SSDT: Scale-Separation Semantic Decoupled Transformer for Semantic Segmentation of Remote Sensing Images

Chengyu Zheng, Yanru Jiang, Xiaowei Lv, Jie Nie, *Member, IEEE*, Xinyue Liang, *Member, IEEE*, and Zhiqiang Wei, *Member, IEEE*

***Abstract*—As we all know, semantic segmentation of remote sensing (RS) images is to classify the images pixel by pixel to realize the semantic decoupling of the images. Most traditional semantic decoupling methods only decouple and do not perform scale-separation operations, which leads to serious problems. In the semantic decoupling process, if the feature extractor is too large, it will ignore the small-scale targets; if the feature extractor is too small, it will lead to the separation of large-scale target objects and reduce the segmentation accuracy. To address this concern, we propose a scale-separated semantic decoupled transformer (SSDT), which first performs scale-separation in the semantic decoupling process and uses the obtained scale information-rich semantic features to guide the Transformer to extract features. The network consists of five modules, scale-separated patch extraction (SPE), semantic decoupled transformer (SDT), scale-separated feature extraction (SFE), semantic decoupling (SD), and multiview feature fusion decoder (MFFD). In particular, SPE turns the original image into a linear embedding sequence of three scales; SD divides pixels into different semantic clusters by K-means, and further obtains scale information-rich semantic features; SDT improves the intraclass compactness and interclass looseness by calculating the similarity between semantic features and image features, the core of which is decoupled attention. Finally, MFFD is proposed to fuse salient features from different perspectives to further enhance the feature representation. Our experiments on two large-scale fine-resolution RS image datasets (Vaihingen and Potsdam) demonstrate the effectiveness of the proposed SSDT strategy in RS image semantic segmentation tasks.**

***Index Terms*—Geophysical image processing, geoscience and remote sensing, semantic segmentation.**

Manuscript received 1 November 2023; revised 13 February 2024; accepted 16 March 2024. Date of publication 9 April 2024; date of current version 1 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62072418 and Grant 62172376, in part by the Fundamental Research Funds for the Central Universities under Grant 202042008, in part by the Major Scientific and Technological Innovation Project under Grant 2019JZZY020705, in part by the Key Research and Development Program of Qingdao Science and Technology Plan (21-1-2-18-xx), and in part by the Central Government Guide Local Science and Technology Development Special Fund Project under GrantYDZX2022028. *(Chengyu Zheng and Yanru Jiang are co-first authors.) (Corresponding authors: Jie Nie; Xinyue Liang.)*

The authors are with the College of Information Science and Engineering, Ocean University of China, Qingdao 266005, China (e-mail: zhengchengyu@stu.ouc.edu.cn; jiangyanru@stu.ouc.edu.cn; lvxiaowei@stu.ouc.edu.cn; niejie@ouc.edu.cn; liangxinyue@ouc.edu.cn; weizhiqiang@ouc.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3383066

I. Introduction

With the rapid development of new technologies such as satellite sensors and aerospace, remote sensing (RS) technology continues to progress and image resolution is increasing. Rational analysis and use of high-resolution RS images are significant to monitoring disaster forecasting, autonomous driving, and national land resource protection [1], [2], [3], [4]. Semantic segmentation [5] is an important topic in computer vision, which aims to achieve region segmentation by determining the class of individual pixels in an image and then recognizing the semantic information of that class to superimpose high-level semantics on the segmentation result. In recent years, deep learning has led to breakthroughs in the field of semantic segmentation, but the task of semantic segmentation of RS images remains challenging due to the differences between RS optical images and ordinary images, and the existence of a large number of target objects with different scales and different semantics in RS images.

In recent years, most of the cutting-edge semantic segmentation models have been based on convolutional neural networks (CNNs), which further extend the scope of semantic segmentation and improve the accuracy of distinguishing recognized objects. The classical deep learning semantic segmentation networks include CNN-based FCN [6], UNet [7], SegNet [8], etc, in which the "encoder-decoder" paradigm [7], [8], [9], [10] is the main network structure framework. Since contextual information is the most critical factor in improving the performance of semantic segmentation, Chen et al. [11], [12], [13], [14] proposed the DeepLab series to explore multiscale contextual information to improve target recognition at different scales. In addition, using attention mechanisms [15], [16], [17], [18] to capture contextual information or feature extraction through graph convolutional networks (GCNs) [19], [20], [21], [22], [23], [24] can further model the relationships between target objects. MSCG-Net [24] is based on GCN to establish connections between pixels by building nodes and edges and can integrate context information to obtain better performance. With Transformer's excellent performance in NLP, the introduction of ViT [25] led the field of computer vision to take a significant step forward, and the authors in [26], [27], [28], [29] fully explored the segmentation capability of ViT and improved the global long-range modeling capability of the network. The most

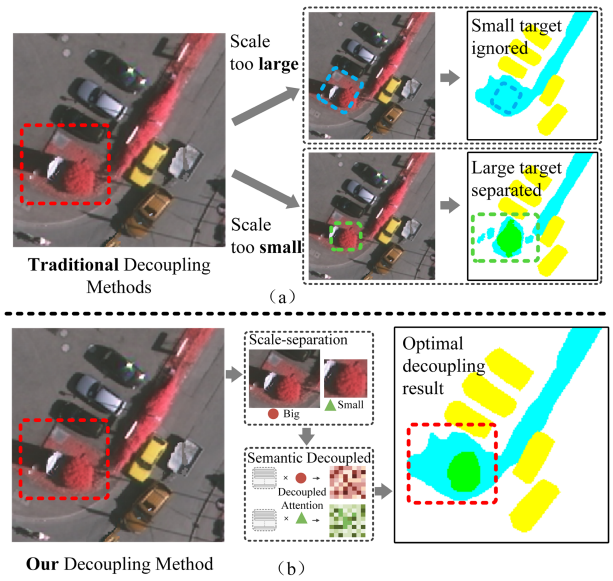

Fig. 1. Here, we illustrate the current problems. (a) Unsatisfactory decoupling result caused by the traditional decoupling method. The red box shows semantic entanglement in RS images: the "Tree" in the green box is coupled above the "Low Vegetation" in the blue box. In the process of decoupling the two semantics, if the feature extractor is too large, such as the convolution kernel of CNN is set large, it will mainly decouple the "Low Vegetation" semantics in the blue box, resulting in ignoring the small-scale target object tree. If the feature extractor is too small, it will decouple the "Tree" semantics in the green box, which will lead to the separation of the "Low Vegetation" of the big-scale target object and decrease the segmentation accuracy. (b) Shows our SD method. In the process of SD, a scale-separation operation is carried out to avoid the interference of scale information on decoupling, and the most ideal decoupling result can be obtained.

