Accordingly, an example was selected from the evaluation set, and we visualized the losses in the heat map during the model training. As shown in Fig. 3, it can be observed that most of the background samples have low losses, while the foreground samples have high losses, especially at the edge parts. Moreover, the number of hard samples is much smaller than the number of easy samples. This observation again reflects the imbalanced sample problem in the HRS imagery semantic segmentation task.

Our goal is to find these hard samples located in the small objects during the training, to correct the optimization direction. The OHEM strategy is modified into a semantic segmentation task, where each pixel is a sample. Differing from the existing OHEM method (e.g., Online Bs.) [40], the proposed SOM strategy fixes the training samples in each minibatch, which is a more suitable approach for the HRS segmentation task. More specifically, the SOM-based network optimization proceeds as shown in Algorithm 1. This algorithm automatically selects hard samples based on the corresponding losses. Thus, the model is updated at exactly the same frequency as the baseline stochastic gradient descent (SGD) approach, without any additional time consumption.

IV. Experiments

A. Dataset Description

In order to evaluate the performance of the proposed FactSeg framework, two benchmark HRS imagery datasets with different environmental settings were adopted.

Algorithm 1 SOM-Based Network Optimization

Data: Training image I, training label L
Result: Converged model M'
Set the sampling ratio r.
Initialize the model M.
while M *not converged* **do**
  Get the mini-batch data X and Y from I and L.
  Forward propagation $P = M(X)$.
  Get the training loss $L = CE(P, Y)$.
  Rank the N losses in L at the pixel level.
  $K = r * N$;
  Select the top K losses L.
  $l = \frac{\sum_{i=1}^{K} l_i}{K}$;
  Optimize M based on l.
end
Get the converged model $M' \leftarrow M$.

1) iSAID: The Instance Segmentation in Aerial Images dataset (iSAID) [11]. This dataset was modified from a large-scale object detection dataset [50]. The dataset contains 655 451 object instances with 15 categories across 2806 HRS optical images. The categories are ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, bridge, large vehicle, small vehicle, helicopter, swimming pool, roundabout, soccer ball field, plane, and harbor. For the dataset split, the training set contains 1411 images, the validation set contains 458 images, and the test set contains 937 images. The size of the images ranges from 12029 × 5014 to 455 × 387. Because the semantic annotations of the test set are not available, the validation set was used to evaluate the performance of the proposed method and the reference methods.

2) Vaihingen: The Vaihingen dataset [51] contains 33 HRS images, each with three bands, corresponding to near-infrared (NIR), red (R), and green (G) wavelengths. The average image size is 2494 × 2064. The corresponding digital surface model (DSM) is provided but was not used in our experiments. Among the images, 16 are manually annotated with pixelwise labels, with each pixel classified into one of five land-cover classes: impervious surface, building, low vegetation, tree, car, and clutter. In order to focus on the small objects, the dataset was reconstructed so that it contained just buildings and cars, and the other classes were merged into the background. The 16 tiles for which the ground truth is available were split into a training subset (tile numbers: 1, 3, 11, 13, 15, 17, 21, 26, 28, 32, 34, and 37) and a hold-out subset for evaluation (tile numbers: 5, 7, 23, and 30).

B. Evaluation Metrics

Accuracy and efficiency are important metrics for the HRS small object semantic segmentation task. The intersection over union (IoU) is chosen for the evaluation accuracy, and the prediction speed is used for efficiency. The IoU is calculated as follows:

$$\mathrm{IoU}_i = \frac{x_{ii}}{\sum_{i=1}^{n} x_{ij} + \sum_{i=1}^{n} x_{ji} - x_{ii}} \tag{7}$$

TABLE II
PERFORMANCE OF THE REFERENCE METHODS AND THE PROPOSED FACTSEG METHOD ON THE VAIHINGEN DATASET

Model	Backbone	mIoU(%)	IoU per class(%)		
			Background	Building	Car
FCN8S	VGG16	79.52	94.05	83.52	61.00
U-Net	-	76.80	94.01	83.57	52.83
DenseASPP	DenseNet121	80.94	95.02	**86.63**	61.19
SFPN	ResNet50	81.48	95.01	86.32	63.11
RefineNet	ResNet50	79.74	94.86	86.02	58.34
PSPNet	ResNet50	81.46	**95.06**	86.33	63.01
DeepLabv3	ResNet50	79.69	94.96	86.33	57.78
DeepLabv3+	ResNet50	80.97	94.99	86.28	61.63
FactSeg	**ResNet50**	**82.09**	95.05	86.37	**63.95**

$$m\mathrm{IoU} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IoU}_i \tag{8}$$

where x_{ij} represents the number of instances of class i predicted as class j, and n is the number of classes.

For the model efficiency, the theoretical and practical indices are both evaluated. The theoretical index is the parameter size, i.e., the sum of the parameters in the operations (convolution, deconvolution, batch normalization, and so on). The practical index is the prediction speed, which is evaluated in a real environment.

C. Vaihingen Dataset Experiments

1) Experimental Settings: In order to prove the effectiveness of the FactSeg framework, several state-of-the-art segmentation networks were chosen for the comparison experiments. The reference networks were FCN8S [52], U-Net [53], DenseASPP [54], the semantic FPN (SFPN) [47], RefineNet [44], PSPNet [42], DeepLabv3 [43], and DeepLabv3+ [55]. All the methods were run under the same settings, for fairness, and comparative experiments on the Vaihingen dataset were conducted. The foreground and background class proportions were as follows: foreground (cars and buildings): 27.90% and background: 72.10%. During the training, the SGD optimizer was adopted with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate was set to 0.007, and a "poly" schedule with power 0.9 was applied. The number of training iterations was set to 20 000, and the base learning rate was set to 0.03. The 512 × 512 patches were randomly cropped from the raw images, with random mirroring and rotation. The batch size was set to 4, and in the evaluation, sliding window inference technology was adopted. The size of the window patch was 512 × 512, and the stride was 256. For FactSeg, the sampling rate in the SOM strategy was set to 0.7. All the networks are implemented under the PyTorch deep learning framework, using NVIDIA's automatic mixed-precision training strategy for acceleration. The results are listed in Table II.

The results for the Vaihingen dataset demonstrate the effectiveness of FactSeg. All the methods achieve better performance on the buildings and poorer performance on the cars. The cars are more difficult to recognize because the cars are much smaller than the buildings. The methods all suffer from weak features and imbalanced samples when processing the small objects. In addition, the smaller the object, the more

obvious these problems are. However, the proposed FactSeg framework achieves good performances on these two object types, obtaining an mIoU of 82.09%. For the car class, FactSeg achieves the best performance among all the methods, with an mIoU of 63.95%.

