Hi-ResNet: Edge Detail Enhancement for High-Resolution Remote Sensing Segmentation

Yuxia Chen*, Pengcheng Fang*, Xiaoling Zhong†, Jianhui Yu, Xiaoming Zhang, Tianrui Li, *Senior Member, IEEE*

***Abstract*—High-resolution remote sensing (HRS) semantic segmentation extracts key objects from high-resolution coverage areas. However, objects of the same category within HRS images generally show significant differences in scale and shape across diverse geographical environments, making it difficult to fit the data distribution. Additionally, a complex background environment causes similar appearances of objects of different categories, which precipitates a substantial number of objects into misclassification as background. These issues make existing learning algorithms sub-optimal. In this work, we solve the above-mentioned problems by proposing a High-resolution remote sensing network (Hi-ResNet) with efficient network structure designs, which consists of a funnel module, a multi-branch module with stacks of information aggregation (IA) blocks, and a feature refinement module, sequentially, and class-agnostic edge aware (CEA) loss. Specifically, we propose a funnel module to downsample, which reduces the computational cost, and extracts high-resolution semantic information from the initial input image. Secondly, we downsample the processed feature images into multi-resolution branches incrementally to capture image features at different scales. Furthermore, with the design of the Window multi-head self-attention, SE attention, and Depth-Wise convolution, the light-efficient IA blocks are utilized to distinguish image features of the same class with variant scales and shapes. Finally, our feature refinement module integrates the CEA loss function, which disambiguates inter-class objects with similar shapes and increases the data distribution distance for correct predictions. With effective pre-training strategies, we demonstrate the superiority of Hi-ResNet over the existing prevalent methods on three HRS segmentation benchmarks.**

***Index Terms*—Remote sensing, Semantic segmentation, Attention, Pre-training**

I. Introduction

In the geomatics community, the advancement of imaging technology allows us to obtain an increasing number of high-resolution remote sensing (HRS) images in real-time. These HRS images can be partitioned into distinct regions through pixel-level semantic segmentation, thereby providing more delicated details and features for applications such as urban planning [1], [2], environmental monitoring [3], and disaster management [4], [5]. Traditional segmentation methods typically use edge-based segmentation [6], [7], threshold-based segmentation [8], [9], and region-based segmentation [10], [11] to extract key information from HRS images. However, with the rapid development of remote sensing technology, traditional methods have gradually become insufficient for complex and diverse image segmentation tasks. Consequently, in order to achieve high-precision segmentation results, many researchers opt to apply street view semantic segmentation algorithms based on convolutional neural networks [12]–[15] and Transformer [16]–[18] to HRS segmentation tasks. However, these methods often perform poorly on HRS images. We attribute this to two primary reasons.

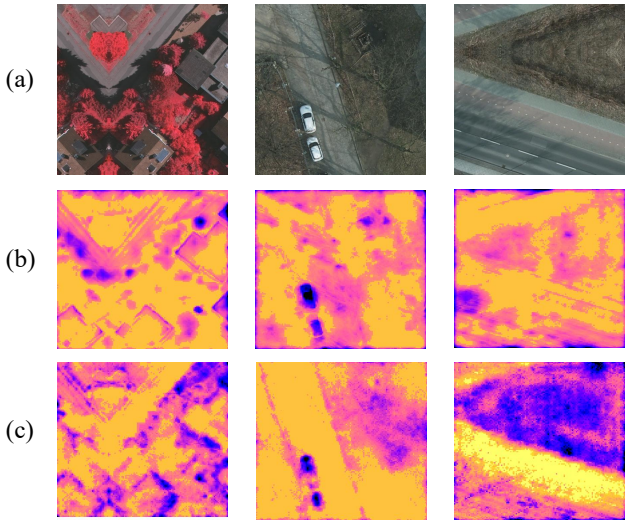


Fig. 1. Comparisons of our model behavior by heatmaps with different images illustrate the feature information obtained by upsampling and merging at the end of each layer for baseline and Hi-ResNet base. The three rows (a)(b)(c) show the original image, and the features of baseline and Hi-ResNet base separately. It is evident from the results that compared to the baseline, the Hi-ResNet base extracts richer and superior feature information.

First, unlike conventional street-level images, objects of the same category in HRS images are often located in different geographical landscapes, leading to more scale, shape, and distribution variations for these objects [19], [20]. For instance, rural environments typically consist of large expanses of tree clusters and relatively narrow roads and rivers, while urban contain orderly arranged trees and wider roads and rivers [21], [22]. Therefore, the abilities to obtain multi-scale image features and distinguish different shapes of the same objects are crucial for HRS image segmentation networks.

Another reason is that, due to the complex background of the HRS images, objects belonging to different categories can

Y. Chen and X. Zhong are with the School of Mechanical and Electrical Engineering, the Chengdu University of Technology, Chengdu 610000 (email: chenyuxia@stu.cdut.edu.cn, zhongxl@cdut.edu.cn). P. Fang is a PhD student at the University of Southampton (email: P.Fang@soton.ac.uk). Y. Jianhui is with the School of Computer Science, University of Sydney, Sydney, Australia (e-mail: jianhui.yu@sydney.edu.au). Z. Xiaoming and L. Tianrui are with the School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China (e-mail: zxmswjtu@163.com; trli@swjtu.edu.cn)
* Equal Contribution.
† Corresponding author.

have a similar appearance, such as flowing streams and narrow roads. Although these diverse complex scenes contribute to richer details, inter-class similarities can easily lead to model error segmentation, and severely impact the performance of semantic segmentation networks [23], [24].

To address the first mentioned issue of scale variation in HRS images, some work [25]–[27] increases the size of the receptive field by introducing the adaptive spatial pooling module, thus capturing features at different scales. Unfortunately, performing a one-to-two feature aggregation at the end of each block often loses spatial information about the features. [28] obtains the feature maps of the low, medium, and high scales in the first convolution of the network, and forms the dense connection modules along the diagonal. However, this approach of consecutive downsampling may lead to feature loss, and the process of dense connection also has the possibility of network structure redundancy and information blocking. In contrast to the aforementioned methods, this paper proposes the funnel module and multi-branch module. The original image passes through the inverted bottleneck (IB) block in the funnel module to obtain reliable high-resolution information. In the multi-branch module, new scale information is obtained by gradual subsampling, and features at different scales are extracted in parallel, forming an efficient and direct feature extraction convolutional stream. At the end of each feature extraction, the feature information from the previous branch was fused with the newly generated branch information. Through multi-scale information interaction,the entire network is able to obtain sufficient complete and reliable low-resolution information while maintaining high resolution. Furthermore, due to the parallel architecture of Hi-ResNet is similar to HRNet [29], we visualize the feature maps extracted from both the baseline (which shares the same architecture as HRNet) and Hi-ResNet base to illustrate their differences. The results are shown in Figure 1. Note that there are only architectural differences between Hi-ResNet base and the baseline, which use exactly the same base blocks stacked by two stride-1, 3×3 convolutions. In comparison to the baseline, our proposed model eliminates the superfluous fourth stage, while simultaneously increasing the depth of the third stage architecture. Obviously, the image features extracted by Hi-ResNet base are far beyond the image features extracted by baseline.

At the same time, in order to alleviate the issue of class distribution disparity, a common approach is to incorporate attention mechanisms into the network. For instance, [30]–[32] utilize spatial attention to optimize class weights and address class imbalance problems. Additionally, [33] and [34] employ parallel channel attention and spatial attention to enhance local features simultaneously. On the other hand, with the proposal of vision transformer (ViT) [35], many subsequent works choose to apply Multi-Head Self-Attention (MHSA) in CNN-based networks [36]–[38]. However, MHSA demands substantial computational resources when the resolution and channels number of the input feature map are large. Several studies attempt to address this issue by using local window attention [39] or reducing the input feature resolution [37], [40]. However, for HRS images, using such modules remains a challenge. Therefore, we present a more lightweight and efficient information aggregation (IA) block. This base block uses window-based multi-head self-attention, performing sliding operations on the channel dimension of the feature graph to capture global contextual information. At the same time, it also applies Squeeze-and-excitation (SE) attention [41] to provide richer location information for the network. The IA block integrates the advantages of both convolution and MHSA, allowing it to aggregate different shapes of the same class, thereby reducing the intra-class distribution distance. Meanwhile, The use of depth-wise separable convolution in IA block reduces the parameter count of Hi-ResNet by fifty percent.

To mitigate the second mentioned of increased error segmentation due to diversification of background in HRS images and to enhance object boundary information, [42] proposes a dual-stream network (one network for segmentation and another for boundary enhancement), designing an independent network to extract boundary information and improve segmentation results. In contrast, we propose a feature refinement module and a class-agnostic edge-aware (CEA) loss, which focuses on module and loss level. Feature refinement module upsamples the three feature maps with different resolutions obtained by muti-branch module to the same size, and concats them into a feature map. By a simple classification convolution and object-contextual representations (OCR) [43], we obtain the results of coarse and refined segmentation of Hi-ResNet respectively and then compute them into the loss function. The outcomes of the loss function will be mixed with a ratio of 1:1. For the design of the loss functions in the HRS task, some work [44], [45] employs dice loss in road extraction by increasing the weights of the key road regions, FactsegNet [46] utilizes collaborative probability loss to merge the outputs of the dual-branch decoders at the probability level, aiming to enhance the utilization of information. Unlike the above losses, the proposed CEA loss in this paper focuses more on the edge information of class objects. The CEA loss expands the original Hausdorff distance (HD) loss [47] to multi-classes and reduces computing resource consumption. This correction of CEA at the edge level improves the model's perception of boundaries and shapes, enhancing its ability to capture accurate object edges. Finally, we evaluate the proposed method on widely used datasets. This study contributes four main points:

(1) We propose the funnel module to reduce computing costs, efficiently extract high-resolution information, and avoid feature loss from the input image.
(2) We apply our proposed IA block to a multi-branch module, integrating the dynamic global modeling capability of the Transformer into CNN-based networks.
(3) We develop the CEA loss, which emphasizes edge information while taking into account multiple classes.
(4) Our Hi-ResNet is validated on several benchmarks with performance better than existing prevalent methods.

The paper is organized as follows: Section II provides an overview of related work, including Semantic Segmentation in Remote Sensing, Attention Mechanisms and Model Pre-training. In Section III, we describe the proposed method,

which includes the Hi-ResNet model, the design of loss functions and the use of unsupervised and supervised pre-training in HRS tasks. Section IV presents a series of ablation experiments, and experimental results and analyses on different datasets. Finally, in Section V, we conclude the paper and provide a summary.

II. RELATED WORK

A. Semantic Segmentation in Remote Sensing

Parallel multi-resolution architectures primarily focus on high-level semantic information, resulting in a semantically richer and spatially more accurate representation, providing an advanced technical reference for HRS semantic segmentation tasks. Among them, HRNet [29] was well known as a parallel semantic segmentation model that could maintain high resolution. It passes through four stages of gradually decreasing resolution and performs multi-scale fusion to enhance high-resolution representations. Subsequently, researchers attempted to combine HRNet with object-contextual representations [43] which distinguishes contextual information for the same target category from different target categories and optimizes feature pixels. This architecture was widely applied in the field of HRS segmentation [48]. However, HRNet primarily focused on high-resolution semantic features of images, while object-contextual representations was more concerned with the relationships between image objects and their pixels, both of which ignore the high-level semantic information. Unfortunately, reliable high-level semantic information that includes target locations is undoubtedly crucial, as HRS images often contain a large amount of complex and unrelated background, and small target objects usually occupy only a few pixels. To obtain richer high-level semantic information, some studies [15], [49]–[51] applied different dilated convolutions to multiple features of traditional CNN networks. By expanding the receptive field of the convolution kernel, these studies have constructed distinctive local semantic representation modules, thereby effectively utilizing multi-scale features. Furthermore, some researchers [52], [53] applied graph convolution on multi-layer features, treating each pixel as a node, and then connecting the extracted graph features with the final global visual features. Despite this, locally aggregating features in the spatial direction might overlook channel and positional information of high-level semantics. An effective solution is to establish an information connection between space and channels in convolutional networks. Recently, MBFANet [54] combined the pooling channel attention module and convolutional coordinate attention module to complement each other, which helped the models focus on more complex background categories. SAPNet [55] joint models both spatial and channel affinity, which allows for preserving spatial details and extracting accurate channel information. Inspired by the parallel architecture of HRNet, we propose Hi-ResNet in this work. The Hi-ResNet utilizes a funnel module and muti-branch module to obtain rich high-resolution semantic information and use feature refinement module to enhance small target features.