popular of these methods is the Swin Transformer [29], which proposes a hierarchical transformer that brings higher efficiency by restricting the self-attentive computation to nonoverlapping local windows, while still allowing crosswindow connections. Swin-UNet [30] adds the idea of UNet on its basis and constructs a purely transformer-based U-shaped structure.

Although the above techniques have contributed to promoting progress in the field of semantic segmentation, the above methods still have shortcomings when used in the field of RS images: they cannot segment the coupled target objects accurately on the scale entangled feature space. Traditional methods such as Swin-UNet [30], although the interrelationship between individual semantic features is extracted using the attention mechanism to achieve semantic decoupling (SD), no scale-separation operation is performed in the SD process. This will lead to a serious problem that the semantics of each scale are mixed and the cluttered scale information will affect the semantic judgment. In addition, due to the different imaging principles, RS images will contain more information than ordinary images, which makes it more difficult to recognize target objects. Therefore, it is not reliable to use only semantic information for RS image semantic segmentation, which limits the semantic extraction ability of the segmentation model. Fig. 1(a) shows the unsatisfactory decoupling result caused by the traditional decoupling method. It is obvious that the RS image has two coupled semantics in the same location space: the "Tree" in the green box is coupled above the "Low Vegetation" in the blue box. In the process of decoupling the two semantics, if the feature extractor is too large, such as the convolution kernel of CNN is set large, it will mainly decouple the semantics in the blue box, and ignore the small target object "Tree" in the decoupling process. If the feature extractor is too small, it will decouple the semantics in the green box, which will cause the big target object "Low Vegetation" to be separated, and the segmentation accuracy will be reduced. From the above two points, it can be seen that semantic segmentation in RS images with seriously entangled semantic information scales is prone to missegmentation.

To solve the above problems, we propose semantic information based on scale-separation to guide Transformer decoupling model, to improve the accuracy of the model description of coupling target objects. Fig. 1(b) shows our SD method. In the process of SD, scale-separation operations are carried out to avoid the interference of scale information on the decoupling, and the most ideal decoupling results can be obtained. First, we still inherit the traditional Swin Transformer module to ensure the interaction of global contextual information, using Swin-UNet [30] as a baseline, unlike Swin Transformer which uses a single scale for feature extract, this article proposes scale-separated patch extraction (SPE), which chunks the original image at different scales sizes to generate three scales of linear mapping embedding sequences for global contextual representation modeling. Second, scale-separated feature extraction (SFE), and SD are proposed to use scale information to divide pixels into different semantic clusters by clustering methods, which can obtain scale information-rich semantic features. It is worth noting that the semantic information we add is the semantic information after scale-separation modeling extracted from the image itself, and it is not supervised by additional word embedding (such as Segmenter [27]), which avoids the difference between the modes between the word and the image. In addition, semantic decoupled transformer (SDT) is proposed to extract intraclass feature interdependencies and reduce interclass feature associations by computing the similarity between semantic features and image features, with decoupled attention at its core. Finally, multiview feature fusion decoder (MFFD) is proposed to fuse salient features from different perspectives to further enhance the feature representation. We compare our method with previous methods on two public datasets. Experiments show that our method outperforms the state-of-the-art semantic segmentation models. The main contributions are as follows.

1) We propose a scale-separation semantic decoupled transformer (SSDT) for semantic segmentation of RS images, which implements a SD module within the Transformer. This can effectively avoid the influence of scale coupling on semantic judgments, and not only helps the Transformer to provide effective semantic features but also helps to compensate for the lack of spatial location of the Transformer.
2) We propose five modules: SPE, SDT, SFE, SD, and MFFD. The total network framework superimposed by each module can obtain the scale information-rich semantic features and use the semantic features to guide and reduce the correlation of interclass features to solve the

problem of severe coupling of semantic information scales within RS images.

3) We validated the validity of the proposed method on the Potsdam dataset and the Vaihingen dataset. Several comparative and ablation experiments were carried out to prove the effectiveness of the scale-separated SD Transformer framework in RS image segmentation.

The rest of this article is organized as follows. Section II introduces the work related to the semantic segmentation of RS images. Section III provides details of the SSDT. Section IV describes the corresponding experimental results and analysis. Finally, Section V concludes this article.

II. Related Work

A. Semantic Segmentation Based on CNNs

In recent years, common semantic segmentation models are based on CNNs, which further extend the scope of semantic segmentation and improve the accuracy of distinguishing recognized objects, among which the classical deep learning semantic segmentation networks include FCN [6], UNet [7], SegNet [8], etc. In particular, FCN [6] brings semantic segmentation into an end-to-end training to achieve the pixel-level classification of images, thus, solving semantic-level image segmentation. But context information is the most crucial factor in improving semantic segmentation performance, as different objects and scenes exhibit different contextual relationships at different scales. To capture scale information between pixels more effectively, an "encoder-decoder" paradigm is proposed [7], [8], [9], [10]. UNet [7] adds low-level spatial features to high-level semantic features through skip-connection to achieve feature fusion at different scales, and PSPNet [10] adds a spatial pyramid pooling module to obtain a set of feature maps with different sensory field sizes to fuse features at different scales. To improve the utilization of global information, the DeepLab series proposed by Chen et al. [11], [12], [13], [14] explores multiscale contextual information to improve target recognition at different scales and increase segmentation accuracy. DeepLabV1 [11] proposed the concept of dilated convolution, increasing the size of the receive field to enable networks to capture a larger range of contextual information. However, the model has two main issues: continuous pooling operations leading to a decrease in spatial resolution, and the invariance of spatial transformations required for the classifier to obtain object-centered decisions. To address these issues, DeepLabV2 [12] introduced atmosphere spatial pyramid pooling, which allows the network to perform feature extraction at multiple different sampling rates, thereby segmenting objects more accurately. DeepLabV3 [13] improved the ResNet structure based on V2, enabling it to use dilated convolutions and pyramid dilated convolutions. This improvement enables the network to maintain the size of the feature map while maintaining the receptive field, thereby better preserving spatial information. DeepLabV3+ [14] introduced dilated spatial pyramid pooling and improves the upsampling method to achieve higher resolution and more accurate segmentation results. However, there are still some challenges and limitations, such as high computational complexity and the need to improve model robustness.