In order to show the differences between the comparative methods, some visualizations on the test set are shown in Fig. 4. As can be seen, all the methods recognize the large buildings correctly. However, FCN8S and U-Net perform poorly in recognizing the edges of buildings and some small buildings due to their shallow layers. The small cars are difficult to recognize, especially when covered by shadow or obscured by trees. The parking lot scene is also shown at the bottom right in Fig. 4. The reference methods do not perform well with this complex scene. The closer cars in the parking lot are easily merged into one object, and the cars close to the buildings are easily misclassified. However, the proposed FactSeg framework addresses this complex scene well because the foreground objects are extracted in the binary outputs, as shown in Fig. 4(k). This relieves the burden on the multiclass branch, which only focuses on recognizing the foreground objects. The close cars are clearly recognized and divided, making the subsequent calculation of the number and area of the objects more accurate.

Fig. 4(k) shows the binary-class probability map, which represents the activated foreground objects for FactSeg. From this visualization, two notable characteristics can be observed:

1) The large objects have higher activation probabilities, while the small objects have lower activation probabilities. This is because the larger objects contain more distinct features.
2) Some fake regions are activated in the binary-class probability map but corrected in the final results. This reflects the fact that the binary-class output is a coarse intermediate result. Guided by the FA probability map, the multiclass branch can not only distinguish the different foreground objects but also refine the binary result.

The qualitative results and visualizations demonstrate the effectiveness of the proposed FactSeg framework on the Vaihingen dataset.

D. iSAID Dataset Experiments

1) Experimental Settings: After evaluating the effectiveness of the proposed FactSeg framework on the Vaihingen dataset, the large-scale iSAID dataset was chosen for further study. With this dataset, the experiments were conducted from three different perspectives: ablation experiments, accuracy comparison experiments, and model efficiency experiments. The foreground and background class proportions were foreground (16 classes): 97.14% and background: 2.86%. In order to ensure fairness, all the experiments were performed under the same settings, i.e., 896×896 patches were randomly cropped from the raw images, with random mirroring and rotation. An SGD optimizer was adopted with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate was set to 0.007, and a "poly" schedule with power 0.9 was applied. The batch size was 8, and all the networks were trained for 60000 steps. All the networks were implemented under the PyTorch deep learning framework. In the accuracy evaluation, due to the large size of each image, sliding window inference technology was adopted. The size of the window patch was 896×896, and the stride was 512. For the overlapping pixels, the mean of the prediction probability was adopted.

2) Ablation Experiments: In this section, we describe the comprehensive experiments performed to analyze the proposed modules and hyperparameters, including three aspects.

1) *Module Analysis:* The dual-branch decoder, CP loss, and SOM were separated from the FactSeg framework, and their effectiveness was analyzed individually.
2) *Optimization Analysis:* Different sampling methods were used to address the imbalanced sample problem and were tested with the CE loss and the proposed CP loss.
3) *Hyperparameter Analysis:* The adjustable parameter r in the SOM strategy was tested with different values. All these experiments were performed on the iSAID HRS dataset under the same settings, if not specified.

Module Analysis: The module analysis settings were given as follows.

1) The baseline method was the same as SFPN [47], with only the SR branch and CE loss optimization. The baseline consisted of a deep residual encoder, which was modified from ResNet50. The SR branch was a variant of an FPN for multiclass segmentation, as described in Section III-B.
2) Baseline with the FA branch (the dual-branch decoder supervised by multitask loss) [48].
3) Baseline with FA branch and CP loss.
4) The full FactSeg framework.

Table III shows the relative gains of each proposed module based on the baseline (1). It can be observed that the baseline obtains a poor performance of 59.31%, indicating its limited ability in addressing small object semantic segmentation in HRS imagery. Notably, the addition of the FA-branch 2) only brings about a slight improvement of 0.60%. This is because the dual-branch decoder supervised with multitask loss does not change the traditional workflow, and it has difficulty in utilizing the two branch outputs effectively. However, the addition of the CP loss 3) results in a qualitative leap of 2.05%. This is because the CP loss effectively fuses the outputs of the dual-branch decoder, formulating the FA representation framework. As shown in Table III (d), the SOM strategy boosts the performance by 2.83% in mIoU, without extra parameters. The results of the module analysis experiments not only confirm the effectiveness of each module but also the compatibility of the overall framework.

Optimization Analysis: In order to further analyze the SOM strategy, as well as the CP loss, several comparative experiments with the existing sampling methods (addressing imbalanced samples) were conducted. Two types of sampling strategies were chosen: 1) inverse weighting [35], where pixel-level losses are weighted according to the inverse proportion of each category in the minibatch and 2) Online Bs. [40], where the imbalanced samples are addressed using the OHEM