B. Attention Mechanisms

Attention mechanisms could help the network to locate the information of interest and inhibit useless information, which has been widely used in convolutional neural networks [13], [41], [58]–[60]. For HRS tasks, some studies utilized the popular attention mechanism Squeeze-and-Excitation to automatically process the features of various scenes and extract more effective features [45], [61]–[64]. However, this attention mechanism only focused on inter-channel information while neglecting spatial and positional information about the features. To simultaneously capture channel and positional information, researchers explored the use of Convolutional Block Attention Module or Bottleneck Attention Module in network architectures [65]–[67]. These modules used spatial attention to obtain the location information and reduce the input channel dimension to save the calculation cost. Due to the limited receptive field of the sliding window in convolutional operations, only local relationships were captured, it could not maintain long-range dependencies between different positions in the image.

Currently, some works apply the Transformer to semantic segmentation models in HRS, thereby obtaining global information [17], [68], [69]. DC-Swin [70] introduced the Swin-Transformer for the encoder in fine image segmentation, while Unetformer [71] uses Unet as an encoder, proposing a global-local attention mechanism to construct Transformer blocks in the decoder. Nevertheless, the substantial computational complexity introduced by self-attention made the training cost of the network expensive. This poses challenge for its application in lightweight convolutional networks. To transfer the dynamic global modeling capability of the Transformer to CNN-based networks while keeping the network lightweight, we propose the efficient IA block. This block combines the long-range interaction capability of the Transformer with the inductive bias of CNNs. By using window-based multi-head self-attention and SE attention, it provides accurate feature information for the model. Meanwhile, the use of depth-wise separable convolution brings less computational quantities and parameters for IA block.

C. Model Pre-training

In addition to the design of the network itself, excellent pre-training programs are also indispensable. Numerous studies showed that applying pre-training can make models more stable and extract more commonalities [72]–[75]. Therefore, we pre-train the Hi-ResNet to enhance the fine-tuning ability for HRS segmentation tasks. Recently, for HRS tasks, some studies [76], [77] used labeled semantic segmentation datasets such as Mapillary [78] for pre-training to improve model performance. However, these large-scale labeled datasets mostly come from natural images, and pre-training on them for HRS tasks often yields poor results. It is worth noting that recent works on unsupervised pre-training [79]–[82] showed that unsupervised pre-training outperforms the supervised way in downstream tasks such as segmentation. MoCo [82], as a mechanism for building dynamic dictionaries for contrastive

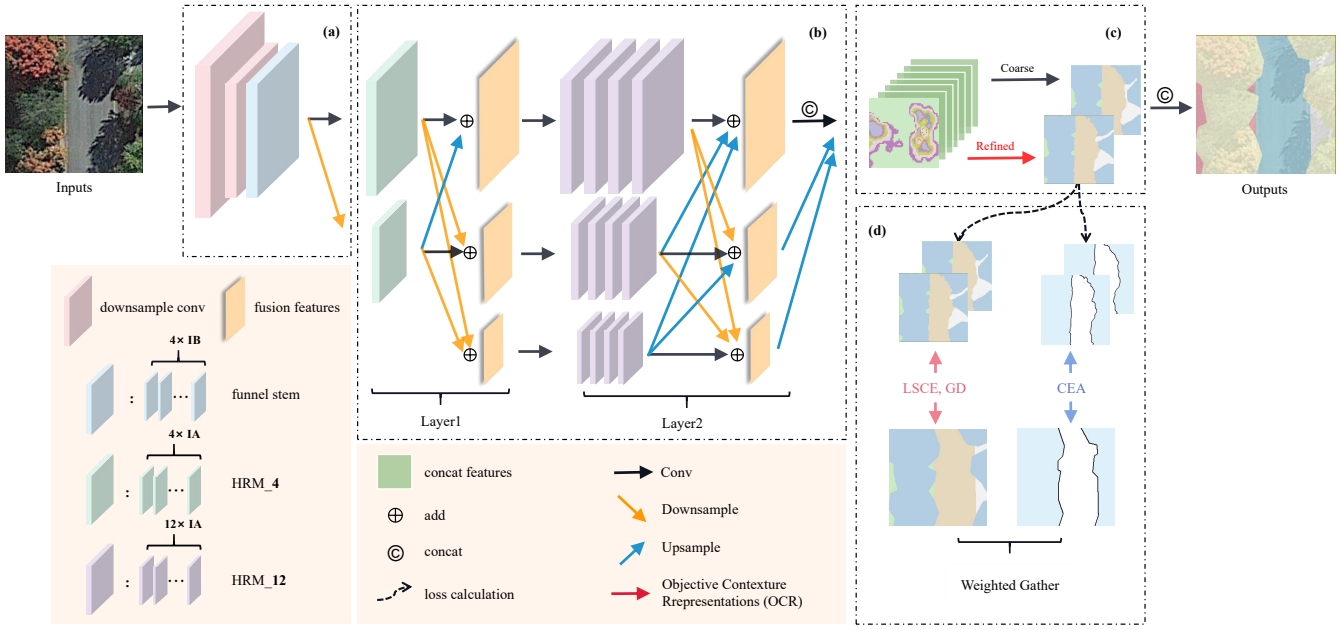


Fig. 2. The comprehensive architecture of Hi-ResNet is partitioned into four components. (a) The funnel module, composed of a downsample part and a funnel stem, is proposed for downsampling input imagery and facilitating feature extraction. (b) The multi-branch module further hones these features via the amalgamation of a multi-resolution convolutions stream. (c) In the feature refinement module, coarse features are computed directly via a convolution layer, with refined features managed through the utilization of OCR [43]. During inference, the coarse results and refined results are added in a 1:1 ratio as the model's output. (d) Multiple loss functions are employed, including LSCE loss [56] and GD loss [57], which are computed in direct relation to the ground truth and predictions. Concurrently, the CEA randomly elects a category, designating all others as background, computing the loss between the two categories.

learning, surpassed its supervised counterpart in seven downstream tasks. [83] illustrated that MoCo mainly transferred low-level and middle-level semantic features, and when performing image reconstruction, the reconstructed images without supervision were closer to the original data distribution. Based on the previous work, we argue that employing supervised pre-training will provide richer and more comprehensive prior information for HRS tasks. At the same time, using the pre-training mechanism of MoCo can effectively compensate for the loss of precise localization information in the network and reduce the emphasis on local object information. Therefore, in this study, we apply both fully supervised and unsupervised pre-training strategies on Hi-ResNet and evaluate the performance of the two methods.

III. PROPOSED METHODS

In this section, we present the framework of Hi-ResNet, including the funnel module, the multi-branch module with information aggregation blocks, and feature refinement module. Then we introduce the class-agnostic edge aware loss for HRS image feature extraction. Finally, we present how to transfer the various pre-training strategies to the HRS segmentation task.

A. Hi-ResNet Framework

The Hi-ResNet proposed in this paper is shown in Figure 2. In the following sections, we will present the funnel module, multi-branch module, and feature refinement module, and the implementation details of each module in turn.

1) Funnel Module: In the funnel module, we start by passing the input image through two stride-2, 3×3 convolutions, which reduce the image resolution to 1/4 of its original size. During the network downsampling, the batch normalization (BN) layer was placed before the convolution operation. It could improve the generalization and stability of the model by applying the BN layer, which makes the pre- and post- samples to different Gaussian distributions. Then, the image goes through a funnel stem with four inverted bottleneck (IB) blocks to obtain high-resolution semantic features. The traditional bottleneck block uses a structure with long heads and a short middle. With consideration of the distribution characteristics of HRS data, to prevent the collapse of activation space and loss of channel information caused by non-linear activation functions in network layers [84], our work adapts the IB block with thin heads and a thick middle. This block is used to extract richer semantic features by performing high-dimensional upsampling on HRS images, followed by residual connection and linear activation function to avoid information loss, thereby preserving more complete information of HRS images. The design of the funnel module is illustrated in Figure 3. For IB block, we use a stride-1, 3×3 convolution in the first layer. In the middle layer, we apply a stride-1, 1×1 convolution to quadruple the number of channels, thus obtaining richer high-resolution semantic information. Subsequently, a stride-1, 1×1 convolution is used in the final layer to revert the channel count to its original number.

2) Multi-branch Module: The multi-branch module consists of multi-resolution convolution streams and repeated feature fusions. First of all, to address the issue of unstable

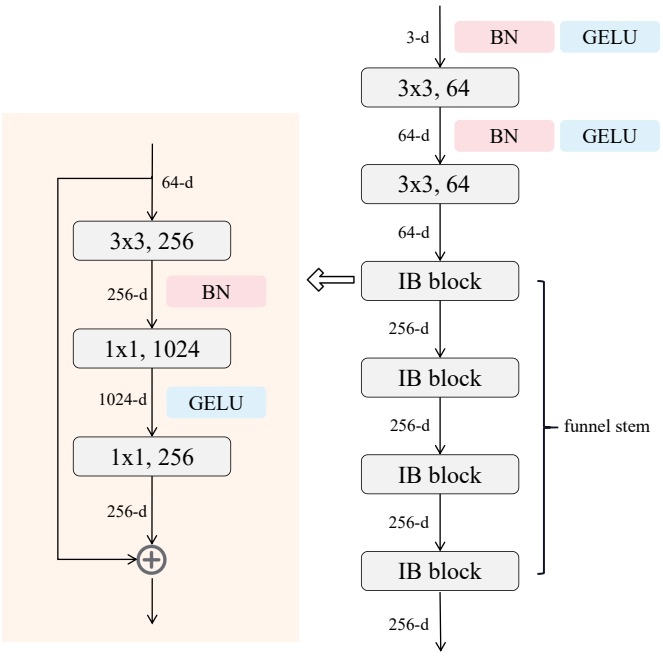


Fig. 3. The structure of the funnel module where IB refers to inverted bottleneck. The number in each block refers to the kernel size and channel numbers respectively.

segmentation accuracy caused by differences in image scales, our network maintains the high-resolution representation of the input image throughout generating a new low-resolution branch at each end of the layer.

We employ the parallel approach to conduct a series of convolution operations in multi-resolution branches, forming the multi-resolution convolution stream. Notably, the minimum resolution of the image in the parallel branch of the second layer is only 1/16 of the original image, indicating that this layer focuses more on the high-level semantic information of the image. Due to the significant layout differences among objects in different areas of HRS images, the shape and contour features in high-level semantic information are crucial. This paper argues that it cannot extract rich high-level semantic features that contain target locations if merely stacking the same number of blocks as in other layers and using the same sliding window sampling. Therefore, we stack 4 IA blocks as high-resolution module_4 (HRM_4) in the first layer, and triple the number of IA blocks in the second layer, i.e., 12 IA blocks as high-resolution module_12 (HRM_12). Figure 4 illustrates the semantic information extracted by the multi-branch module before and after the extension. More abundant and reliable high-level semantic information can be obtained through the second layer after expansion, which not only effectively alleviates the problem of class distribution inconsistency and reduces intra-class variance but also avoids the loss of positional information of small target objects that occupy only a few pixels in the image, thereby enhancing the weak features of small target objects.