Further, some researchers perform feature extraction through graph convolution [19], [20], [21], [22], [23] to model the relationship between target objects. Other researchers have concentrated on attention mechanisms [15], [16], [17], [18] to capture contextual information, which allows networks to ignore irrelevant information and focus on priority information, including spatial domain attention, channel domain attention, layer domain attention, hybrid domain attention, temporal domain attention, and self-attention mechanisms. For example, Ding et al. [16] proposed local attention network (LANet), which combines patch attention module and attention embedding module to obtain the degree of interlocation dependency by calculating the pixel-to-pixel similarity and incorporating the neighboring pixel information into the computed pixels, thus, freeing the fusion of global information from the limitation of image distance. SENet [17] can be interpreted as making the model focus more on a certain aspect of features and correspondingly will assign weights to each channel, thus distinguishing the importance of different features and achieving the effect of reinforcing a certain feature. DANet [18] attached two attention modules to the dilated FCN based on a self-attentive mechanism, modeling semantic dependencies in the spatial and channel dimensions, respectively. However, the multiscale features extracted by atrous convolution or pyramidal pooling are limited, so LoG-CAN [31], which improves the segmentation performance of RS images by utilizing local details and global semantic information of the image is proposed. Hang et al. [32] used a "multiscale progressive segmentation network" to gradually segment objects into small, large, and other scales, which solves the problem that due to the limited learning capacity of each CNN, it tends to make tradeoffs when segmenting objects of different scales. Wang et al. [33] proposed structure-driven relation graph networks, which utilize graph networks to model complex relationships between objects and capture subtle differences between them through a structure-driven approach, thereby improving the accuracy of fine-grained recognition.

In addition, models for hyperspectral data are also growing. SDEnet [34] utilized single-source hyperspectral data and unlabeled data in the target scene to design a generator that includes a semantic coder and a morphological coder to classify hyperspectral images in the target scene and achieves good performance on the cross-scene hyperspectral image classification task. The FHS-SSL [35] algorithm is a self-supervised learning method for hyperspectral image classification using unlabeled data. The algorithm introduces migration learning and meta-learning to further improve the classification accuracy. However, the above model requires a large amount of unlabeled data for training, and some labeled data need to be manually selected for supervised learning. Therefore, GACP [36] using graph neural networks, ARMA filters, and parallel CNNs has been proposed to be able to combine the spatial structure and spectral features of hyperspectral data to improve the accuracy of hyperspectral image classification. However, the GACP algorithm has high computational complexity and requires a long training time. Then, the cross-scene hyperspectral image classification method

LDGnet [37] was proposed, which is a joint modeling method using linguistic modalities and visual modalities with prior knowledge of remotely sensed features. The core idea is to align visual and linguistic features category-by-category to output the classification prediction probability of visual modalities, which can improve classification accuracy and reduce the dependence on domain knowledge.

Although discriminable feature learning methods based on attention mechanisms can extract discriminative information in images. However, there are still two problems: First, in the field of semantic segmentation of RS images, there is no relevant research that considers both intraclass compactness and interclass looseness of RS objects; Second, although the existing multiscale based semantic segmentation methods are relatively mature, the scale coupling phenomenon of semantic information is serious, which makes the results of multiscale prediction models unreliable. Therefore, the above methods cannot maximize the accuracy of semantic segmentation.

B. Semantic Segmentation Based on Transformer

With the excellent performance of Transformer in NLP, the introduction of ViT [25] led the field of computer vision to take a big step forward. Zheng et al. [26] proposed SETR to analyze image segmentation from a sequence perspective and thoroughly explored the segmentation capability of ViT. Strudel et al.[27] proposed a semantic segmentation method (Segmenter) using only Transformer for context modeling, which has the advantage of capturing global interactions between scene elements using the global image context at each layer of the model and improving global dependencies between features by using predefined class embeddings in the decoding part to capture semantic information by masking Transformer to obtain class labels. SegFormat proposed by Xie et al. [28] used a new position-free coded and hierarchical Transformer encoder with a lightweight ALL-MLP decoder design to achieve better results. Swin Transformer [29] is another collision of Transformer in the field of vision, whose main idea is to divide the feature map into multiple disjoint regions and to perform self-attentive computation only within this window to reduce the computation, especially when the shallow feature map is large. Still, at the same time, it also reduces the information transfer between spatial locations. MPViT [38] proposed a multiscale block coding and multipath structure, where blocks of different sizes are coded simultaneously by overlapping convolution operations to produce features with the same sequence length, and then the resulting features are fed into the Transformer structure in parallel to produce better results. SOT-Net [39] utilizes ultrahigh resolution RS and LiDAR data for structured analysis, which strengthens the semantic association of multisource information, and achieves an adaptive fusion of multimodal data through the crossattention mechanism, which can achieve higher classification accuracy.