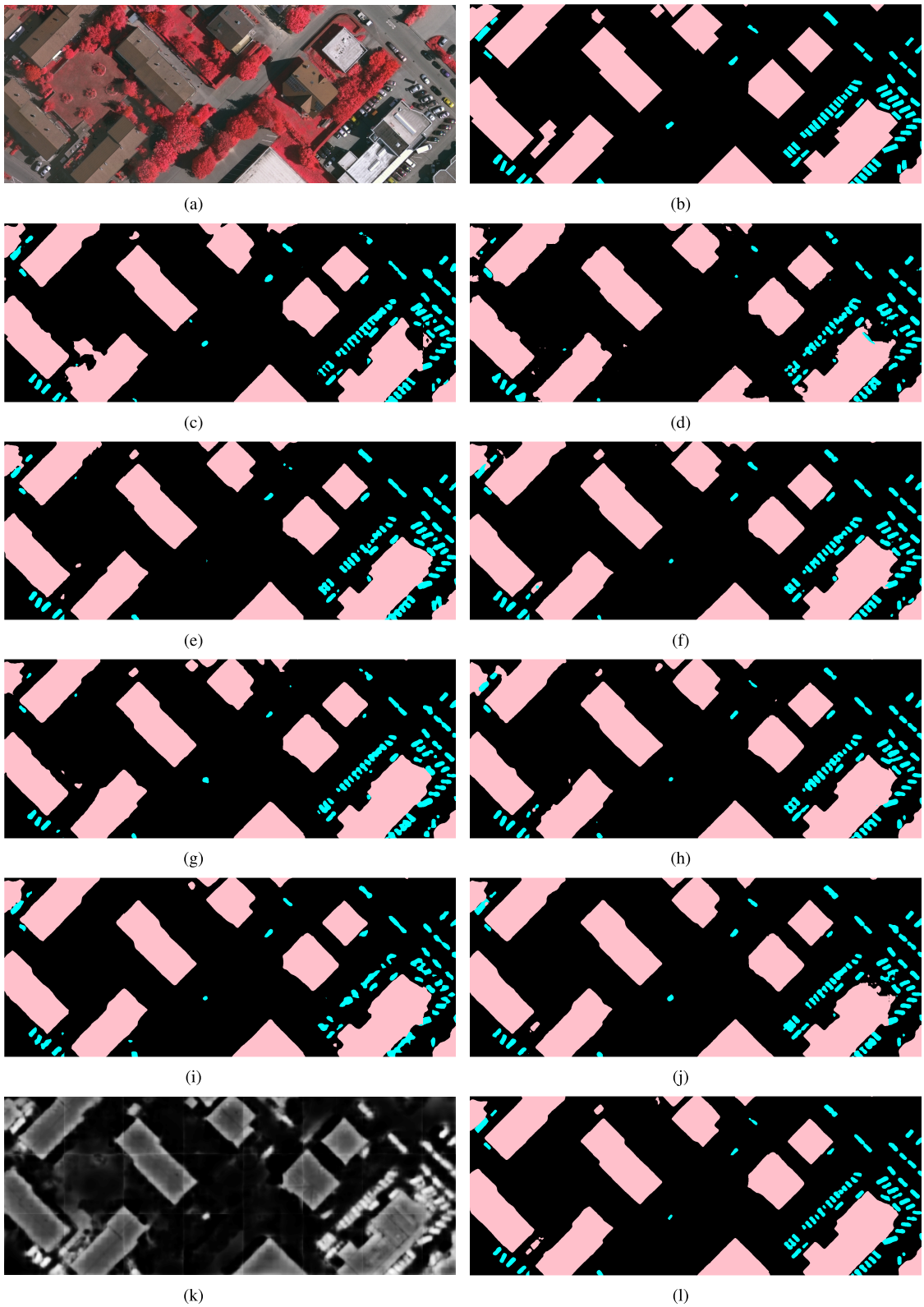


Fig. 4. Visualization results for the Vaihingen validation set. **Legend: background**, building, and car. Class distribution: background (70.60%), building (26.51%), and car (2.89%). All figures in this article are best viewed digitally with zoom. (a) HRS image. (b) Ground truth. (c) FCN8S. (d) U-Net. (e) DenseASPP. (f) SFPN. (g) RefineNet. (h) PSPNet. (i) DeepLabv3. (j) DeepLabv3+. (h) FactSeg binary-class probability map. (i) FactSeg.

strategy for the segmentation task and the hard samples are filtered based on the probability threshold. The baseline was the same as (b) in Table III.

As shown in Table IV, the inverse weighting strategy obtains the worse results, which are 2.92% lower than the baseline. Because the number of samples varies greatly in the different

TABLE III
RESULTS OF THE MODULE ANALYSIS EXPERIMENTS

Model	FA branch	CP loss	SOM	mIoU (%)
(a) Baseline	-	-	-	59.31
(b) Baseline w/FA branch	✓			59.91
(c) Baseline w/FA branch and CP loss	✓	✓		61.96
(d) FactSeg	✓	✓	✓	64.79

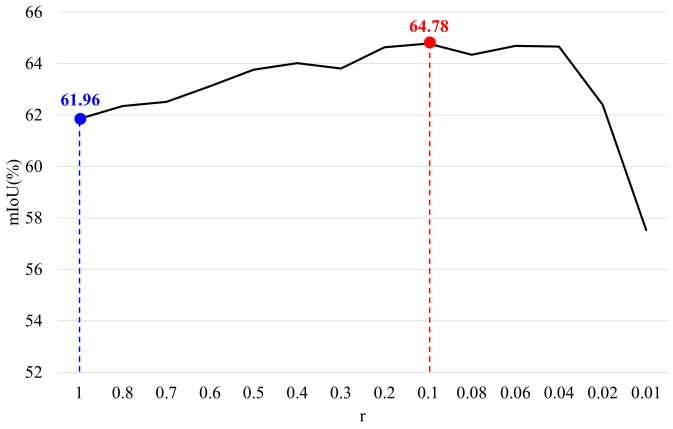


Fig. 5. Correlation curve between the sampling rate and the model accuracy.

TABLE IV
RESULTS OF THE OPTIMIZATION ANALYSIS EXPERIMENTS

Model	mIoU (%)
(a) CE loss	59.91
(b) CE loss w/inverse weight	56.99
(c) CE loss w/Online Bs.	62.73
(d) CE loss w/SOM	63.95
(e) CP loss	61.96
(f) CP loss w/Online Bs.	63.99
(g) CP loss w/SOM	64.79

categories, the inverse weighting approach generates weights that differ by orders of magnitude. This indicates that the inverse weighting strategy is not suitable for addressing the imbalanced sample problem in the HRS imagery small object semantic segmentation task. The Online Bs. method achieves gains of 2.82% and 2.03% according to (c) and (d), which proves the effectiveness of the OHEM strategy. Compared with Online Bs., the proposed SOM strategy further improves the performance by 1.22% and 0.80% on both the CE and CP losses. This demonstrates that the SOM strategy is a more suitable approach for the small object semantic segmentation task for HRS imagery.

Hyperparameter Analysis: The SOM strategy has the adjustable parameter sampling ratio r, which is an important parameter for FactSeg. Parameter r represents the strength of the SOM strategy. When r is lower, more samples are discarded. A series of parameter sensitivity analysis experiments were performed to explore the influence of this parameter. The other settings were kept the same, and the values of the scale were varied from 1.0 to 0.01.

As shown in Fig. 5, the model without the SOM strategy ($r = 1.0$) achieves a low accuracy. This reveals the fact that the abundant easy samples mislead the model optimization. When the sampling ratio is 0.1, the optimization achieves the best effect. In this experimental environment, the effective hard samples were only about one-tenth of the entire dataset. When the sampling ratio is less than 0.1, the optimization gradually becomes worse as the ratio decreases and finally collapses. This is because the training samples are not sufficient for the model optimization.

3) Comparative Experiments: In order to prove the effectiveness of the FactSeg framework, several state-of-the-art segmentation networks were chosen for the comparison experiments. The reference networks were FCN8S [52], U-Net [53], DenseASPP [54], SFPN [47], RefineNet [44], PSPNet [42], DeepLabv3 [43], DeepLabv3+ [55], FarSeg [32], DenseU-Net [33], and Mask-RCNN [46]. For Mask-RCNN, the implementation and experimental setting were the same as for the iSAID benchmark [11]. The performance of Mask-RCNN is reported as the IoU and instance segmentation and object detection indices [11]. All the experimental settings were the same as those described in Section IV-D1.

The experimental results are listed in Table V. It can be observed that FCN8S and U-Net achieve relatively poor performances due to their shallow layers. FCN8S and U-Net only have 16 and 18 layers in the backbone, respectively, and are not able to fit such a complex and large-scale dataset. FactSeg achieves the best performance and outperforms the other methods in 13 categories. All the models obtain a high accuracy for the tennis court category because this object category features a large area of artificial green turf, which results in low intraclass variance. The helicopter class is difficult to recognize because the helicopters are small and scarce in the training dataset. In order to intuitively display the differences between the comparative methods, some typical validation results are visualized in Fig. 6, where it can be seen that the key objects are scarce and separately distributed. As for the existing small object semantic segmentation methods, FarSeg achieves a relatively good performance because it enhances the small object features via the foreground-aware relations. DenseU-Net outperforms DenseASPP under the same feature extractor (DenseNet121) and keeps richer details in the objects.