Numerous studies propose methods for multi-scale feature fusion [12], [50], [85]. Classic semantic segmentation networks like UNET [86] and SegNet [87] extract feature maps

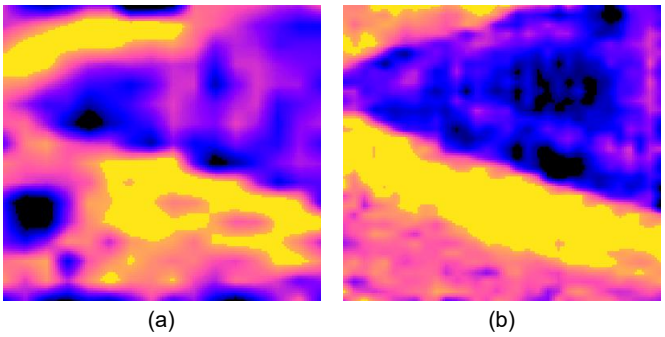


Fig. 4. (a) the output features of the multi-branch module in Hi-ResNet before the extension. (b) the output features of the multi-branch module in Hi-ResNet after the extension.

of different resolutions during the downsampling phase. In the model's upsampling phase, these are combined with feature maps of the corresponding resolution, serving to prevent feature loss. In contrast, our approach performs cross-layer fusion between parallel branches with different resolutions, capturing features of different sizes by repeatedly exchanging information on different scales at each layer. Figure 5 illustrates the fusion process for layer2, where the input consists of three images with different resolutions. Different sampling methods are used depending on the resolution of the input and output. The upsampling stage includes bilinear upsampling, BN layer, and a stride-1, 1×1 convolution, while the downsampling stage includes BN layer and a stride-2, 3×3 convolution. We sum the images sampled at the same resolution to produce the final output for that resolution. The multi-branch module process ultimately outputs three feature maps with different resolutions.

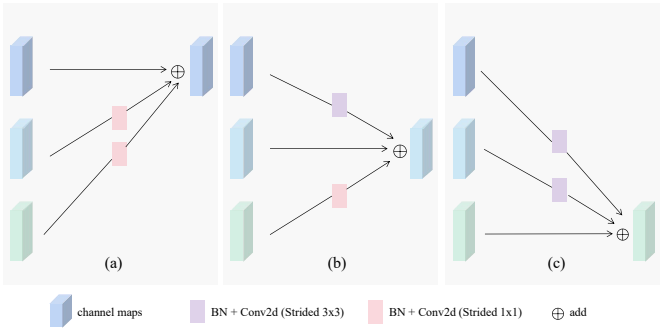
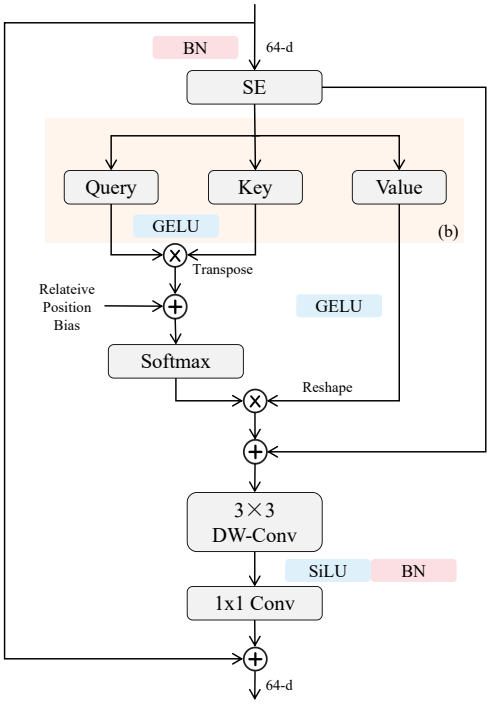


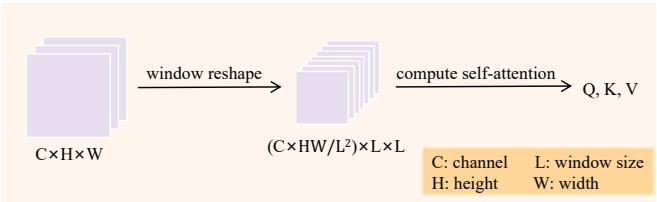
Fig. 5. This figure illustrates the process of feature information aggregation across various resolutions in the fusion layer of the network. Furthermore, we exchange the sequence of the BN and the conv here.

3) Information Aggregation Block: HRS images provide rich details and features but also bring more irrelevant background objects. To suppress the impacts brought by the irrelevant background information and to enhance the spatial and positional feature representations, we propose a lightweight block: Information Aggregation (IA) block. This block refers to few parameters, easy operators, and high efficiencies. In IA block, the SE attention [41] is utilized to sufficiently sketch the hidden information from the input features. Then, we consider that the attention mechanism shares a wider perception field

than the convolutional kernel, thereby multi-head self-attention (MHSA) applied in the block. However, MHSA consumes huge computation resources, especially in image calculations. Therefore, Window-MHSA (WMHSA) and Depth-Wise Convolution (DW-Conv) both with a skip connection are utilized to trade-off model cost and accuracy. Unlike sliding windows of the Swin-Transformer [39], the WMHSA here simply resize the tensors from $C\times H\times W$ to $(C\times H\times W/L^2)\times L\times L$ (shown in Figure 6(b)) and then conduct MHSA. We employ another skip connection acting on the whole block, which enables feature reuse and prevents loss. In the IA block, instead of RELU, we prefer GELU and SILU to obtain a relatively slight change during the minus. Moreover, different from general convolutional blocks, we utilize fewer activation functions and normalizations. The IA block is illustrated in Figure 6.



(a) The structure of Information Aggregation Block



(b) Illustration of the window partition operation

Fig. 6. The IA block mainly consists of two attentions: a convolutional self-attention and a SE attention, a Depth-Wise convolution with a relatively large kernel, and a 1×1 convolution.

4) Feature Refinement Module: We use the three different resolution feature maps output by the multi-branch module as inputs to the feature refinement module. In the feature refinement module, we combine the three input images to the same size using bilinear upsampling, which serves as the coarse segmentation of the network. By leveraging OCR [43], we first treat a category in the coarse segmentation result as a region and estimate the comprehensive feature representation within that region by aggregating the representations of each pixel. Then, we compute the pixel-region relationships to obtain corresponding weights, which are used to enhance the representation of each pixel by weighting all the regions. The weighted feature representation serves as the refined segmentation result of the model. Lastly, Hi-ResNet outputs both coarse segmentation and refined segmentation.

TABLE I
THE MAIN ARCHITECTURE CONFIGURATION OF HI-RESNET

Input Size	Funnel Module	Multi-branch Module	
		Layer1	Layer2
$H\times W\times 3$	$[3\times3, stride=2]\times2$ $[IB\ Block]\times B_1$		
$\frac{4}{H}\times\frac{4}{W}\times C_1$		$[IA\ Block]\times B_2\times M_1$	$[IA\ Block]\times B_3\times M_2$
$\frac{8}{H}\times\frac{8}{W}\times C_2$		$[IA\ Block]\times B_2\times M_1$	$[IA\ Block]\times B_3\times M_2$
$\frac{16}{H}\times\frac{16}{W}\times C_3$			$[IA\ Block]\times B_3\times M_2$

TABLE II
THE CONFIGURATIONS OF HI-RESNET INSTANCES

Model	Channels (C_1, C_2, C_3)	Blocks (B_1, B_2, B_3)	Modules (M_1, M_2)
Hi-ResNet	(48, 96, 192)	(4, 4, 12)	(1, 4)

The main architecture configuration of Hi-ResNet is shown in Table I. Here, H and W denote the height and width of the input image, respectively. C1, C2, and C3 denote the number of channels. B1, B2, and B3 denote the number of blocks on each branch of the module, respectively. M1 and M2 denote the number of modules in each layer, respectively. Table II displays the detailed configuration of (C1, C2, C3), (B1, B2, B3) and (M1, M2) instances in Hi-ResNet.

B. Loss Design

In HRS tasks, semantic segmentation typically involves more than two labels, with significant differences in the number and pixel range of objects for different categories, leading to sample imbalance and sub-optimal performance. Therefore, the appropriate loss function is crucial. In this work, we propose a new class-agnostic edge aware (CEA) loss, which is combined with the Generalised Dice loss (GD) [57] and Label Smoothing Cross-Entropy loss (LSCE) [56] as the training loss.

1) Generalised Dice Loss: Weighted cross-entropy and Sensitivity-Specificity approaches are designed to address imbalanced problems only in binary classification tasks. In contrast, the GD loss method can weight various pixel classes, allowing for a more comprehensive approach to imbalanced sample issues. The loss calculation for GD loss can be expressed as:

$$L_{GD} = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n r_{ln} + p_{ln}} \tag{1}$$

The equation for GD loss involves using r_{ln} to represent the label of each pixel in the reference foreground segmentation for class l, and p_{ln} to denote the predicted probabilistic map for the foreground label of class l over N image elements p_n. The weighting factor w_l is used to provide invariance to different label set properties. Its calculation is expressed as:

$$w_l = \frac{1}{\sum_{i=1}^{N} r_{ln}^2} \tag{2}$$

During the calculation process, overlapping r_n and p_n for each category l are added according to their weights and then divided by the weighted sum of the union part. This effectively suppresses the interference of complex background classes, enhances the features of small targets, and alleviates the problem of imbalanced image samples.

2) Label Smoothing Cross-Entropy Loss: The label smooth technique proposed in [56] as a training strategy can adjust the extreme values of the loss and improve the model's generalization ability when combined with Cross-Entropy loss. The equation of Label Smoothing Cross-Entropy loss is as follows:

$$\mathcal{L}_{lsce} = -\sum_{k=1}^{K}\sum_{n=1}^{N} y_k^{(n)} \log \hat{y}_k^{(n)} \tag{3}$$

$$y_k^{(n)} = \begin{cases} 1-\varepsilon & \text{if } n = k, \\ \varepsilon/(K-1) & \text{otherwise} \end{cases} \tag{4}$$

In the above equation, N and K indicate pixel values and categories, respectively. $y_k^{(n)}$ and $\hat{y}_k^{(n)}$ represents the sample label following the label smoothing operation and the corresponding softmax output belonging to the category k, respectively. Equation 4 shows the calculation process of $y_k^{(n)}$. When the pixel representation class n is the same as the input class k, $y_k^{(n)}$ equals $1-\varepsilon$, where ϵ is the smoothing factor. Otherwise, $y_k^{(n)}$ equals $\varepsilon/(K-1)$, where K is the total number of classes.

Considering that HRS image datasets usually have a small amount of data, we argue that using this loss can prevent overfitting of the network and provide the correct optimization direction for the model.

3) Class-agnostic Edge Aware Loss: Our proposed CEA loss enhances the original Hausdorff distance (HD) [47] loss in two crucial stages. Initially, we extend the Hausdorff loss to accommodate multiple classes. Subsequently, the HD loss utilizes the Scipy library to compute the Euclidean distance transform, an approach that has proven to be inefficient in the context of multi-class loss calculation. Hence, we employ cascaded convolutional operations to approximate the Manhattan distance transform of images, thereby addressing the observed inefficiency. It is shown below:

$$\mathcal{P} = Softmax(s_\theta(p)) \tag{5}$$

$$\mathcal{T} = Onehot(t) \tag{6}$$

$$\mathcal{L}_{CEA} = \int_{\Omega} (\mathcal{T}-\mathcal{P})^2 (D_G(\mathcal{T})^\beta + D_S(\mathcal{P})^\beta) \mathrm{d}\mathcal{P} \tag{7}$$

We get $\mathcal{P}$ from Equation 5 and $\mathcal{T}$ Equation 6, where p refers inputs of our model s_θ, and t refers the ground truth. In Equation 7, Ω denotes the spatial domain of the training images. The distance function from the predicted boundary S, after applying thresholds s_θ, is represented by D_S. The hyper-parameter β, set to 2 by the authors of [47] through a grid search, is also a part of this process.