In the field of computer vision such as semantic segmentation of RS images, CNNs show excellent performance, mainly due to convolution operations. The ability to collect local features of different layers for better feature representation. However, the use of global information needs to be improved. At the same time, Transformer is being applied to computer vision, enabling self-attention mechanisms and multilayer sensing machine structures to reflect complex spatial transformations and long-distance feature dependencies. Hence, some works [40], [41], [42], [43] proposed combining the two for feature extraction to fuse global and local information interactively. UNetFormer [40] used only the UNet architecture and simply splices the CNN with the Transformer-based decoder, aiming to improve the global feature extraction capability of neural networks on raw images for efficient semantic segmentation of RS urban scene images. Valanarasu et al. [44] solved the problem of lack of long-range dependencies in the model due to the inherent inductive bias of the convolutional architecture by using the Transformer as a baseline and improving the self-attention mechanism into a gated axial-attention model. TransUnet [41] is a combination of Transformer and UNet, which uses Transformer to process the CNN feature map into a sequence, captures the global information with the help of a self-attention operation, upsamples this information, and then fuses it with the high-resolution feature map, which effectively improves the segmentation task and achieves accurate localization. Inspired by the UNet architecture, the authors of Swin-UNet [30] constructed a U-shaped structure based purely on the Swin transformer, which replaces the traditional convolutional feature extraction encoder with Swin transformer to extract features. ST-UNet [42] not only embedded the Swin transformer into the classical CNN-based UNet but also proposed three new strategies to enhance the feature representation of the occluded objects and reduce the loss of detailed information. Conformer [43] is a hybrid network structure that relies on feature coupling units to enhance the learning of feature representations by combining convolutional operations and self-attentive mechanisms interactively and using a parallel structure to fuse local and global feature representations at different resolutions.

C. Semantic Segmentation of RS Images Based on Decoupling

In the era of Big Data, deep learning, known for its efficient autonomous implicit feature extraction capability, has triggered a boom in the new generation of artificial intelligence, yet the unexplainable black box behind it has become a key bottleneck problem limiting its further development. Therefore, the idea of decoupling is crucial, and decoupled representation learning decouples multilevel and multiscale data information from different perspectives, prompting deep learning models to perceive data autonomously like humans, and gradually becoming an important new research method. We classify the decoupling strategies in the semantic segmentation domain into the following three main types: Decoupling using the coarse-to-fine paradigm, decoupling using intraclass and extraclass relations, and decoupling using edge supervision. To decouple the paradigm from coarse-to-fine, ACFNet [45] used attention to first perform coarse segmentation results on the original image and then uses the coarse segmentation results to calculate the class center of each class to correct the misclassified classes. CDGCNet [46] used coarse segmentation predictions as class masks to extract node features and performs dynamic graph convolution

to learn interclass feature aggregation, which can effectively exploit long-term contextual dependencies and aggregate usage information to better predict pixel labels. CCANet [47] proposed a new class constraint following a coarse-to-fine paradigm of attention depth network that enables the formation of class information constraints with explicit remote contextual information. For decoupling using intra and extraclass relationships, CGFDN [48] encoded cooccurrence relationships between different class objects in a scene as convolutional features and infers segmentation results based on the decoupled features. Glove [49] could consider cooccurrence relationships between different words in the encoding process. Michieli and Zanuttigh[50] proposed a continuous learning scheme shaping the latent space to reduce forgetting while improving the recognition of new classes. For decoupling using edge supervision, Li et al.[51] proposed a new semantic segmentation paradigm by explicitly sampling pixels from different parts (body or edge) and further optimizing the body features and remaining edge features of the target object obtained under decoupled supervision. BGCNet [20] used the BGC module to guide the construction of graphs using node features and boundary predictions After convolving the graph, the inferred features and the input features are fused to obtain the segmentation results. Nie et al. [52] proposed scale-relation joint decoupling network by simultaneously considering decoupling scales and decoupling relationships to excavate more complete relationships of multiscale RS objects.

Inspired by these excellent works, we adopt a semantic clustering module with scale-separation to provide semantic features to guide the SD of the Swin-UNet, which ensures the interaction of global contextual information while fully exploiting the coupled target object features. To the best of the authors' knowledge, the proposed SSDT is the first to extract the semantic information within the image to guide the Swin-UNet network applied to the RS image segmentation task, which makes up for the shortcomings of the traditional Swin-UNet and improves the segmentation accuracy.

III. Proposed Method

In this section, we first introduce the general structure of the proposed SSDT and describe the motivation and architecture involved. Next, five important modules in SSDT are introduced, namely, SPE, SFE, SD, SDT, and MFFD. Finally, we explain the loss function used for network training.

A. Overview of the Proposed SSDT

To address the challenge of spatial scale coupling of semantic information features presented in Chapter 1, we propose SSDT, a network that uses scale-separated semantic information for decoupling as a way to guide the Swin Transformer in extracting features. The network consists of five modules, SPE, SDT, SFE, SD, and MFFD. In particular, SPE transforms the original image into a linear mapping embedding sequence of three scales; SD divides pixels into different semantic clusters based on the scale information extracted by SFE through clustering methods to further obtain scale information-rich semantic features; SDT extracts intraclass feature interdependencies and reduces the association of interclass features by calculating the similarity between semantic features and image features, the core of which is Decoupled Attention. Finally, MFFD is proposed to fuse salient features from different perspectives to further enhance the feature representation.

Specifically, the framework SSDT is shown in Fig. 2. For the input original image X, it is divided into two branches, which are, respectively, used to construct global context information and extract semantic features of scale-separation. In the upper branch, the original image X is sliced at different scale sizes by SPE to generate three scales of linear embedding sequences E^i for global contextual representation modeling, where $i = \{\text{small}, \text{medium}, \text{big}\}$. Subsequently, E^i will undergo two stages of processing to obtain the attention feature F^Z: Stage I is to input the linear mapping embedding sequence E^i of three scales generated by SPE to the traditional Transformer block can get the feature F^i, and the feature F^i of three scales, big, medium and small, are merged to get multipatch feature F^1, which is used to extract the deep representative information of the image. The core of Stage 2 is decoupled attention, the details of which are the similarity calculation between the linear mapping embedding sequence E^i and output of SD module semantics S_c^i to obtain scale semantic information-rich attention features Z_c^i, after merged into attention features F^Z for saliency extraction of semantic information on different scale features. Second, the lower branch goes through DCNN and then enters SFE to get multiscale feature F^X, which can be modeled subscale in the feature space of scale entanglement and solve the scale entanglement problem effectively. After that, the SD module is based on F^X, the pixel is divided into different semantic clusters by clustering method to generate semantics S_c^i (the cth semantic feature on the ith scale), and then concat according to different semantics to get multisemantic feature F^S. Finally, MFFD was used to integrate significant features from different perspectives to further improve the representation ability of features, including the output attention feature F^Z of the SDT module, the output multiscale feature F^X of the SFE module, and the output multisemantic feature F^S of the SD module, to obtain the final output feature $F = [F^Z, F^X, F^S]$.