Mask-RCNN achieves a relatively poor performance, especially for dense and extreme-scale objects. As shown in Fig. 6(m), the small vehicles in the parking lot are not recognized well. This is because errors exist in the region proposals, which affects the segmentation results. Compared with the reference models, FactSeg achieves the best visual segmentation results, especially in restoring the details and

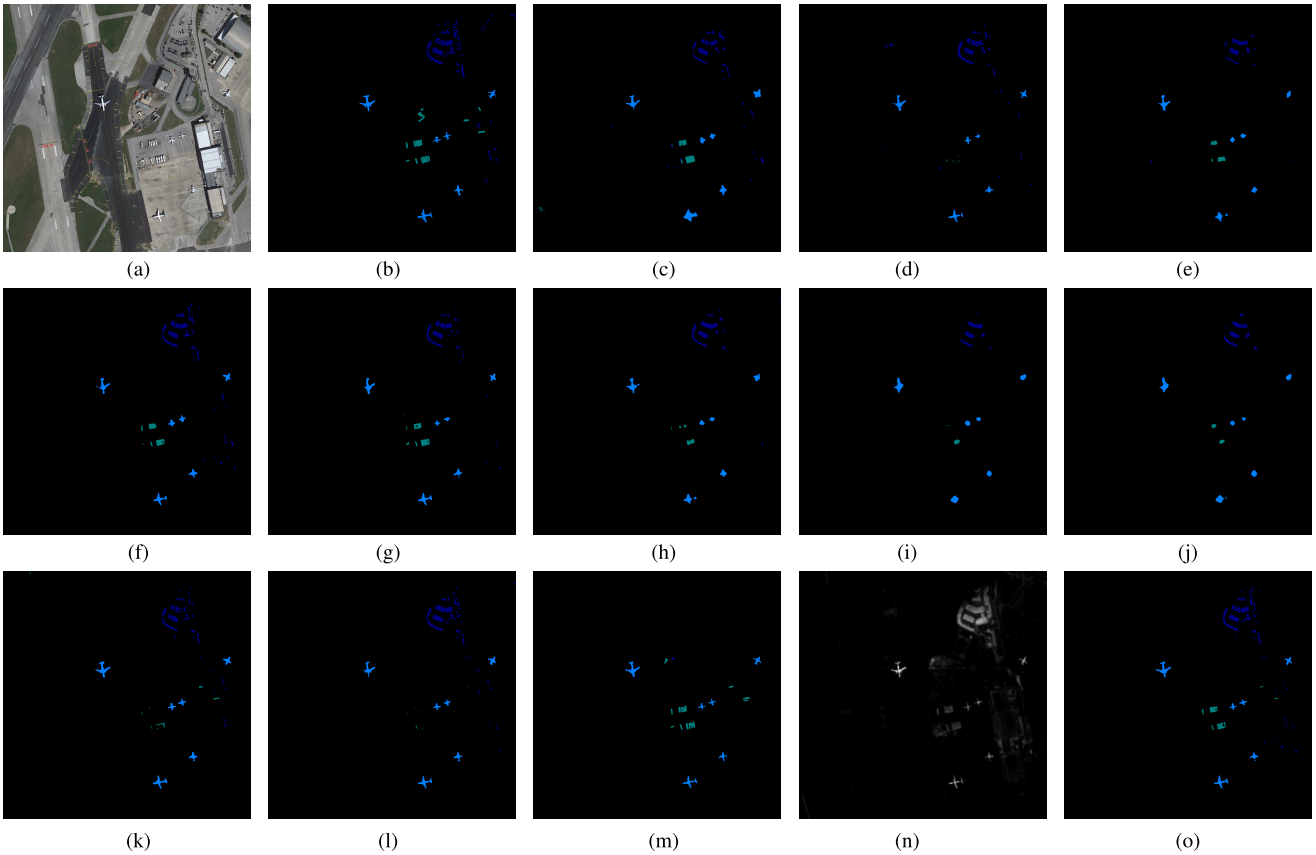


Fig. 6. Visualization results on the iSAID validation set. **Legend: background**, **plane**, **large vehicle**, and **small vehicle**. Class distribution: background (98.65%), plane (0.36%), large vehicle (0.30%), and small vehicle (0.69%). All figures in this article are best viewed digitally with zoom. (a) HRS image. (b) Ground truth. (c) FCN8S. (d) U-Net. (e) DenseASPP. (f) SFPN. (g) RefineNet. (h) PSPNet. (i) DeepLabv3. (j) DeepLabv3+. (k) FarSeg. (l) DenseU-Net. (m) Mask-RCNN. (n) FactSeg binary-class probability map. (o) FactSeg.

TABLE V

PERFORMANCE OF THE REFERENCE METHODS AND THE PROPOSED FACTSEG FRAMEWORK ON THE iSAID DATASET

Model	Backbone	mIoU(%)	IoU per class(%)															
			BG	Ship	ST	BD	TC	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor
FCN8S	VGG16	41.65	98.31	51.74	22.90	26.43	74.80	30.23	27.85	8.17	49.34	37.04	0	30.74	51.90	52.06	62.89	42.01
U-Net	-	39.20	98.21	48.99	0	36.51	78.59	22.89	5.51	7.47	49.88	35.62	0	38.02	46.48	9.67	74.74	45.63
DenseASPP	DenseNet121	56.80	98.77	61.14	50.01	67.53	86.08	56.55	52.27	29.61	57.09	38.44	0	43.26	64.80	74.10	78.12	51.09
SFPN	ResNet50	59.31	98.84	63.68	59.49	71.75	86.61	57.78	51.64	33.99	59.15	45.14	0	46.42	68.71	73.58	80.83	51.27
RefineNet	ResNet50	60.20	98.82	63.79	58.55	72.30	85.27	61.09	52.77	32.62	58.22	42.35	22.97	43.40	65.63	**74.41**	79.89	51.09
PSPNet	ResNet50	60.25	98.82	65.2	52.1	75.7	85.57	61.12	**60.15**	32.46	58.03	42.96	10.89	46.78	68.6	71.9	79.5	54.26
DeepLabv3	ResNet50	59.04	98.71	59.74	50.49	76.98	84.20	57.91	59.56	32.87	54.79	33.74	31.28	44.74	66.03	72.13	75.83	45.68
DeepLabv3+	ResNet50	60.82	98.84	63.89	52.45	72.80	84.89	56.53	58.86	32.24	59.12	42.88	31.43	46.10	67.71	72.91	79.82	52.62
FarSeg	ResNet50	63.71	98.84	65.38	**61.80**	77.73	86.35	62.08	56.70	**36.70**	60.59	46.34	35.82	51.21	71.35	72.53	82.03	53.91
DenseU-Net	DenseNet121	58.70	98.85	66.05	50.42	76.13	86.16	57.68	49.45	33.87	54.73	46.20	0	45.14	65.86	71.88	82.18	54.58
Mask-RCNN*	ResNet101	54.18	98.43	56.62	36.05	59.67	86.47	60.44	41.70	29.81	57.99	33.97	31.82	41.64	67.40	56.12	74.61	34.06
FactSeg	**ResNet50**	**64.79**	**98.93**	**68.34**	56.83	**78.36**	**88.91**	**64.89**	54.60	36.34	**62.65**	**49.53**	**42.72**	**51.47**	**69.42**	73.55	**84.13**	**55.74**