4) Loss function: According to the results of the ablation study, the overall loss can be formulated as:

$$\mathcal{L}_{coarse/refined} = \alpha\mathcal{L}_{GD} + \beta\mathcal{L}_{LSCE} + \gamma\mathcal{L}_{CEA} \tag{8}$$

$$\mathcal{L} = \mathcal{L}_{coarse} + \mathcal{L}_{refined} \tag{9}$$

where $\mathcal{L}_{coarse}$ and $\mathcal{L}_{refined}$ represent the coarse segmentation and refined segmentation of the model, respectively. Both $\mathcal{L}_{coarse}$ and $\mathcal{L}_{refined}$ are calculated by Equation 8, where α, β, γ denote the weights of each loss. In our model, they are set to 0.3923, 0.3923, 0.2153 respectively.

C. Remote Sensing Pre-training

Given that this paper introduces a new backbone, it is necessary to employ different strategies for pre-training of Hi-ResNet. In this section, we demonstrate two pre-training strategies designed for the HRS semantic segmentation task.

1) Dataset:

- The Mapillary dataset [78] is presently the largest publicly available street view dataset, with specific instance annotations and a high degree of diversity. This dataset encompasses 25,000 high-resolution RGB images, captured by a variety of imaging devices, and includes fine-grained labels for 66 categories.
- Million-AID [88] is a comprehensive benchmark dataset designed for remote sensing scene classification. This dataset obtains images with resolutions ranging from 0.5m to 153m from multiple satellites of Google Earth. The scene labels are obtained through the geographical coordinate information, resulting in over one million images labeled with 51 semantic scene categories.

2) Pre-training Details: The section introduces two distinct methods for pre-training models. For the supervised training, the Mapillary dataset is selected as it shares the same downstream task as this paper. We got 2 million 256×256 images after clipping images of Mapillary. To address the issue of imbalanced data, we filter the cropped images based on the proportion of pixel classes within each labels, ultimately resulting in 400,000 images for training. The supervised pre-training process is the same as for training our own model.

The unsupervised training utilizes the Million-AID dataset. Since the images in Million-AID have varying resolutions, they are partitioned into 400×400, while images size less than 400 are dropped from the dataset. We use contrast learning MoCoV2 [82] as the unsupervised pre-training method. The process of MoCoV2 is shown in Figure 7, and the primary training settings for both pre-training approaches are presented in Table III.

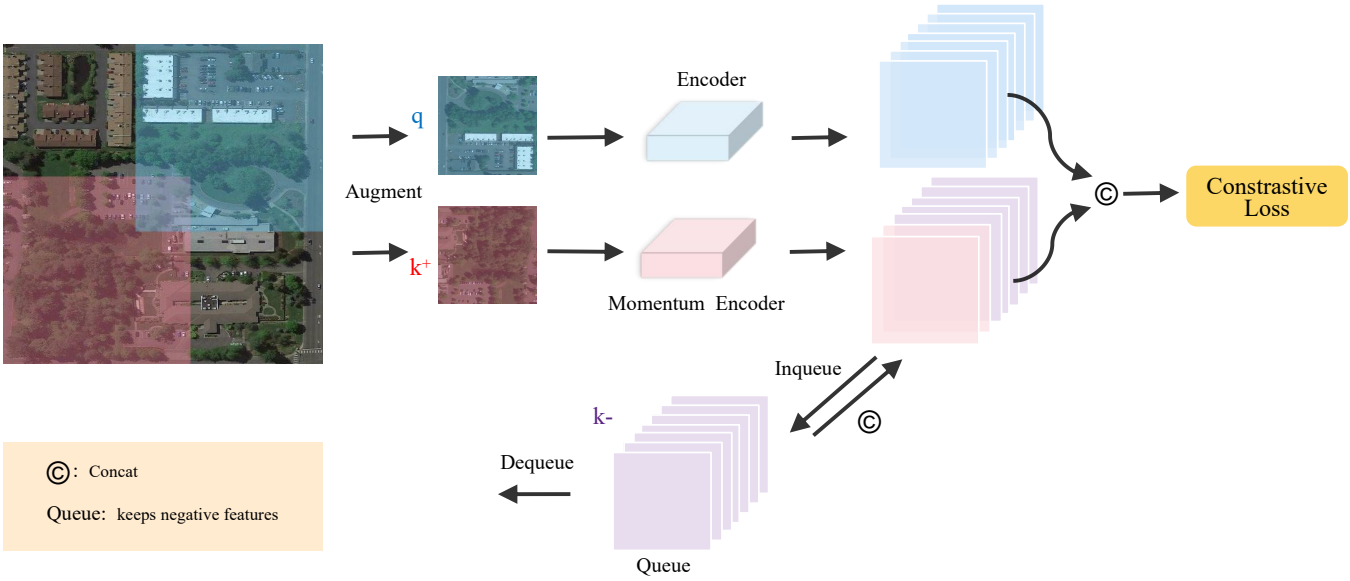


Fig. 7. The figure outlines the entire process of MoCoV2 pre-training. $\boldsymbol{q}$ and $\boldsymbol{k}^+$ are the positive sample and the negative respectively, augmented from the same image. while $\boldsymbol{k}^-$ refers to the past negative features stored in the queue.

TABLE III
HYPER-PARAMETER SETTINGS OF DIFFERENT PRE-TRAINED MODELS

Method	dataset	lr	image size	batch size	quantity
supervised	Mapillary	5e-5	256×256	80	400,000
MoCoV2	MillionAid	0.015	400×400	64	1,000,000

MoCoV2 pre-training commences with the augmentation of a batch of images twice, to generate the positive samples, denoted as q, and the batch negative samples, denoted as k^+. Following this, the logits of q and k^+ are procured by introducing q and k^+ into the standard encoder and the momentum encoder, respectively. In addition, the input negative samples are amalgamated with k^+ and k^-, which are retrieved from the queue. Subsequently, the logits of the positive and negative samples are concatenated, following which the InfoNCE loss function is computed to update the standard encoder:

$$\mathcal{L}_{q,k^+,k^-} = -\lg \frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum_{k^-} \exp\left(q \cdot k^-/\tau\right)} \tag{10}$$

where $\mathbf{q}$ is a query representation, $\boldsymbol{k}^+$ is a representation of the positive (similar) key sample, and $\boldsymbol{k}^-$ are representations of the negative (dissimilar) key samples. $\boldsymbol{\tau}$ is a temperature hyper-parameter. During training, only the normal encoder updates while the momentum encoder is updated with the function below:

$$\theta_k = m\theta_k + (1-m)\theta_q \tag{11}$$

Here $\mathbf{m} \in [0, 1)$ is a momentum coefficient. Only the parameters θ_q are updated by back-propagation. The momentum update in Equation 11 makes θ_k evolve more smoothly than θ_q. Finally, k^+ will be added to the queue, and features earlier in the queue will be dequeued.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we evaluate the performance of our proposed model on multiple remote sensing datasets, including LoveDA, Potsdam, and Vaihingen. We first conduct a series of ablation studies to analyze and identify a suitable framework for our proposed model. Next, we compare our Hi-ResNet with current state-of-the-art (SOTA) methods on public benchmarks. Additionally, we demonstrate the performance of our proposed model and existing popular frameworks in terms of computational complexity, inference speed, and memory usage on three datasets.

A. Datasets

1) LoveDA: The LoveDA dataset [89] comprises 5987 HRS images (GSD 0.3m) from three different cities, each containing 166768 annotated objects. Each image is 1024×1024 pixels and includes 7 land cover categories, namely building, road, water, barren, forest, agriculture, and background. The dataset provides 2522 images for training, 1669 images for validation, and 1796 official images for testing. The images consist of two scenes, urban and rural, from three Chinese cities, namely Nanjing, Changzhou, and Wuhan. Consequently, the dataset presents a significant research challenge due to the presence of multi-scale objects, complex backgrounds, and inconsistent class distributions.

2) Potsdam: Potsdam is a historic city with complex buildings, narrow streets, and dense settlement structures. The Potsdam dataset is composed of 38 images, each size is 6000×6000, and containing a true orthophoto (TOP) extracted from a larger TOP mosaic. The dataset has been manually classified into the six most common land cover categories (impervious surfaces, background, buildings, low vegetation, trees, and cars) , and the ground sampling distance of the TOP is 5cm. In this paper, we follow the approach used in [71] and

use 23 images (excluding image 7_10 with error annotations) for training and 14 images for Validation.

3) Vaihingen: The village of Vaihingen comprises many individual buildings and small multi-story houses. This dataset includes 33 HRS images, and the average size of images is 2494×2064. Each image consists of three channels: near-infrared, red, green, and a single-band DSM (note that the DSM is not used in our experiments). All images are labeled into the same six classes as the Potsdam dataset. For the experiment, we follow [71] to select the remote sensing images with ID 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38 for Validation, while the remaining 16 images are used for training. Table IV provides detailed information about each dataset.

TABLE IV
THE DETAILS OF DIFFERENT SEMANTIC SEGMENTATION DATASETS.

Datasets	Training	Validation	Testing	Category	Input Size
LoveDA	2,522	1,669	1,796	7	1,024×1,024
Potsdam	24	-	14	6	6,000×6,000
Vaihingen	16	-	17	6	2,494×2,064

B. Implementation Details and Evaluation Metrics

1) Implementation Details: For the LoveDA dataset, the training and validation sets are both used for training. These images are cropped into patches with 512×512 resolution for input. During training, various enhancement techniques such as random vertical flip, random horizontal flip, and random scaling with ratios of [0.5, 0.75, 1.0, 1.25, 1.5] are employed. The training process last for 400 epochs with a batch size of 16. During the testing phase, we use 1796 images provided by the official for prediction. As for the Potsdam and Vaihingen datasets, the images are cropped into 512×512 for model input. The training epoch set to 200 with a batch size of 16.

In the experiment, learning rate warmup combined with cosine annealing is used to adjust the learning rate, where warmup set to 3 epochs. Moreover, AdamW [90] optimizer was selected to accelerate model convergence, with the learning rate and weight decay set to 1×10^{-4} and 1×10^{-8} respectively. All models are trained using NVIDIA GTX 3090 GPUs and implemented on the PyTorch framework.

2) Evaluation Metrics: We evaluate models based on two aspects: accuracy and performance. In terms of model accuracy, we employ metrics including the mean F1 score (F1), overall accuracy (OA), and mean intersection over union (mIoU). Regarding performance, we assess the model size using the number of model parameters (M), the model complexity through GPU memory usage (MB) and the floating point operation count (FLOPs), and the inference speed by measuring frames per second (FPS).

C. Ablation Study and Comparison Experiments

1) Architecture Analysis: To determine the foundational network architecture of Hi-ResNet, we conduct a series of ablation studies on the Vaihingen dataset. For fair comparison, all ablation studies share the same settings, except for the experimental variable. Table V shows the experimental setup for architecture analysis.