B. Scale-Separated Patch Extraction

Since context information is the most critical factor to improve semantic segmentation performance, this article proposed that the SPE module cut patches of different sizes to match different sensitivity fields of the SFE module. SPE divides the input image $X \in R^{H\times W\times C}$ into three scales p^i can obtain $N = HW/(p^i)^2$ image blocks, where $i = \{\text{small}, \text{medium}, \text{big}\}$, p^i is the length and width of small image blocks, and they are mapped into a linear projection sequence, represented as follows:

$$E^i = [e_1^i + p_1^i, e_2^i + p_2^i, \ldots, e_N^i + p_N^i] \tag{1}$$

where e_1^i is image embedding and p_1^i is image position embedding. Finally, we take the output of SPE: The sequence of embedding at three scales (big, medium, and small), as the input of the SDT for the next operation.

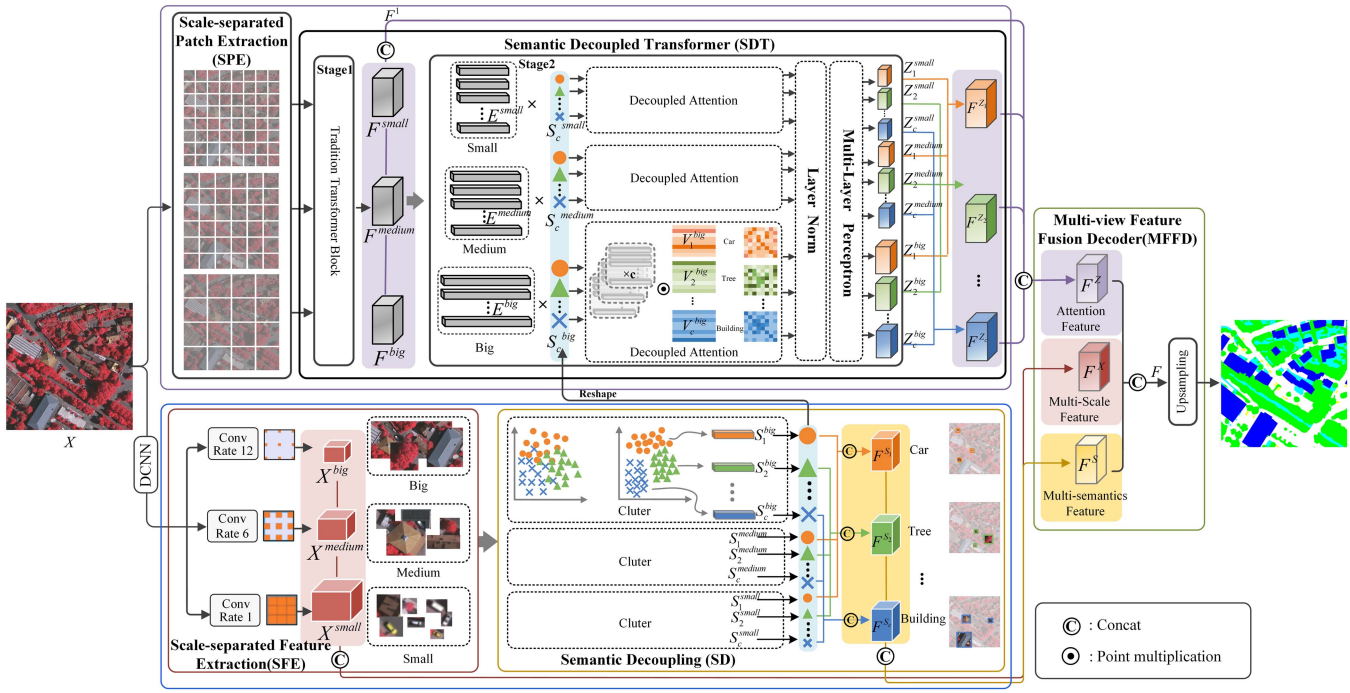

Fig. 2. Architecture of the proposed SSDT. The framework is divided into two branches, which are used to construct global contextual information and extract scale-separated semantic features, respectively. First, the upper branch slices the original image X into patches of different scales by SPE to generate a linear mapping of three scales. After that, the semantic features obtained from SD are imported into SDT for guided modeling to obtain attention features. Second, the lower branch goes into SFE after DCNN to get the scale feature, and the pixels are divided into different semantic clusters by the K-means clustering method based on this scaling feature in the SD module to get the semantic features rich in scale information. Finally, MFFD is utilized to integrate the meaning features from different perspectives, including the attention features output from the SDT module, the multiscale features output from the SFE module, and the multisemantic features output from the SD module, to obtain the final output features.

C. Semantic Decoupled Transformer

SDT is proposed to solve the problem that the scale of semantic information in RS images is seriously entangled. Based on the deep network characterization of DCNN, multiple scale features can be obtained through convolutional layers with different void convolution rates. Traditional feature extraction methods can directly extract semantic information in RS images, but their methods cannot solve the problem of severe coupling of various semantic scales. Compared with traditional methods, our method uses scale modeling and separate SD for each scale information, which can obtain multiperspective and multiscale semantic information and further assist the Swin-UNet encoder to carry out accurate feature characterization.