* Mask-RCNN instance segmentation results: $AP = 31.80, AP_{50} = 54.60, AP_{75} = 32.60, AP_S = 18.80, AP_M = 38.70, AP_L = 43.30$
* Mask-RCNN object detection results: $AP = 37.30, AP_{50} = 58.80, AP_{75} = 40.50, AP_S = 22.10, AP_M = 45.60, AP_L = 49.10$
The abbreviations are as follows: BG - background, ST - storage tank, BD - baseball diamond, TC - tennis court, BC - basketball court, GTF - ground track field, LV - large vehicle, SV - small vehicle, HC - helicopter, SP - swimming pool, RA - roundabout, SBF - soccer ball field.

edge parts. In both the quantitative and visual experimental results, FactSeg outperforms the reference semantic segmentation methods.

4) Efficiency Analysis Experiments: In order to evaluate the efficiency of FactSeg and the reference methods, a series of efficiency analysis experiments were performed. The theoretical and practical indices for the model were tested in the same real environment. The efficiency tests were carried out on a single NVIDIA Tesla P100 graphics card with double floating-point precision operations, with the inputs being 896 × 896 images under the PyTorch deep learning framework, without any additional optimization. The efficiency experiment results are listed in Table VI.

It can be concluded that U-Net has a minimal number of parameters because it contains lightweight convolutions. SFPN has the fastest prediction speed because the convolutions in the decoder have only a few channels, requiring fewer operations. FactSeg has relatively lightweight parameters of 33.44 M, and a high prediction speed of 18.48 samples/s.

Fig. 7 is the visualization of the prediction speed versus accuracy on the iSAID validation set. This indicates that FactSeg achieves a better tradeoff between speed and accuracy,

TABLE VI

EFFICIENCY OF THE REFERENCE NETWORKS AND THE PROPOSED FACTSEG FRAMEWORK ON THE ISAID DATASET

Model	Backbone	Params(M)	Speed(samples/sec)
FCN8S	VGG16	50.26	18.27
U-Net	-	7.85	19.80
DenseASPP	DenseNet121	9.19	11.89
SFPN	ResNet50	28.47	25.31
RefineNet	ResNet50	94.88	7.89
PSPNet	ResNet50	53.31	5.60
DeepLabv3	ResNet50	39.04	17.08
DeepLabv3+	ResNet50	39.16	15.29
FarSeg	ResNet50	29.59	22.01
DenseU-Net	DenseNet121	13.61	14.83
FactSeg	**ResNet50**	**33.44**	**18.48**

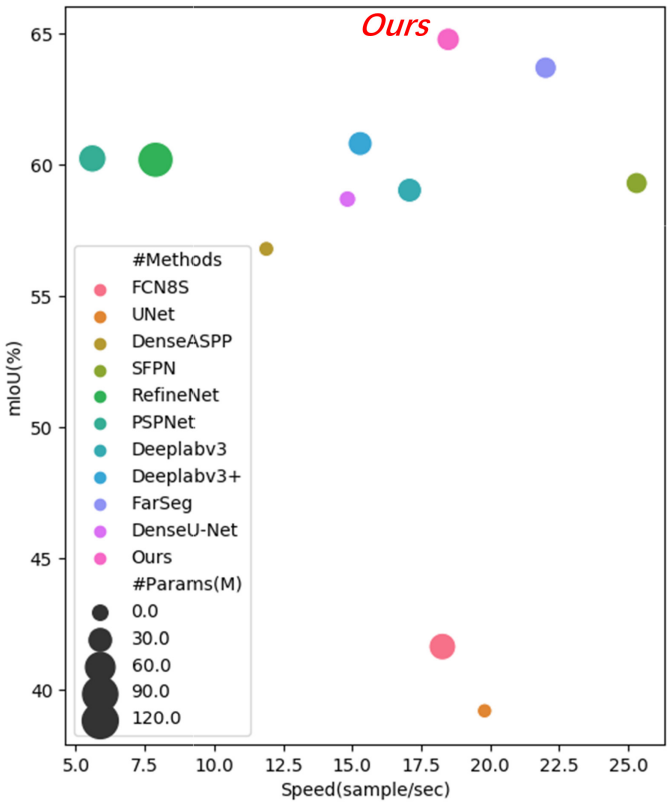


Fig. 7. Visualization of the prediction speed versus validation accuracy on the iSAID validation set. The radius of the circles represents the number of parameters.

as it benefits from the lightweight architecture and FA object semantic segmentation workflow.

V. DISCUSSION OF THE APPLICATION POTENTIAL

The results obtained with the two HRS datasets confirm the effectiveness of FactSeg. Three large-scale complex scenes were selected to investigate the small object semantic segmentation mapping performance. These scenes were the airport scene (6313 × 3098 pixels) and port terminal scene (3240 × 1158 pixels) in the iSAID evaluation set, as well as the residential area scene (2562 × 1729) in the Vaihingen evaluation set. The IoU was chosen for the initial mapping result evaluation. Moreover, the statistics were calculated after the postprocessing. For the postprocessing, the seed filling algorithm was used for obtaining the object instances. The number of objects and the area in each class are also reported.

TABLE VII

STATISTICAL RESULTS FOR THE AIRPORT SCENE

		LV	SV	HC	Plane
Area	Groundtruth	186	3719	5757	1291893
	FactSeg	1854	3074	6839	1278963
	Diff(%)	−89.96	−17.34	+18.79	+1.00
Number	Groundtruth	6	20	9	281
	FactSeg	1	19	10	297
	Diff(%)	−83.33	−5.00	+11.11	−5.69
IoU(%)	(mean)64.63	10.03	48.80	77.48	87.84

The abbreviations are: LV - large vehicle, SV - small vehicle, HC - helicopter.