TABLE V
EXPRIMENTAL SETUP FOR THE DIFFERENT ARCHITECTURE

Method	Multi-branch Module			
	Stage1	Stage2	Stage3	Stage4
HRNet	$[Block] \times^{1} B_4$	$[Block] \times B_4 \times^{2} M_1$	$[Block] \times B_4 \times M_4$	$[Block] \times B_4 \times M_3$
V1	[3]		$[Block] \times \mathbf{B_{12}} \times M_4$	
V2				$[Block] \times \mathbf{B_{12}} \times M_3$
Hi-ResNet base			$[Block] \times B_{12} \times M_4$	✘

[1] 4 blocks are represented by B_4.
[2] 1 Module are represented by M_1.
[3] Blank refers to keep the same settings of the baseline.

Given that the overall structure of this paper inspires from HRNet [29], the baseline architecture is the same as the original HRNet, which consists of four stages. These four stages contain 1 to 4 branches respectively, each branch is composed of several high resolution modules with stacks of bottleneck blocks or basic blocks. Apart from the baseline, we set V1 to extends stage3 threefold, while V2 tripled stage4. Finally, our unique Hi-ResNet base not only triples the length of stage3 but also eliminates stage4 entirely.

Table VI displays the results of the ablative experiments. Surprisingly, although extending stage4 leads to more than doubling the number of network parameters, the mIoU increased by only 0.8%, which is less than the mIoU obtained by extending stage3 with less increase in parameters. Through sampling analysis of the original stage3 and stage4 feature output images, we argue that there is some feature loss when extracting information in stage3, thereby reducing the medium and low-resolution semantic information that stage4 can acquire. Lengthening stage3 effectively addresses this issue, allowing for the extraction of richer and more accurate spatial information and better fitting of the features of HRS images in tasks. Consequently, we decide to lengthen stage3 by three times. After determining the stage size, we attempt to eliminate redundant stage within the framework. We argues that lengthening stage3 can fully encompass the information extracted by stage4, so we try to remove stage4. This decision significantly reduces the number of parameters and FLOPs, speeds up the training process, and further enhances the model's efficiency and accuracy.

TABLE VI
RESULTS OF THE ARCHITECTURE ANALYSIS EXPERIMENTS ON VAIHINGEN DATASET

Method	Params (M)	FLOPs (G)	Memory (MB)	mIoU
HRNet	68.6	140	**1569**	69.3
V1	96.5	205	2177	70.4
V2	153.3	214	2379	70.1
Hi-ResNet base	**46.5**	**139**	2116	**70.8**

Ultimately, we expanded the third stage of HRNet threefold and removed the fourth stage to establish the foundational architecture for Hi-ResNet. Different from the four-stage network architecture composed of HRNet with a fixed number of

convolutions, our network extends specific stages to ensure it acquires more high-level semantic information. Compared to the baseline, Hi-ResNet base achieves more than 1.5% increase in mIoU on the Vaihingen dataset and reduces the number of parameters by 30%. In addition, in order to align the various modules proposed in Hi-ResNet, we renamed stage1 as the funnel stem, and stage2 and stage3 are referred to as layer1 and layer2, respectively.

2) Module Analysis: To evaluate the performance of each proposed component in Hi-ResNet, we cast ablation experiments on our modified modules such as Funnel Module, Multi-branch Module, Feature Refinement Module, and CEA loss, which bases on the Hi-ResNet base model. To prevent overfitting during the training process, all ablation studies employ both GD and LSCE loss, with a balanced weighting ratio of 1:1.

TABLE VII
RESULTS OF THE MODULE ANALYSIS EXPERIMENTS ON VAIHINGEN DATASET

Method	Funnel	Multi-branch	Feature Refinement	CEA	Params	mIoU
Hi-ResNet base					46.5	71.0
Hi-ResNet v1	✔				50.2	72.2
Hi-ResNet v2	✔	✔			**22.3**	74.7
Hi-ResNet v3	✔	✔	✔		26.0	75.2
ours	✔	✔	✔	✔	26.0	**76.2**

As shown in Table VII, Hi-ResNet base acquires only 71.0% of mIoU, indicating a limited ability to segment HRS images. The addition of the funnel module increases model precision by 1.2%, demonstrating its effectiveness in capturing more high-resolution semantic information during the down-sampling process. Notably, the muti-branch module greatly reduces the number of parameters of the model (more than 50 percent), while providing a significant increase of at least 2.5 % for the model. This fully demonstrates the excellent performance of IA block, which can effectively suppress the interference of irrelevant background and extract more accurate feature information. This also illustrates the effectiveness of feature fusion. By feature refinement module, the model outputs both coarse and refined segmentations. However, the model's mIoU only increases by 0.5%. This is because the combination of GD and LSCE loss has difficulty in utilizing the two outputs of the model. Therefore, we add the CEA loss to the model training to form the final Hi-ResNet. Without any increase in additional parameters, CEA loss improves model precision by 1%, proving its capacity to balance model outputs effectively and its compatibility with the model. Compared to Hi-ResNet base, the final Hi-ResNet achieves 76.2% mIoU without pre-training, demonstrating the effectiveness of each components of the model.

3) Loss Analysis: We use LSCE, GD, and CEA loss for training the Hi-ResNet, aiming to improve the accuracy and generalization ability of the model. Nevertheless, imbalanced weights among the various loss functions could potentially instigate gradient conflicts and destabilize the training process. Therefore, to eliminate this instability, we conduct ablation studies on the weights between the loss functions across the Potsdam, Vaihingen and LoveDA datasets. All loss calculations in ablation studies were performed exclusively with the results from the refined segmentation.

We sum the GD, LSCE, and CEA loss with a default 1:1:1 ratio as V1. In the V2, we weight the LSCE, GD, and CEA loss in a 1:1:0.4 ratio to ensure that the values of the three loss functions are in the same scale. Subsequently, we select Random Weighting [91] as V3. This approach employs dynamic loss weights in each training iteration, samples loss weights from a potentially normalized distribution, and minimizes the aggregate loss weighted by these randomly sampled weights. Finaly, we take the softmax computation result of 1:1:0.4, i.e., (0.3923, 0.3923, 0.2153) as V4.

TABLE VIII
RESULTS OF LOSS ANALYSIS ON THE THREE DATASETS

Method	LoveDA	Potsdam	Vaihingen
V1	[1]49.7	81.8	75.1
V2	50.0	82.3	75.7
V3	49.8	82.0	75.3
V4 (Ours)	**50.1**	**82.5**	**75.8**

[1] All number in this rectangle refers to mIoU.

As shown in Table VIII, V1 achieves the lowest mIoU on all three datasets, indicating that the unreasonable weighting of the loss functions hinders model optimization. It is noteworthy that the mIoU obtained with fixed weights V2 and V4 are both higher than the mIoU obtained with the Random Weighting strategy. This is because although the use of dynamically weighted loss gives the model more diversity, it ignore the optimal weight of the loss. Meanwhile, compared to V2, V4 after softmax function performes best on all three datasets. This suggests that normalized loss weights can effectively utilize all loss functions, avoiding instability in training due to improper weight settings. We apply the optimal weights of V4 to the model.

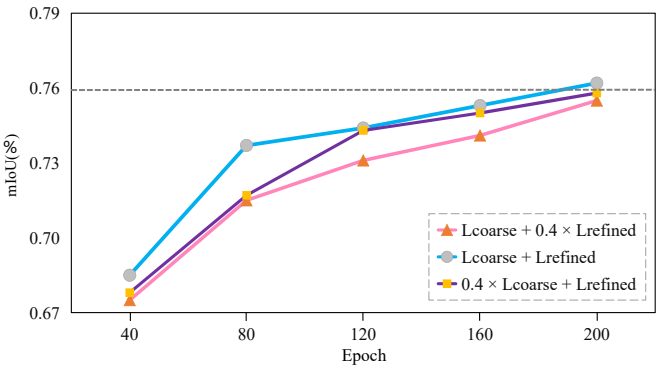


Fig. 8. Results on the Vaihingen dataset using different ratios to weight the losses of coarse segmentation and refined segmentation.

As mentioned before, Hi-ResNet finally outputs two results: coarse segmentation and refined segmentation. These two outputs will calculate three losses respectively according to the conclusion in Table VIII. However, what weight to add up the losses of the two outputs is still a question we need to consider.

Therefore, we test different weights for the losses of the two outputs. Specifically, we add the loss of coarse segmentation and the loss of refined segmentation using weights of 1:0.4, 1:1, and 0.4:1 respectively (note that the ratio of 1:0.4 is the same as the original OCR setup [43]). As shown in Figure 8, it can be seen that when the losses of the two outputs are added in a 1:1 ratio, the model can achieve the highest mIoU of 76.2%, which is superior to the settings of the original OCR network. This suggests that the target location information contained in the coarse segmentation and the object edge information contained in the refined segmentation can complement each other well, guiding the optimization direction of the model together, thereby improving the model's accuracy.

4) Pre-training Comparison: To illustrate the impact of pre-training strategies on downstream HRS tasks, we separately finetune the two pretrained models on three datasets. For supervised pre-training, the original Mapillary dataset [78] is randomly cropped into 256×256 pixels, and 400,000 category-balanced HRS images are selected as the pre-training dataset. The batch size is set to 80, and the base learning rate is set to 5×10^{-5}. The maximum iteration number of this pre-training is 3,125,000, achieving 51.8 on the Mapillary validation set. For unsupervised pre-training, the MillionAID dataset [88] with randomly cropped 400×400 pixels is used. The learning rate is set to 0.015, the batch size is 64, and the maximum iteration number is the same as supervised pre-training. The final top-1 accuracy on MillionAID is 78.9, and the top-5 accuracy is 94.1. We conduct a comprehensive evaluation of HRS pre-training models on three datasets, and the detailed information presentes in Table IX.

TABLE IX
RESULTS OF THE PRETRAINING COMPARISON EXPERIMENTS

Method	Max iteration	LoveDA	Potsdam	Vaihingen
[1]NP	-	[2]50.4	82.8	76.2
SP	3,125,000	51.8	85.2	78.6
MoCo	3,125,000	**52.6**	**87.6**	**79.8**

[1] NP: No pre-training, SP: supervised pre-training, MoCo: unsupervised pre-training using MoCoV2.
[2] All number in this rectangle refers to mIoU.

Two pre-trained model weights are loaded onto the Hi-ResNet, and the result shows that unsupervised pre-training of HRS images using MoCoV2 [82] can provides a 3.6% increase in mIoU, while supervised pre-training only increases mIoU by 2.4% under the same number of iterations. Therefore, we use the unsupervised pre-training weights in the subsequent experiments. The experiment demonstrates that unsupervised remote sensing pre-training can significantly improve the performance of the model on a small data set and make the model converge faster. Furthermore, using MoCoV2 provides a deeper feature representation for HRS downstream tasks. In this sense, pre-trained models using contrastive learning methods can offer competitive backbones for future research in the field of HRS.

Moreover, due to limited computing resources, we perform only 3-4 complete pre-training processes (150 epochs), which makes pre-training the depth provided by Hi-ResNet compares with the pre-training weights provided by other officials, there is a certain gap between the hierarchical representation information. However, the result on the LoveDA dataset in Table XII fully demonstrates the excellent performance of Hi-ResNet in terms of performance and accuracy.

5) Model Efficiency Comparison: In addition to accuracy and precision, the complexity and speed of a model are equally important for HRS tasks. Therefore, we use a single NVIDIA GTX 3090 GPU to compare our proposed model with classic CNN-based networks and Transformer-based networks in terms of the model parameters, GPU memory usage, and FLOPs. We chose to train on the large-scale, i.e., 1024×1024 LoveDA dataset, and the comparison results can be seen in Table X.