The module includes two stages, the traditional Transformer block, and decoupled attention. Stage 1: The sequence $E^i \in R^{N\times(p^i)^2\,C}$ obtained from SPE can be input into the traditional Transformer block to obtain output featuresF^i, where $i = \{\text{small}, \text{medium}, \text{big}\}$. And the feature F^i of three scales, big, medium, and small, are merged to get multipatch feature F^1, as shown below:

$$F^1 = [F^{\text{Small}}, F^{\text{Medium}}, F^{\text{Big}}]. \tag{2}$$

The essential module in the traditional Transformer block is the MSA module, which consists of multiple self-attention mechanisms, whose inputs include three vectors $Q^i, K^i, V^i \in R^{N\times(p^i)^2\,C}$, as shown below:

$$Q^i = E^i W_Q, K^i = E^i W_K, V^i = E^i W_V \tag{3}$$

where W_Q, W_K, and W_V are learnable parameters and E^i is a scale-separated linear mapping embedded sequence. The self-attention mechanism consists of a calculation between three vectors, calculated as follows:

$$\text{MSA}(Q^i, K^i, V^i) = \text{softmax}\left(\frac{Q^i\left(K^T\right)^i}{\sqrt{d}}\right)V^i \tag{4}$$

where d is the dimension of vector K. The output features A^i_{l-1} of the multiple attention mechanism at layer $l-1$ is sent into the multi layer perceptron (MLP), and the layer norm (LN) is applied before each block. After the residual connection, the output feature of the coding region can be obtained by loop l times. The calculation formula is as follows:

$$A^i_{l-1} = \text{MSA}(\text{LN}(F^i_{l-1})) + F^i_{l-1} \tag{5}$$

$$F^i_l = \text{MSA}(\text{LN}(A^i_{l-1})) + A^i_{l-1}. \tag{6}$$

The network architecture of decoupled attention is shown in Fig. 3. It is different from the traditional Transformer module. For the traditional Transformer module, the three vectors Q^i, K^i, and V^i of the attention mechanism are composed of multistream linear mapping embedded sequences, whereas the vectors $\hat{Q}^i, \hat{K}^i, \hat{V}^i \in R^{N\times(p^i)^2\,C}$ of decoupled attention add

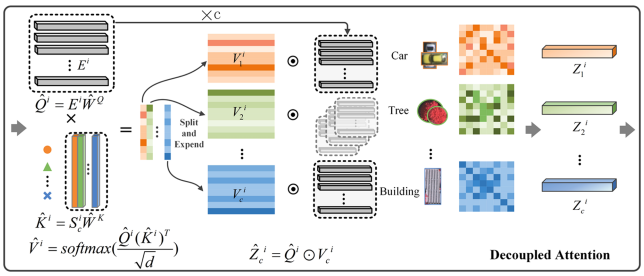

Fig. 3. Architecture of the proposed decoupled attention. This module is different from the traditional MSA module, the vectors $\hat{Q}^i$, $\hat{K}^i$, and $\hat{V}^i$ of decoupled attention add the semantic information of scale-separation extracted from the original image. Based on the multistream original image sequence $\hat{Q}^i$ and the scale-information-rich semantic sequence V_c^i, dot product calculation is carried out to obtain the scale-semantic information-rich attention feature Z_c^i. Then, the three scales Z_c^i of each semantic are combined into c attention features F^{Zc}.

the semantic information of scale-separation extracted from the original image, the calculation process is as follows:

$$\hat{Q}^i = E^i \hat{W}_Q, \hat{K}^i = S_c^i \hat{W}_K, \hat{V}^i = \text{softmax}\left(\frac{\hat{Q}^i \left(\hat{K}^i\right)^T}{\sqrt{d}}\right) \tag{7}$$

where $\hat{W}_Q$ and $\hat{W}_K$ are learnable parameters, E^i is a scale-separated linear mapping embedded sequence, and $S_c^i \in R^{N\times(p^i)^2\,C}$ is the c semantics with rich scale information output by the SD module after reshaping to the linear embedding sequence with the same width as E^i. $\hat{V}^i \in R^{N\times c}$ is a linear embedding sequence of c semantic information obtained from the calculation of similarity between $\hat{Q}^i$ and $\hat{K}^i$. Different from the traditional Transformer module, we conduct Split and Expend on vector $\hat{V}^i$, that is, divide $N \times c$ into $N \times 1$ of c 1-D semantic information and then expand it to $N \times (p^i)^2\,C$. The extended linear embedding sequence $V_c^i \in R^{N\times(p^i)^2\,C}$ of each semantic can be obtained, and the calculation process is as follows:

$$V_c^i = \text{Expend}(\text{Split}(\hat{V}^i)) \tag{8}$$

Based on the scale-separated original image sequence $\hat{Q}^i$ and the scale-information-rich semantic sequence V_c^i, dot product calculation is carried out, as shown in (14), to obtain the scale-semantic information-rich attention feature Z_c^i. Then, the three scales Z_c^i of each semantic are combined into c attention features F^{Zc}, and the calculation process is as follows:

$$Z_c^i = \hat{Q}^i \odot V_c^i \tag{9}$$

$$\begin{cases} F^{Z^1} = [Z_1^{\text{Small}}, Z_1^{\text{Medium}}, Z_1^{\text{Big}}] \\ \cdots \\ F^{Z^c} = [Z_c^{\text{Small}}, Z_c^{\text{Medium}}, Z_c^{\text{Big}}] \end{cases} \tag{10}$$

$$F^Z = [F^{Z^1}, \ldots, F^{Z^c}]. \tag{11}$$

D. Scale-Separated Feature Extraction

The SFE module is proposed for scale adaptation with SPE, classifying the scale sentences as Big, Medium, and Small, and the module is to obtain multiple-scale features $X \in R^{H\times W\times C}$ by passing the original image through the convolution layer of different void convolution rates after passing through DCNN. The multiscale feature F^X can be obtained by combining the three scale features X^i as follows:

$$X^i = \text{Atrous}(X) \tag{12}$$

$$F^X = [X^{\text{Small}}, X^{\text{Medium}}, X^{\text{Big}}] \tag{13}$$

where AtrousConvRate $= \{1, 6, 12\}$.