TABLE VIII

STATISTICAL RESULTS FOR THE PORT TERMINAL SCENE

		Ship	Bridge	SV	SP	Harbor
Area	Groundtruth	6503	52557	7562	608	61525
	FactSeg	5654	58837	4855	419	41603
	Diff(%)	−13.05	+11.94	−35.79	−31.08	−32.38
Number	Groundtruth	21	1	32	1	18
	FactSeg	21	1	31	1	24
	Diff(%)	0.00	0.00	−3.12	0.00	+33.33
IoU(%)	(mean)70.89	67.12	81.28	52.67	68.91	56.48

The abbreviations are: SV - small vehicle, SP - swimming pool.

The FactSeg prediction mask was stacked over the raw images, as shown in Fig. 8. The results for the airport scene are reported in Table VII. In this scene, the number of planes is much more than the number of the other objects, but FactSeg still achieves the highest IoU of 87.84%. For this typical scene, FactSeg obtains a lower statistical error on the planes, i.e., area: +1.00% and number: −5.95%. Although the helicopters are the most difficult objects to recognize, as shown in Table V, FactSeg again obtains relatively accurate statistics, i.e., area: +18.79% and number: +11.11%. However, most of the large vehicles are misclassified into small vehicles because they have similar sizes in this airport scene. This results in the area and number for these two classes being inaccurate. However, overall, for this airport scene, the results of the proposed FactSeg framework could provide a good basis for aircraft statistics.

The prediction mask for the port terminal scene is visualized in Fig. 9. The spectra between the bridge and road are similar, making them difficult to distinguish. However, it is notable that the bridge is clearly separated from the road, with the area error being +11.94% (Table VIII). This is because FactSeg contains the multiscale fusion module, allowing it to obtain the water context, to distinguish the bridge from the road. The numbers of ships, bridges, and swimming pools are predicted perfectly. However, the number of predicted harbors is more than the ground truth, due to some fractures.

The residential area scene was chosen from the Vaihingen dataset. The prediction mask for this scene is visualized in Fig. 10. Compared with the airport and port terminal scenes, the residential area scene has more homogeneous objects, and FactSeg achieves a higher mIoU of 86.75%. As is exhibited in Table IX, the number of cars is predicted perfectly but with a −15.99% error in area, due to the small sizes. FactSeg

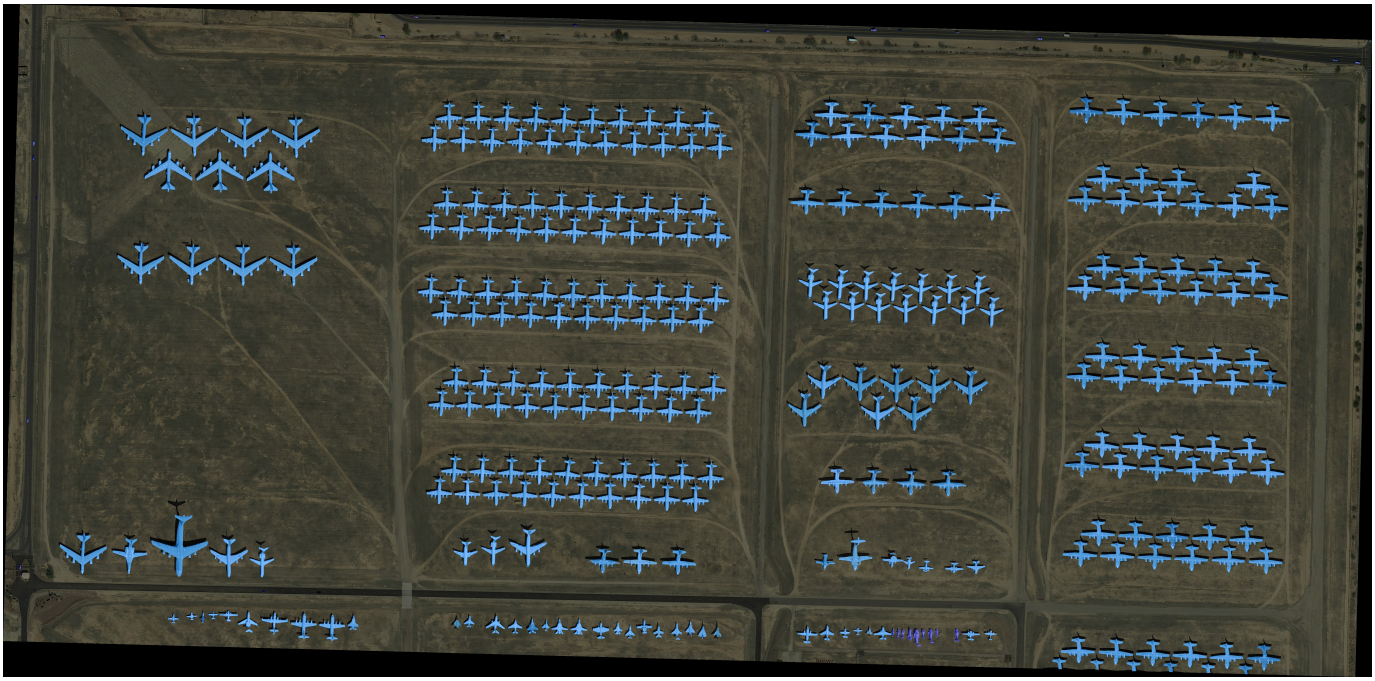


Fig. 8. Airport scene in the iSAID dataset. Class distribution: background (93.35%), large vehicle (0.01%), small vehicle (0.02%), helicopter (0.03%), and plane (6.60%). **Legend: background, small vehicle, large vehicle, helicopter, and plane.** All figures in this article are best viewed digitally with zoom.