TABLE X
EFFICIENCY OF THE DIFFERENT NETWORKS ON THE LOVEDA DATASET

Method	backbone	Params(M)	Memory(MB)	FLOPs(G)	mIoU
Segmenter [68]	ViT-T	**6.7**	3495	**26**	47.1
UperNet [73]	ViT-B + RVSA	114.0	25343	407	51.9
SegFormer [17]	MiT-B1	13.7	3933	63	51.1 *-
DeepLabV3+ [50]	ResNet50	59.3	**1063**	355	47.6
HRNet [29]	HRNet-W48	75.9	1969	559	49.8
Hi-ResNet	Hi-ResNet	26.0	2116	402	**52.6**

Compared with the classic CNN based DeeplabV3+ [50], our model has relatively higher memory usage and FLOPs. This is because Hi-ResNet maintains the high resolution of the input, while simultaneously extracting image features at multiple resolutions in parallel. However, our model has almost half the number of parameters compared to DeeplabV3+, while achieving a nearly 5% higher mIoU. It is worth mentioning that although HRNet [29] uses the same parallel structure as our proposed model, this structural improvement in our ablation studies allows our model to achieve higher mIoU with fewer parameters as well as FLOPs. Due to the global attention mechanism, Transformer-based semantic segmentation networks like SegFormer [17] often require expensive computational resources, while having a relatively smaller number of parameters. In contrast, our model maintains a balance between GPU memory and FLOPs, allowing it to achieve superior accuracy within a reasonable complexity.

TABLE XI
RESULTS OF DIFFERENT INPUT SIZE ON THE VAIHINGEN DATASET

Input Size	[1]Imp.surf	Building	Lowveg	Tree	Car	mIoU
256x256	[2]85.3	90.1	72.1	80.2	70.1	79.5
256x512	84.1	90.1	71.1	80.2	69.9	79.1
512x512	85.5	90.6	72.5	80.3	70.2	**79.8**
512x1024	84.8	90.2	71.8	80.1	69.8	79.3
1024x1024	85.5	90.2	72.1	80.1	69.9	79.6

[1] Imp. surf: impervious surfaces, Lowveg: low vegetation.
[2] All number in this rectangle refers to mIoU.

6) Stability Analysis: To validate the stability of the proposed model, we conduct experiments on the Vaihingen dataset using various input sizes, including square sizes of 256×256,

TABLE XII
PERFORMANCE OF THE REFERENCE METHODS AND THE PROPOSED HI-RESNET METHOD ON THE LOVEDA DATASET

Method	Backbone	Pretrain	IoU per class(%)							mIoU	FLOPs(G)	FPS
			Background	Building	Road	Water	Barren	Forest	Agriculture			
PSPNet [12]	ResNet50	Y	44.4	52.1	53.5	76.5	9.7	44.1	57.9	48.3	738	27
DeepLabV3+ [50]	ResNet50	Y	43.0	50.9	52.0	74.4	10.4	44.2	58.5	47.6	355	46
UNet++ [15]	ResNet50	Y	42.8	52.6	52.8	74.5	11.4	44.4	58.8	48.1	544	30
SemanticFPN [51]	ResNet50	Y	42.9	51.5	53.4	74.7	11.2	44.6	58.7	48.2	589	37
FarSeg [23]	ResNet50	Y	43.1	51.5	53.9	76.6	9.8	43.3	58.9	48.2	350	47
BANet [92]	ResT-Lite	Y	43.7	51.5	51.1	76.9	16.6	44.9	62.5	49.6	67	**84**
TransUNet [93]	ViT-R50	Y	43.0	56.1	53.7	78.0	9.3	44.9	56.9	48.9	803	13
Segmenter [68]	ViT-Tiny	Y	38.0	50.7	48.7	77.4	13.3	43.5	58.2	47.1	**26**	61
SwinUperNet [39]	Swin-Tiny	Y	43.3	54.3	54.3	78.7	14.9	45.3	59.6	50.0	349	19
FactSeg [46]	ResNet50	Y	42.6	53.6	52.8	76.9	16.2	42.9	57.5	48.9	267	46
DC-Swin [70]	Swin-Tiny	Y	41.3	54.5	56.2	78.1	14.5	47.2	62.4	50.6	107	60
UperNet [73]	VITAE-B + RVSA	Y	46.7	58.1	57.1	79.6	16.5	46.4	62.4	52.4	413	11
UperNet [73]	VIT-B + RVSA	Y	45.2	59.8	55.2	79.4	18.4	46.2	59.2	51.9	407	19
RSSFormer [69]	RSS-B	Y	52.3	60.7	55.2	76.2	18.7	45.3	58.3	52.3	413	6
AerialFormer-S [60]	AerialFormer	Y	46.6	57.4	57.3	80.5	15.6	46.8	62.8	52.4	-	-
HRNet(Baseline) [29]	HRNet-W48	Y	44.6	55.3	57.4	78.0	11.0	45.3	60.9	49.8	559	25
Hi-ResNet	Hi-ResNet	Y	46.8	58.3	55.9	80.1	17.0	46.7	62.7	**52.6**	402	35

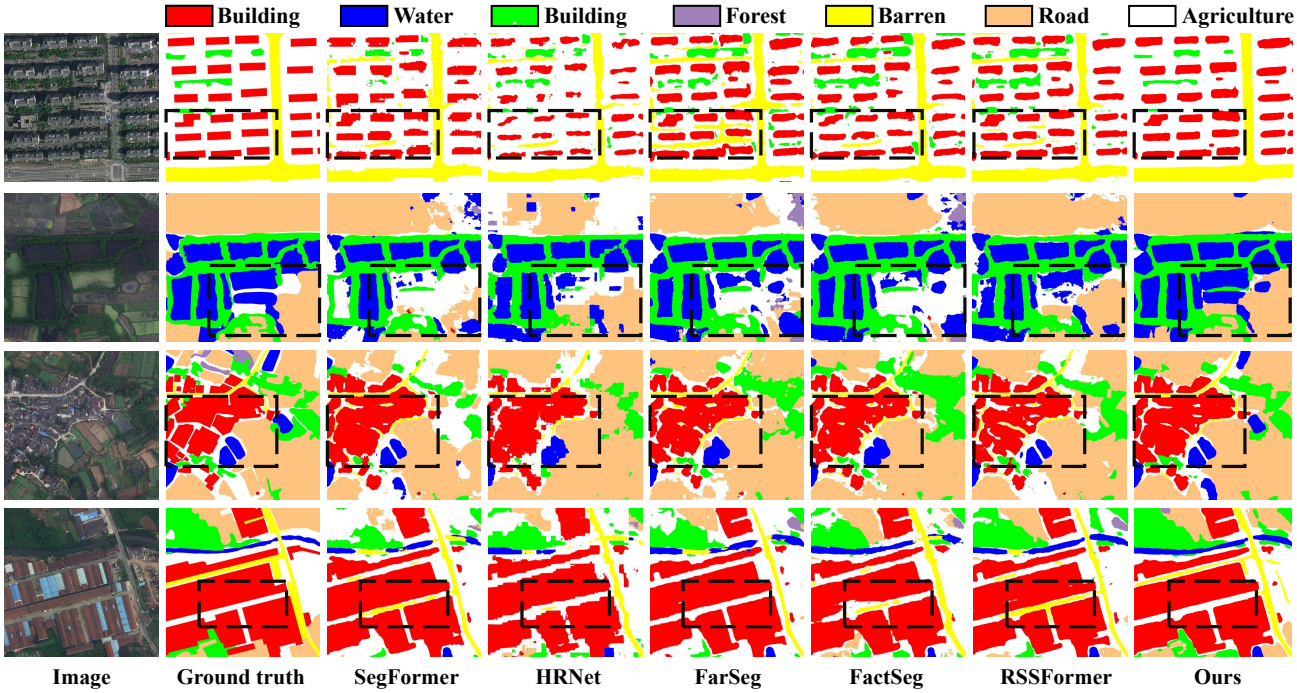


Fig. 9. Visual results of different methods on the LoveDA dataset. From left to right: original image, ground truth, results of SegFormer [17], results of HRNet [29], results of FarSeg [23], results of FactSeg [46], results of RSSFormer [69], and results of our Hi-ResNet.

512×512, and 1024×1024, as well as rectangular sizes of 256×512 and 512×1024.

As shown in Table XI, the Hi-ResNet presents a mIoU deviation of less than 0.8% for inputs of different sizes, with the best performance observed for input size of 512×512. When training with large-scale HRS images of 1024×1024, our network maintains its accuracy on the "Cars" class, which substantiates its effectiveness in segmenting smaller objects within HRS images. Correspondingly, with an input of smaller 256×256 images, the mIoU attained by the model is insignificantly different from the optimal results, suggesting that the proposed model has a larger receptive field. Furthermore, even with rectangular inputs such as 512×1024, the model still attaines the mIoU exceeding 79%, demonstrateing the stability of Hi-ResNet.

D. Results on The Dataset

1) LoveDA: The LoveDA dataset is recognized as a challenging HRS dataset for land cover domain adaptive semantic segmentation. This dataset presents three significant challenges for large-scale remote sensing mapping, namely multi-scale targets, complex background samples, and inconsistent class distributions. As a result, achieving high scores on this dataset is quite difficult.

Table XII demonstrates the results of different methods on the LoveDA dataset, where both FPS and FLOPs are evaluated on a single NVIDIA GTX 3090 GPU with an input size of 1024×1024. In addition, the official backbone pretraining weights are used for all the networks in Table XII.

TABLE XIII
PERFORMANCE OF THE REFERENCE METHODS AND THE PROPOSED HI-RESNET METHOD ON THE POTSDAM DATASET

Method	Backbone	Pretrain	F1 per class(%)					MeanF1	OA	mIoU	FLOPs(G)	FPS
			[1]Imp.surf	Building	Lowveg	Tree	Car					
ERFNet [94]	ERF	Y	88.7	93.0	81.1	75.8	90.5	85.8	84.5	76.2	**11**	142
‘ BiSeNet [95]	ResNet18	Y	90.2	94.6	85.5	86.2	92.7	89.8	88.2	81.7	20	130
DANet [13]	ResNet18	Y	89.9	93.2	83.6	82.3	92.6	88.3	86.7	79.6	58	**157**
ShelfNet [16]	ResNet18	Y	92.5	95.8	86.6	87.1	94.6	91.3	89.9	84.4	98	123
FANet [96]	ResNet18	Y	92.0	96.1	86.0	87.8	94.5	91.3	89.8	84.2	79	118
Segmenter [68]	ViT-Tiny	Y	91.5	95.3	85.4	85.0	88.5	89.2	88.7	80.7	12	138
SwinUperNet [39]	Swin-Tiny	Y	93.2	96.4	87.6	88.6	95.4	92.2	90.9	85.8	-	-
UperNet [73]	ResNet50	Y	92.4	96.1	85.7	85.5	89.9	89.9	90.6	-	-	-
UperNet [73]	Swin-Tiny	Y	92.6	96.3	86.0	85.4	89.7	90.1	90.8	-	215	58
DC-Swin [70]	Swin-Tiny	Y	94.2	97.6	88.6	89.6	96.3	**93.3**	92.0	87.5	23	72
BSNet [97]	ResNet50	Y	92.4	95.6	86.8	88.1	94.6	91.5	90.7	77.5	-	-
ST-UNet [98]	ResNet50	Y	-	-	-	-	-	86.1	-	75.9	-	9
RSSFormer [69]	RSS-B	Y	93.8	96.0	86.8	86.7	96.8	92.0	91.0	-	84	11
EfficientUNets [99]	EfficientB7	N	91.5	96.3	79.4	90.9	88.1	89.2	90.8	80.5	-	-
UperNet [72]	VIT-G	Y	92.7	96.9	85.8	89.0	96.0	92.1	92.5	-	-	-
HRNet(Baseline) [29]	HRNet-W48	Y	88.7	93.4	83.0	81.5	91.1	87.5	86.1	78.1	279	121
Hi-ResNet	Hi-ResNet	Y	93.2	96.5	87.9	88.6	96.1	92.4	**92.6**	**87.6**	57	131

[1] Imp. surf: impervious surfaces. Lowveg: low vegetation.