E. Semantic Decoupling

The SD module is proposed to improve the compactness of intraclass features and expand the dispersion of interclass features. The feature information based on scale-separation divides pixels into different semantic clusters by clustering method. In the process of clustering, for intraclass features, pixels in the same class are close to the clustering center and the intraclass distance is shortened, thus realizing robust intraclass modeling. For interclass features, since there are different clustering centers between classes, pixels repel each other, and the distance between classes is elongated. Therefore, semantic features can be fully utilized to excavate the differences between classes. Compared with the traditional feature extraction method using a CNN, the clustering method is used to extract semantic information. Our method can realize parameterless training and generate feature representations of each category in the context of less computation, thus realizing SD better. SD module is based on scale features $X^i \in R^{H\times W\times C}$. The original image at each scale is divided into different semantic clusters by the clustering method. The c semantic S_c^i on the ith scale is output, and then the three scales S_c^i of each semantic are combined into c semantic features F_c^S.

$$S_c^i = \text{Cluter}(X^i) \tag{14}$$

$$\begin{cases} F^{S^1} = [S_1^{\text{Small}}, S_1^{\text{Medium}}, S_1^{\text{Big}}] \\ \cdots \\ F^{S^c} = [S_c^{\text{Small}}, S_c^{\text{Medium}}, S_c^{\text{Big}}] \end{cases} \tag{15}$$

$$F^S = [F^{S^1}, \ldots, F^{S^c}]. \tag{16}$$

F. Multiview Feature Fusion Decoder

The multiview feature fusion decoding module integrates the saliency features of different views, including the output attention feature F^Z of the SDT module, the output multiscale feature F^X of the SFE module, and the output multisemantic feature F^S of the SD module, to obtain the final output $F = [F^Z, F^X, F^S]$.

The SSDT architecture proposed in this article extracts the scale semantic information-rich attention features by calculating the similarity between the semantic features rich in scale information and the original image features, representing the pixel-level features with the similarity features, and decoupling the coupled semantic information one by one. It can be seen that, compared with the traditional Transformer model, the feature representation of the SSDT guided by the semantic information of similarity is no longer a single image pixel-level information,

but establishes the correlation between semantic features and image features. The model is more robust and improves the accuracy of the model's description of coupling target objects.

G. Loss Function

In this article, the standard multiclassification crossentropy loss function is adopted, which is expressed as follows:

$$L_{\mathrm{CE}} = \frac{1}{n}\sum_{i=1}^{n}(-\hat{Y}_i \log(Y_i) - (1-\hat{Y}_i)\log(1-Y_i)) \tag{17}$$

where n refers to all pixels of the RS image, Y_i is the prediction result generated by the model, and $-\hat{Y}_i$ is the multiclassification label.

IV. Experimental Results and Analysis

A. Datasets

The effectiveness of the SSDT is tested using the International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam dataset and the ISPRS Vaihingen dataset [54].

Potsdam: The Potsdam dataset contains 38 true orthophotos (TOP) images and their corresponding DSMs, which are obtained from a historic city. The spatial sizes of the data files are 6000 × 6000 pixels, and the ground sampling distance (GSD) is 5 cm. There are four spectral bands in each TOP image, including near-infrared, red, green, and blue bands, and one band in each DSM. Note that we use only the red, green, and near-infrared channels in our experiments. We utilize 24 images for training and the remaining 14 images for testing. The Potsdam dataset is labeled according to seven semantic types, which include Impervious Surfaces (white), Buildings (blue), Low Vegetation (cyan), Trees (green), Cars (yellow), Clutter (red), and Undefined (black).

Vaihingen: The Vaihingen dataset contains 33 TOP images and their corresponding DSMs, which are obtained from a small village. The spatial sizes of the data files are 2494 × 2064 pixels, and the GSD is 5 cm. Different from the Potsdam dataset, there are three spectral bands in each TOP image, including near-infrared, red, and green bands, and one band in each DSM. We utilize TOP tiles in our experiments without the DSMs. We utilize 16 images for training and the remaining 17 images for testing. The Vaihingen dataset is split into the same seven categories as those of the Potsdam dataset.

LoveDA: The LoveDA (Land-cOVE dataset for domain adaptation) [55] dataset was created by the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing at Wuhan University. The purpose of this dataset is to promote semantic segmentation and transfer learning tasks. It contains 5987 high-resolution images with 0.3 m resolution and 166 768 annotated semantic objects from three different cities: Nanjing, Changzhou, and Wuhan. The LoveDA dataset involves two different domains, namely, urban and rural, which results in complexity and diversity, such as multiscale objects, complicated background samples, and inconsistent class distributions.

B. Evaluation Metrics

The performance of the SSDT is evaluated by using pixel accuracy (PA), mean pixel accuracy (MPA), and mean intersection over union (mIoU). In addition, we use the F1 score (F1) and the frequency-weighted intersection over union (FWIoU) to further evaluate the network performance, where F1 is a comprehensive indicator considering both precision and recall, and FWIoU is an improvement of mIoU, taking into account the frequency of each class. Among all evaluation metrics, mIoU is the most commonly used metric due to its simplicity and strong representation.

C. Implementation Details

In this section, we focus on the implementation details of the proposed method. To train our network, we crop the image into 1000 random patches of 256 × 256 space size. Random flip or mirror for data expansion to better train the network. In addition, we use argumentation library [56] for enhanced data, and all the training images belonging to one become [0.0, 1.0]. The K-means algorithm is unsupervised learning, with K=7 per data set. In addition, p parameters corresponding to small and big in MIPS are 4/8/16, respectively. All models were implemented using the PyTorch and Adam optimizer with a learning rate of 1e-5. We set the batch size to 10 and train the model with about 300 epochs. All experiments were conducted on an NVIDIA 2080Ti GPU server.

D. Baselines

FCN [6]: This network can save spatial information, can realize point-to-point learning and end-to-end training, and greatly reduce the time cost compared with CNNs.

UNet [7]: This network is an encoder-decoder structure that consists of a contracting path and an expanding path. The contracting path can extract abstract features, whereas the expanding path restores the location information.