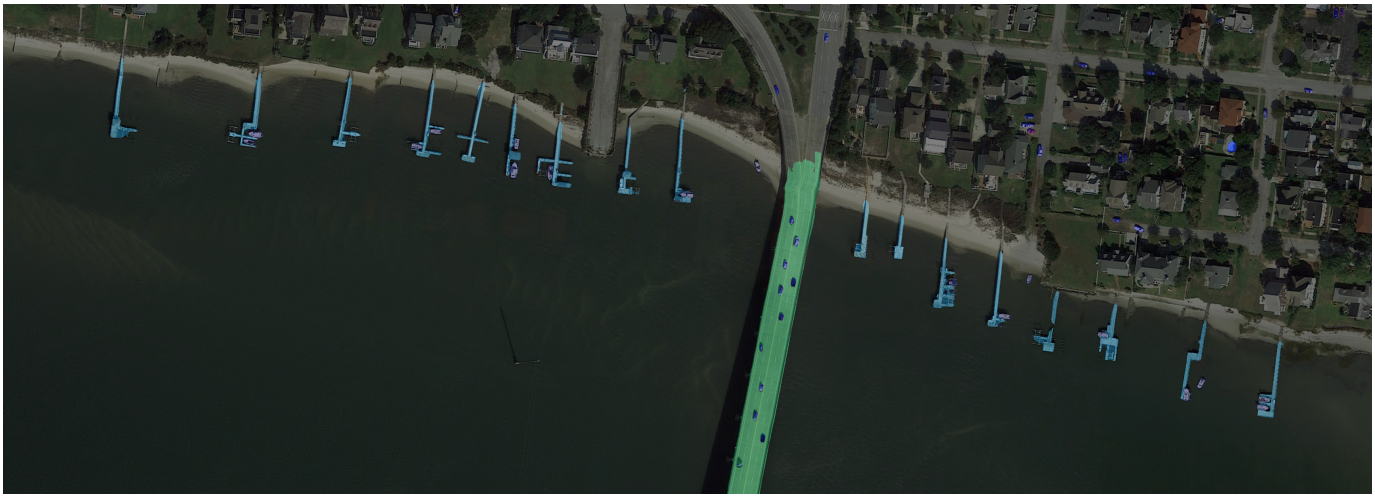


Fig. 9. Port terminal scene in the iSAID dataset. Class distribution: background (98.23%), ship (0.17%), bridge (1.40%), small vehicle (0.20%), swimming pool (0.02%), and harbor (1.63%). **Legend: background, ship, bridge, and swimming pool.** All figures in this article are best viewed digitally with zoom.

TABLE IX
STATISTICAL RESULTS FOR THE RESIDENTIAL SCENE

		Building	Car
Area	Groundtruth	964846	28745
	FactSeg	994469	24146
	Diff(%)	+3.07	−15.99
Number	Groundtruth	55	33
	FactSeg	61	33
	Diff(%)	+10.90	0.00
IoU(%)	(mean)86.75	89.31	74.23

achieves a low error rate in the area of buildings, at +3.07%, but it overestimates the number because of the fractures.

When the proposed FactSeg framework is applied to the three typical scenes, it shows strong potential, but some limitations are also apparent. However, FactSeg again outperforms the other methods in obtaining the key object information in these typical large-scale complex scenarios. Although the model may overestimate the number of objects, morphological methods could be applied to merge the fractures. The results indicate that the FactSeg framework could be applied to many small object semantic segmentation tasks, providing a basis for airport and port terminal monitoring, cadastral surveying, and traffic control.

VI. CONCLUSION

In this article, an FA-driven small object semantic segmentation (FactSeg) framework has been proposed for HRS imagery. Specifically, the proposed framework is made up of

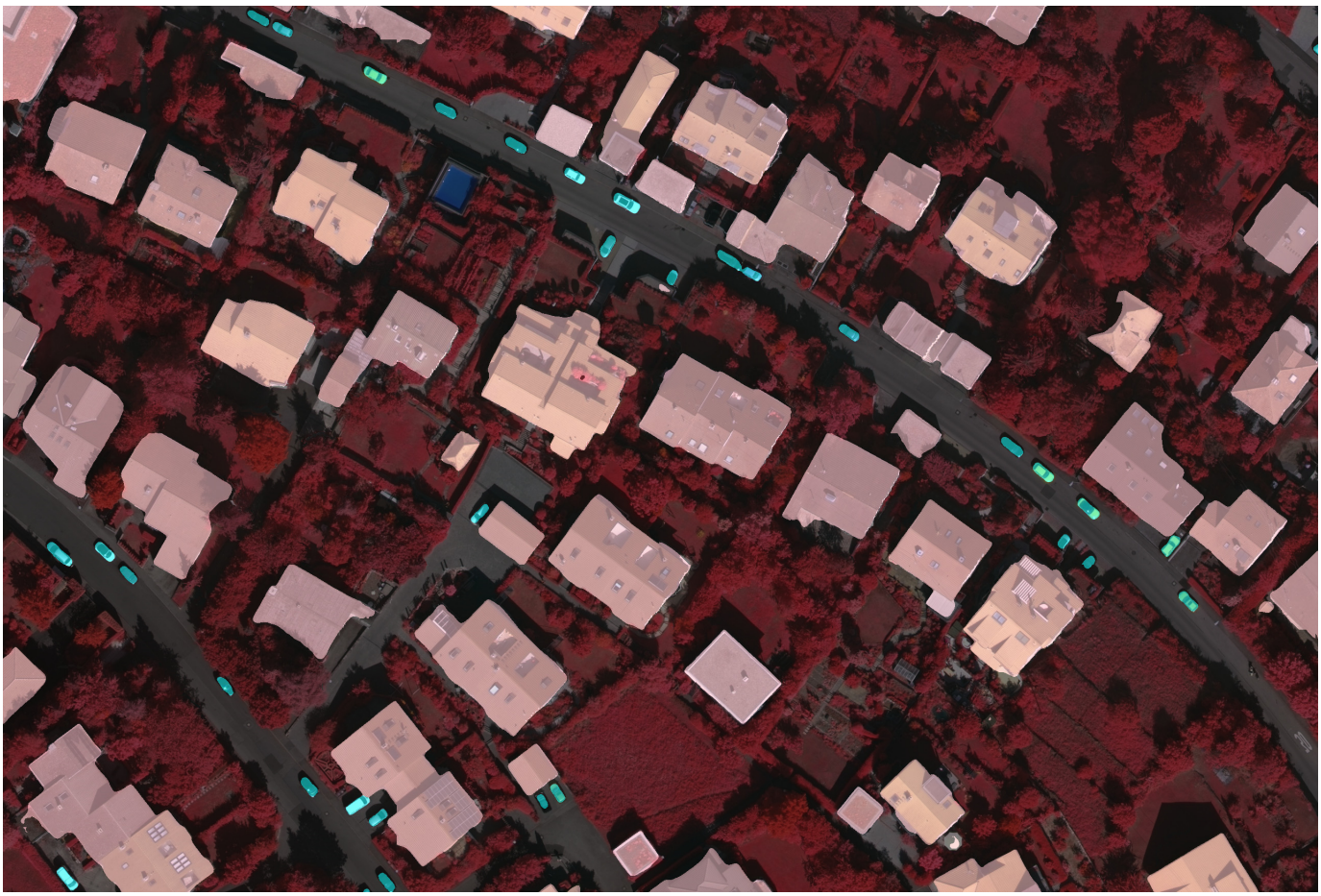


Fig. 10. Residential area in the Vaihingen dataset. Class distribution: background (77.61%), building (21.75%), and car (0.64%). **Vaihingen legend: background, building, and car.** All figures in this article are best viewed digitally with zoom.

a dual-branch decoder, CP loss, and SOM-based network optimization. After obtaining the deep semantic features through the deep residual encoder, the dual-branch decoder is used to separately extract the activation and refinement features. The activation and refinement features are then fused using the CP loss. For the optimization, pixel-level SOM-based network optimization is adopted to address the imbalanced sample problem.