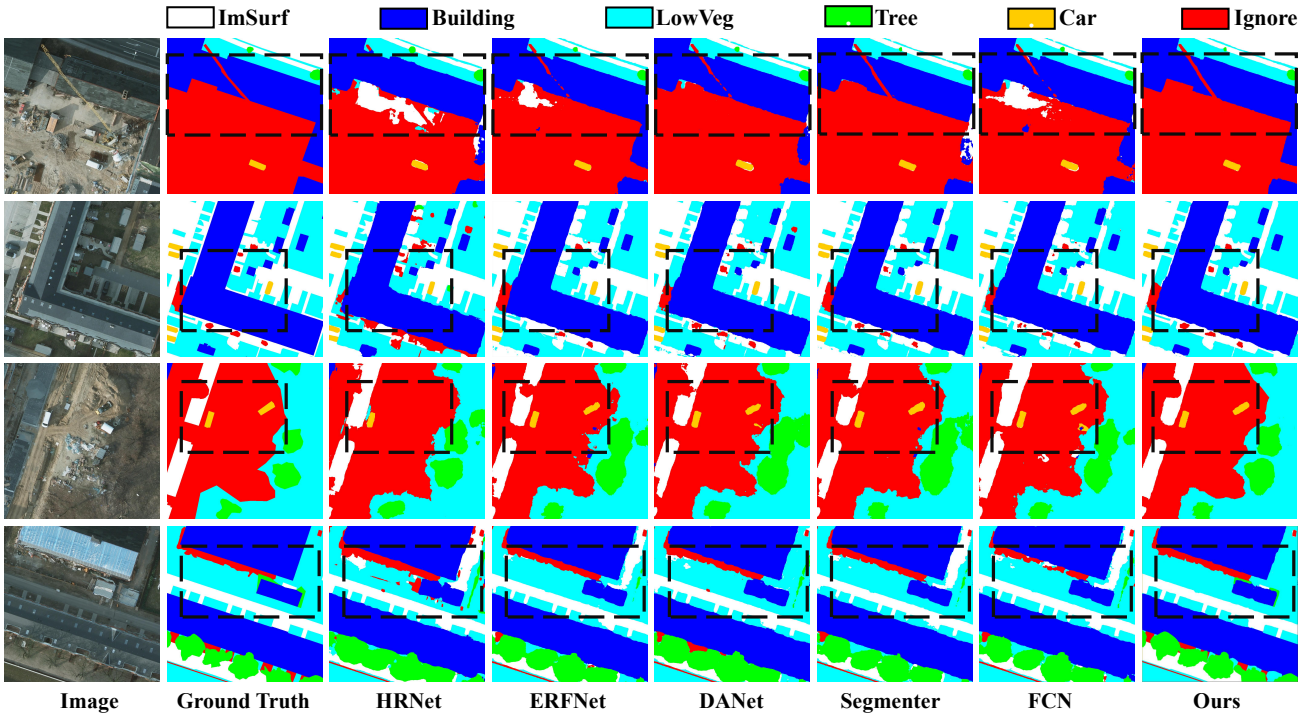


Fig. 10. Visualization results for the Potsdam validation set. From left to right: original image, ground truth, results of HRNet [29], results of ERFNet [94], results of DANet [13], results of Segmenter [68], results of FCN [85], and results of our Hi-ResNet.

Thanks to the precise loss strategy, we can handle complex samples of different backgrounds well on LoveDA and achieve the mIoU of 52.5% on the official test set. Our network outperforms the HRNet [29], loaded with officially provided pre-trained weights, by 2.5% on mIoU and FactSegNet [46], an excellent small-object semantic segmentation network, by 3.5%. It is worth noting that for the “Barren” class, where most networks underperform, our mIoU is 3% higher than most methods. Whether in urban or rural scenarios, sparse or dense distribution, our network can accurately segment objects with high confidence.

However, We admit that compare to models like Segmenter [68], Hi-ResNet underperforms in terms of model complexity and FPS. We attribute it to two reasons. First, the use of high-resolution 1024×1024 input images significantly increases the computation of our model’s attention mechanism, which in turn reduces the inference speed. The second reason is that maintaining high-resolution image feature computation throughout the network consumes more computational resources. Nevertheless, our proposed model still holds advantages over CNN-based models, managing to achieve superior mIoU on difficult datasets with a relatively small increase in complexity. We provide visual comparison results with other methods in Figure 9.

2) Potsdam: As a widely-used dataset for segmentation tasks, Potsdam can comprehensively demonstrates the im-

TABLE XIV
PERFORMANCE OF THE REFERENCE METHODS AND THE PROPOSED HI-RESNET METHOD ON THE VAIHINGEN DATASET

Method	Backbone	Pretrain	F1 per class(%)					MeanF1	OA	mIoU	FLOPs(G)	FPS
			Imp.surf	Building	Lowveg	Tree	Car					
PSPNet [12]	ResNet18	Y	89.0	93.2	81.5	87.7	43.9	79.0	87.7	68.6	53	112
BiSeNet [95]	ResNet18	Y	89.1	91.3	80.9	86.9	73.1	84.3	87.1	75.8	20	128
DABNet [18]	DAB	N	87.8	88.8	74.3	84.9	60.2	79.2	84.3	70.2	**7**	146
DANet [13]	ResNet18	Y	90.0	93.9	82.2	87.3	44.5	79.6	88.2	69.4	58	153
ShelfNet [16]	ResT-Lite	Y	91.8	94.6	83.8	89.3	77.9	87.5	89.8	78.3	98	122
FANet [96]	ResNet18	Y	90.7	93.8	82.6	88.6	71.6	85.4	88.9	75.6	79	118
EaNet [14]	ResNet18	Y	91.7	94.5	83.1	89.2	80.0	87.7	89.7	78.7	-	-
ABCNet [59]	ResNet18	Y	92.7	95.2	84.5	89.7	85.3	89.5	90.7	81.3	16	**185**
BoTNet [100]	BoTNet50	Y	89.9	92.1	81.8	88.7	71.3	84.8	88.0	74.3	102	-
Segmenter [68]	ViT-Tiny	Y	89.8	93.0	81.2	88.9	67.6	84.1	88.1	73.6	12	130
BSNet [97]	ResNet50	Y	92.1	94.4	83.1	88.3	86.7	88.9	90.3	80.2	-	-
DC-Swin [70]	Swin-S	Y	93.6	96.1	85.7	90.3	87.6	**90.6**	91.6	**83.2**	23	80
ST-UNET [98]	ResNet50	Y	-	-	-	-	-	82.1	-	70.2	-	7
RSSFormer [69]	RSS-B	Y	93.7	96.8	81.3	91.7	89.2	90.5	90.8	-	84	10
EfficientUNets [99]	EfficientB7	N	91.4	96.3	79.4	90.8	88.1	89.0	90.8	73.1	-	-
HRNet(Baseline) [29]	HRNet-W48	Y	89.8	92.8	81.0	86.8	79.5	86.0	87.6	75.8	279	120
Hi-ResNet	Hi-ResNet	Y	92.3	95.1	84.9	89.5	83.5	90.1	**91.7**	79.8	57	123

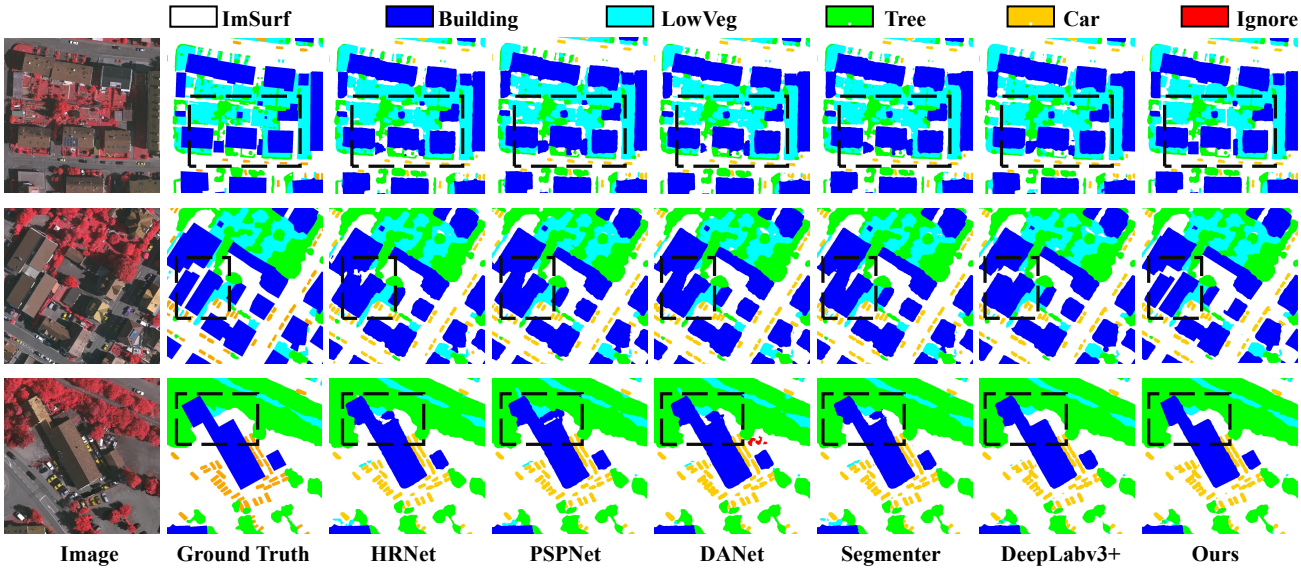


Fig. 11. Visualization results for the Vaihingen validation set. From left to right: original image, ground truth, results of HRNet [29], results of PSPNet [12], results of DANet [18], results of Segmenter [68], results of DeepLabv3+ [50], and results of our Hi-ResNet.

provement of the accuracy of HRS images by the model proposed in this paper. Table XIII shows the scores achieved on the Potsdam dataset. The FLOPs and FPS are measured by 512×512 inputs on a single NVIDIA GTX 3090 GPU. The network proposed in this paper achieves the F1 of 92.4% and mIoU of 87.6% on the Potsdam dataset. Hi-ResNet outperforms the lightweight convolutional network FANet [96] and the lightweight transformer-based network Segmenter [68]. Notably, Hi-ResNet performs well among all methods for the "Lowveg" class, achieving a score of 87.9%. The car category also achieved a high score with a mean F1 of 96.1%. This result fully demonstrates that the Hi-ResNet has better performance for small target segmentation in HRS images.

Moreover, when the input image resolution is reduced from 1024×1024 to 512×512, the complexity of Hi-ResNet significantly decreases, and the model's inference speed accelerates nearly sixfold. At this point, our proposed model can achieve faster inference speeds than Transformer-based networks such as RSSFormer [69] and DC-Swin [70], even under conditions of higher complexity. This validates the performance advantage of our network.

We present Potsdam segmentation results to showcase the effectiveness of Hi-ResNet for small object segmentation. As displayed in Figure 10, most networks perform poorly in the segmentation of object edges. To overcome this limitation, Hi-ResNet employs CEA loss in the loss calculation to maximize the distance between the two boundaries, ensuring good connectivity of the extracted edge features during loss calculation, thus avoiding category boundary blur.