Comparative experiments were conducted on two benchmark HRS imagery datasets, where FactSeg outperformed the state-of-the-art deep learning-based semantic segmentation methods of FCN8S, U-Net, DenseASPP, SFPN, RefineNet, PSPNet, DeepLabv3, and DeepLabv3+. In addition, ablation experiments and model efficiency analyses were also carried out on the iSAID HRS imagery dataset. In the ablation experiments, the effectiveness of each proposed module was proven. In the model efficiency analyses, the efficiency of FactSeg was evaluated, and it was found that the FactSeg framework achieved the highest accuracy, with a relatively high prediction speed. Three typical complex scenes were also chosen to evaluate the potential of FactSeg in practical applications. Moreover, the model application conditions and generalization ability were explored in foreground–background ratio experiments. In the future, because the FactSeg framework can provide a general object semantic segmentation workflow, which could be applied in other lightweight architectures, we will extend this to in-orbit satellite data processing [56], [57]. FactSeg will also be extended to other related tasks, such as instance segmentation [46] and object detection [48].

REFERENCES

[1] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1–9.

[2] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1442–1450.

[3] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 277–293, Aug. 2017.

[4] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

[5] C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu, "Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas," *Remote Sens. Environ.*, vol. 187, pp. 156–168, Dec. 2016.

[6] H. Wang, C. Wang, and H. Wu, "Using GF-2 imagery and the conditional random field model for urban forest cover mapping," *Remote Sens. Lett.*, vol. 7, no. 4, pp. 378–387, Apr. 2016.

[7] Z. Huang, F. Xu, L. Lu, and H. Nie, "Object-based conditional random fields for road extraction from remote sensing image," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 17, no. 1, 2014, Art. no. 012276.

[8] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.

[9] J. Wang, Y. Zhong, Z. Zheng, A. Ma, and L. Zhang, "RSNet: The search for remote sensing deep neural networks in recognition tasks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2520–2534, Mar. 2021.

[10] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017.

[11] S. Waqas Zamir *et al.*, "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 28–37.

[12] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer, "A survey on moving object detection for wide area motion imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[15] J. M. Pe na-Barragán, M. K. Ngugi, R. E. Plant, and J. Six, "Object-based crop identification using multiple vegetation indices, textural features and crop phenology," *Remote Sens. Environ.*, vol. 115, no. 6, pp. 1301–1316, Jun. 2011.

[16] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3278–3285.

[17] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[18] K. Wang and D. Ming, "Road extraction from high-resolution remote sensing images based on spectral and shape features," *Proc. SPIE*, vol. 7495, Oct. 2009, Art. no. 74953R.

[19] X. Huang, W. Yuan, J. Li, and L. Zhang, "A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017.

[20] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jan. 2015.

[21] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[22] M. Dickenson and L. Gueguen, "Rotated rectangles for symbolized building footprint extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 18–22.

[23] T.-S. Kuo, K.-S. Tseng, J.-W. Yan, Y.-C. Liu, and Y.-C.-F. Wang, "Deep aggregation net for land cover classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 252–256.

[24] S. Aich, W. van der Kamp, and I. Stavness, "Semantic binary segmentation using convolutional networks without decoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 197–201.

[25] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," 2018, *arXiv:1807.05713*. [Online]. Available: http://arxiv.org/abs/1807.05713

[26] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, p. 765, Mar. 2019.

[27] R. Lalonde, D. Zhang, and M. Shah, "ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4003–4012.

[28] C. Q. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[29] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[30] R. Dong, D. Xu, J. Zhao, L. Jiao, and J. An, "Sig-NMS-based faster R-CNN combining transfer learning for small target detection in VHR optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8534–8545, Nov. 2019.

[31] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

[32] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4096–4105.

[33] R. Dong, X. Pan, and F. Li, "DenseU-Net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65347–65356, 2019.

[34] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2017.

[35] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[36] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[37] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[38] H. Yu *et al.*, "Loss rank mining: A general hard example mining method for real-time detectors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[39] B. Kellenberger, D. Marcos, S. Lobry, and D. Tuia, "Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep CNNs and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9524–9533, Dec. 2019.

[40] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," 2016, *arXiv:1605.06885*. [Online]. Available: http://arxiv.org/abs/1605.06885

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[44] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[45] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[47] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6399–6408.

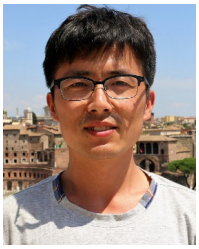
[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[49] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[50] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[51] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS," *ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Jul. 2016.

[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[54] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.
[55] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.
[56] D. Li, M. Wang, Z. Dong, X. Shen, and L. Shi, "Earth observation brain (EOB): An intelligent earth observation system," *Geo-Spatial Inf. Sci.*, vol. 20, no. 2, pp. 134–140, Feb. 2017.
[57] Y. Zhong, W. Li, X. Wang, S. Jin, and L. Zhang, "Satellite-ground integrated destriping network: A new perspective for EO-1 hyperion and chinese hyperspectral satellite datasets," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111416.



Ailong Ma (Member, IEEE) received the B.S. degree from the China University of Petroleum, Qingdao, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017.

He is a Research Associate with Wuhan University. His major research interests are remote sensing image processing, evolutionary computing, and pattern recognition.

Junjue Wang (Student Member, IEEE) received the B.S. degree from the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, in 2019. He is pursuing the master's degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan.

His major research interests are high-resolution remote sensing imagery semantic segmentation and computer vision.

Mr. Wang won the Second Place Prize in the Single-View Semantic 3D Challenge of the 2019 IEEE GRSS Data Fusion Contest and the Fourth Place in the Multitemporal Semantic Change Detection Challenge of the 2021 IEEE GRSS Data Fusion Contest.

Yanfei Zhong (Senior Member, IEEE) received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

Since 2010, he has been a Full professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. He organized the Intelligent Data Extraction, Analysis and Applications of Remote Sensing (RSIDEA) Research Group. He has published more than 100 research articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include hyperspectral remote sensing information processing, high-resolution remote sensing image understanding, and geoscience interpretation for multisource remote sensing data and applications.

Dr. Zhong is also a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). He won the Second-Place Prize in the 2013 IEEE GRSS Data Fusion Contest and the Single-View Semantic 3-D Challenge of the 2019 IEEE GRSS Data Fusion Contest, respectively. He is serving as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and the *International Journal of Remote Sensing*.

Zhuo Zheng (Graduate Student Member, IEEE) received the B.S. degree from the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, in 2018. He is pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan.

His major research interests multisource remote sensing imagery panoptic parsing and computer vision.

Mr. Zheng won the Second Place Prize in the Single-View Semantic 3D Challenge of the 2019 IEEE GRSS Data Fusion Contest.