3) Vaihingen: The Vaihingen dataset has a large number of houses obscured by tree branches and multi-story small villages, so the dataset requires the network to identify and segment small targets more accurately. Table XIV shows the results of different methods on the Vaihingen dataset. (Note

that the calculation for FLOPs and FPS is the same as for Potsdam). Our proposed network achieves an OA of 91.7% and the mIoU of 79.8% on the Vaihingen dataset. In the low vegetation category, Hi-ResNet secured first place with the same performance on the Potsdam dataset. Significantly, our network improves the results for the categories of "Building" and "Car" by 4% compared to the HRNet network. This is because Hi-ResNet effectively solves the sample imbalance problem caused by small targets occupying small pixels in HRS images by using the CEA loss and GD loss to weigh each category.

We show some typical segmentation results in vaihingen in Figure 11. Most networks present misclassification on the "Tree" class and the "Lowveg" class. At the same time, the within-class distance segmentation redundancy of the dense small target object "Car" class, and the edge segmentation of the small cars is not clear. Hi-ResNet can obtain the position and edge information of small cars more accurately when the global receptive field is increased, thereby avoiding misclassification in complex scenes.

V. Conclusion

Our study centers on the semantic segmentation of HRS, specifically focusing on addressing the inherent challenges of object scale and shape variance, and complex background environments. These issues often lead to object misclassification and sub-optimal outcomes with current learning algorithms. We respond by developing Hi-ResNet, which stands out due to an efficient network structure that includes a funnel module, a multi-branch module embedded with IA blocks, and a feature refinement module. Additionally, we introduce the CEA loss function. In our approach, the funnel module functions to downsample and extract high-resolution semantic information from the input image. The process then moves to the multi-branch module with stacks of IA blocks, enabling the capture of image features at different scales and distinguishing variant scales and shapes within the same class. Our study concludes with the integration of the CEA loss function within our feature refinement module. This innovative step effectively disambiguates inter-class objects with similar shapes and increases the data distribution distance for accurate predictions. The superiority of Hi-ResNet is proven through a comparative evaluation with leading methodologies across LoveDA benchmarks. The results underscore the value of our contributions to advancing HRS semantic segmentation and demonstrate the sensitivity of parallel architecture of the input size.

References

[1] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017. I

[2] X. Gao, M. Wang, Y. Yang, and G. Li, "Building extraction from rgb vhr images using shifted shadow algorithm," *Ieee Access*, vol. 6, pp. 22 034–22 045, 2018. I

[3] P. Qin, Y. Cai, J. Liu, P. Fan, and M. Sun, "Multilayer feature extraction network for military ship detection from high-resolution optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 058–11 069, 2021. I

[4] A. J. Cooner, Y. Shao, and J. B. Campbell, "Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 haiti earthquake," *Remote Sensing*, vol. 8, no. 10, p. 868, 2016. I

[5] C. Xiong, Q. Li, and X. Lu, "Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network," *Automation in Construction*, vol. 109, p. 102994, 2020. I

[6] A. Wanto, S. D. Rizki, S. Andini, S. Surmayanti, N. Ginantra, and H. Aspan, "Combination of sobel+ prewitt edge detection method with roberts+ canny on passion flower image identification," in *Journal of Physics: Conference Series*, vol. 1933, no. 1. IOP Publishing, 2021, p. 012037. I

[7] R. Tian, G. Sun, X. Liu, and B. Zheng, "Sobel edge detection based on weighted nuclear norm minimization image denoising," *Electronics*, vol. 10, no. 6, p. 655, 2021. I

[8] P. A. Rogerson, "Change detection thresholds for remotely sensed images." *Journal of Geographical Systems*, vol. 4, no. 1, 2002. I

[9] J. Yang, Y. He, and J. Caspersen, "Region merging using local spectral angle thresholds: A more accurate method for hybrid segmentation of remote sensing images," *Remote sensing of environment*, vol. 190, pp. 137–148, 2017. I

[10] Z. Wang, J. R. Jensen, and J. Im, "An automatic region-based image segmentation algorithm for remote sensing applications," *Environmental Modelling & Software*, vol. 25, no. 10, pp. 1149–1165, 2010. I

[11] X. Zhang, X. Feng, P. Xiao, G. He, and L. Zhu, "Segmentation quality evaluation using region-based precision and recall measures for remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 102, pp. 73–84, 2015. I

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890. I, III-A2, XII, XIV, 11

[13] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154. I, II-B, XIII, 10, XIV

[14] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 170, pp. 15–28, 2020. I, XIV

[15] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11. I, II-A, XII

[16] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "Shelfnet for fast semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0. I, XIII, XIV

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021. I, II-B, X, IV-C5, 9

[18] G. Li, I. Yun, J. Kim, and J. Kim, "Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," *arXiv preprint arXiv:1907.11357*, 2019. I, XIV, 11

[19] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 60–77, 2018. I

[20] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–9. I

[21] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1102–1110. I

[22] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 96–107, 2018. I

[23] 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 *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, 2020, pp. 4096–4105. I, XII, 9

[24] J. Bai, J. Ren, Y. Yang, Z. Xiao, W. Yu, V. Havyarimana, and L. Jiao, "Object detection in large-scale remote-sensing images based on time-frequency analysis and feature optimization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021. I

[25] S. Yin, H. Li, L. Teng, M. Jiang, and S. Karim, "An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images," *International Journal of Image and Data Fusion*, vol. 11, no. 2, pp. 201–214, 2020. I

[26] L. Li, Z. Zhou, B. Wang, L. Miao, and H. Zong, "A novel cnn-based method for accurate ship detection in hr optical remote sensing images via rotated bounding box," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 686–699, 2020. I

[27] X. Wang, M. Kang, Y. Chen, W. Jiang, M. Wang, T. Weise, M. Tan, L. Xu, X. Li, L. Zou *et al.*, "Adaptive local cross-channel vector pooling attention module for semantic segmentation of remote sensing imagery," *Remote Sensing*, vol. 15, no. 8, p. 1980, 2023. I

[28] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9201–9222, 2019. I

[29] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020. I, II-A, IV-C1, X, IV-C5, XII, 9, XIII, 10, IV-D1, XIV, 11

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. I

[31] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote sensing scene classification via multi-branch local attention network," *IEEE Transactions on Image Processing*, vol. 31, pp. 99–109, 2021. I

[32] W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, and L. Wang, "Gcsanet: A global context spatial attention deep learning network for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1150–1162, 2022. I

[33] Q. Bi, K. Qin, H. Zhang, and G.-S. Xia, "Local semantic enhanced convnet for aerial scene recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 6498–6511, 2021. I

[34] X. Zhao, J. Zhang, J. Tian, L. Zhuo, and J. Zhang, "Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image," *Remote Sensing*, vol. 12, no. 11, p. 1887, 2020. I

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. I

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357. I

[37] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578. I

[38] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019. I

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. I, III-A3, XII, XIII

[40] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31. I

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. I, II-B, III-A3

[42] X. Li, L. Xie, C. Wang, J. Miao, H. Shen, and L. Zhang, "Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images," *GIScience & Remote Sensing*, vol. 61, no. 1, p. 2356355, 2024. I

[43] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019. I, II-A, 2, III-A4, IV-C3

[44] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 182–186. I

[45] Y. Lin, D. Xu, N. Wang, Z. Shi, and Q. Chen, "Road extraction from very-high-resolution remote sensing images via a nested se-deeplab model," *Remote sensing*, vol. 12, no. 18, p. 2985, 2020. I, II-B

[46] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021. I, XII, 9, IV-D1

[47] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019. I, III-B3, III-B3

[48] Z. Cheng and D. Fu, "Remote sensing image segmentation method based on hrnet," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 6750–6753. II-A

[49] 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 *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 1442–1450. II-A

[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818. II-A, III-A2, X, IV-C5, XII, 11

[51] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6399–6408. II-A, XII

[52] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020. II-A

[53] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5751–5765, 2021. II-A

[54] J. Shi, W. Liu, H. Shan, E. Li, X. Li, and L. Zhang, "Remote sensing scene classification based on multibranch fusion attention network," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023. II-A

[55] S. Zheng, C. Lu, Y. Wu, and G. Gupta, "Sapnet: Segmentation-aware progressive network for perceptual contrastive deraining," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 52–62. II-A

[56] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019. 2, III-B, III-B2

[57] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248. 2, III-B

[58] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 096–10 105. II-B

[59] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 84–98, 2021. II-B, XIV

[60] K. Yamazaki, T. Hanyu, M. Tran, A. Garcia, A. Tran, R. McCann, H. Liao, C. Rainwater, M. Adkins, A. Molthan *et al.*, "Aerialformer: Multi-resolution transformer for aerial image segmentation," *arXiv preprint arXiv:2306.06842*, 2023. II-B, XII

[61] T. Tian, L. Li, W. Chen, and H. Zhou, "Semsdnet: A multiscale dense network with attention for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5501–5514, 2021. II-B

[62] X. Zhang, J. Li, and Z. Hua, "Mrse-net: multiscale residuals and se-attention network for water body segmentation from satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5049–5064, 2022. II-B

[63] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022. II-B

[64] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2020. II-B

[65] W. Wang, X. Tan, P. Zhang, and X. Wang, "A cbam based multiscale transformer fusion approach for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817–6825, 2022. II-B

[66] L. Zhu, X. Geng, Z. Li, and C. Liu, "Improving yolov5 with attention mechanism for detecting boulders from planetary images," *Remote Sensing*, vol. 13, no. 18, p. 3776, 2021. II-B

[67] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021. II-B

[68] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272. II-B, X, XII, XIII, 10, IV-D1, XIV, 11, IV-D2

[69] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1052–1064, 2023. II-B, XII, 9, XIII, XIV, IV-D2

[70] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022. II-B, XII, XIII, XIV, IV-D2

[71] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022. II-B, IV-A2, IV-A3

[72] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," *arXiv preprint arXiv:2304.05215*, 2023. II-C, XIII

[73] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer towards remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2022. II-C, X, XII, XIII

[74] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, 2022. II-C

[75] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 181–10 190. II-C

[76] S. Workman, A. Hadzic, and M. U. Rafique, "Handling image and label resolution mismatch in remote sensing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3709–3718. II-C

[77] G. Kong and H. Fan, "Enhanced facade parsing for street-level images using convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 519–10 531, 2020. II-C

[78] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999. II-C, III-C1, IV-C4

[79] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022. II-C

[80] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. II-C

[81] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020. II-C

[82] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738. II-C, II-C, III-C2, IV-C4

[83] N. Zhao, Z. Wu, R. W. Lau, and S. Lin, "What makes instance discrimination good for transfer learning?" *arXiv preprint arXiv:2006.06606*, 2020. II-C

[84] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. III-A1

[85] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. III-A2, 10

[86] O. Ronneberger, P. Fischer, and T. Brox, "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 18*. Springer, 2015, pp. 234–241. III-A2

[87] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. III-A2

[88] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 14, pp. 4205–4230, 2021. III-C1, IV-C4

[89] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021. IV-A1

[90] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. IV-B1

[91] B. Lin, F. Ye, Y. Zhang, and I. W. Tsang, "Reasonable effectiveness of random weighting: A litmus test for multi-task learning," *arXiv preprint arXiv:2111.10603*, 2021. IV-C3

[92] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sensing*, vol. 13, no. 16, p. 3065, 2021. XII

[93] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. XII

[94] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017. XIII, 10

[95] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341. XIII, XIV

[96] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff, "Real-time semantic segmentation with fast attention," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 263–270, 2020. XIII, XIV, IV-D2

[97] J. Hou, Z. Guo, Y. Wu, W. Diao, and T. Xu, "Bsnet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–22, 2022. XIII, XIV

[98] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022. XIII, XIV

[99] H. AlMarzouqi and L. S. Saoud, "Semantic labeling of high resolution images using efficientunets and transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2023. XIII, XIV

[100] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 519–16 529. XIV