SegNet [8]: The network consists of an encoder network, a corresponding decoder network, and a pixel-level classification layer. The decoder uses the pooled index calculated in the maximum pooling step of the corresponding encoder to perform nonlinear upsampling, which reduces the number of parameters and computation and eliminates the need for learning upsampling compared with deconvolution.

PSPNet [10]: The pyramid scene parsing network is applied to capture different subregion representations, followed by upsampling and concatenation layers to form the final predictions.

DeepLabV3+[14]: To introduce multiscale information in Deeplabv3+, the main body of the encoder is DCNN with hole convolution. Then, there is the atrous spatial pyramid pooling module with atrous convolution. Compared with DeepLabv3, v3+ introduces the Decoder module, which further integrates the low-level features with the high-level features.

LoG-CAN [31] is a local-global class-aware (GCA) network for semantic segmentation of RS images, which combines local and global contextual information to improve the segmentation performance of RS images.

TABLE I
COMPARATIVE EXPERIMENT RESULTS ON POTSDAM DATASET

Method	Impervious Surface	Building	Low Vegetation	Tree	Car	PA(%)	MPA (%)	mIoU (%)	FWIoU (%)	F1 score
FCN [6]	78.09	85.80	70.63	66.58	75.07	84.26	73.04	75.23	73.01	70.80
UNet [7]	75.33	80.67	70.92	66.70	75.88	83.18	75.05	73.90	71.45	70.15
SegNet [8]	78.61	84.73	67.79	63.72	78.13	84.08	73.23	74.60	73.34	70.88
PSPNet [10]	78.00	85.80	70.58	57.72	76.02	83.10	77.28	73.62	74.11	65.47
DeeplabV3+ [14]	76.97	90.99	76.54	72.41	80.18	87.20	74.48	80.62	78.75	72.37
LoG-CAN [31]	83.59	**92.35**	75.97	71.31	80.18	88.01	78.53	81.49	78.97	**76.54**
MSCG-Net [24]	76.66	84.93	69.63	65.65	77.98	86.25	**83.93**	73.18	74.97	72.47
DANet [18]	76.54	81.34	69.62	62.95	77.20	82.90	72.70	73.53	71.87	69.38
LANet [16]	81.66	87.99	74.97	67.64	81.66	86.56	75.58	78.78	75.94	74.52
CCANet [53]	81.97	86.04	74.34	70.19	70.68	85.10	68.73	76.64	75.75	67.78
ViT [25]	71.19	77.18	72.50	51.98	56.79	79.53	63.67	65.93	69.38	60.67
UNetformer [40]	83.33	90.66	77.05	70.54	82.53	87.54	76.12	80.82	77.17	75.97
Swin Transformer [29]	67.61	74.88	58.78	45.08	56.38	75.80	60.74	60.54	62.68	58.57
Swin-UNet [30]	84.79	89.45	78.01	73.77	**83.42**	88.35	78.90	81.89	79.30	76.02
SSDT(Ours)	**84.88**	91.90	**78.79**	**75.20**	82.97	**88.45**	77.49	**82.75**	**79.55**	74.78

The bold font represent the optimal values for the experiment.

MSCG-Net [24]: This network is proposed based on GCN, which uses multiple views to explicitly utilize rotation invariance in airborne images, fuses global context information of multiple views, and verifies the influence of multiple angles on RS image segmentation.

DANet [18]: This network captures feature dependencies in spatial dimension and channel dimension based on the self-attention mechanism, and adds two kinds of attention modules to dilated FCN to model semantic dependencies in spatial dimension and channel dimension, respectively.

LANet [16]: This network obtains the degree of interlocation dependency by calculating the pixel-to-pixel similarity and incorporating the neighboring pixel information into the computed pixels, thus freeing the fusion of global information from the limitation of image distance.

CCANet [53]: This network proposes a new attention depth network that follows a coarse-to-fine paradigm for class constraints, which enables the formation of class information constraints to obtain clear remote contextual information.

ViT [25]: ViT is a pioneering work based on Transformer for computer vision tasks, completely changing the field of computer vision. Although ViT is an excellent substitute for CNN, it lacks the inherent inductive bias of CNNs, such as translation, which makes its generalization ability poor when training on insufficient data.

UNetFormer [40] uses only the UNet architecture and simply splices the CNN with the Transformer-based decoder, aiming to improve the global feature extraction capability of neural networks on raw images for efficient semantic segmentation of RS urban scene images.

Swin Transformer [29]: The main idea of Swin Transformer is to divide the graph into disjoint regions and only perform self-attention calculations in this window to reduce computational complexity. Especially when the shallow feature map is large, it reduces computational complexity and isolates information transmission between different windows.

Swin-UNet [30]: This network was inspired by the UNet architecture and replaced the feature extraction encoder with a Swin Transformer to extract features, constructing a pure Swin Transformer based U-shaped encoding and decoding structure.

E. Comparison With State-of-The-Art Methods

In this section, we compare our method with nine baselines on the ISPRS Potsdam and Vaihingen datasets. The experimental results are listed in Tables I and II, where the first five columns of data are the results of mIoU in each category, and the last five columns are the experimental results of common indicators. We can conduct the following analysis.

First, the first three rows are FCN-based methods, such as UNet and SegNet, which perform the worst due to the extraction of features without considering low-level semantic features, ignoring shallow detail information and spatial information, as well as the fusion method is too simple and crude. As shown in Table I, the mIoU of UNet is only 73.90% on the Potsdam dataset, and our proposed SSDT is 8.85% higher than it. Second, DeepLabv3+ performs better than the FCN-based approach because it utilizes atrous convolution to achieve multiscale feature mining. However, the multiscale features extracted by atrous convolution or pyramidal pooling are limited, so graph convolution-based methods, such as MSCG-Net are proposed, which are based on GCN to interact pixels with each other by constructing nodes and edges to establish connections and fuse contextual information to obtain better performance. In addition, attention-based mechanisms such as DANet, LANet, and CCANet are proposed, where CCANet proposes a new class constraint following the coarse-to-refine paradigm of attention-depth networks for SD. The above methods have achieved specific effects, but generally not as good as Swin Transformer, Swin Transformer has the best performance among all baselines. Therefore, we proposed a SSDT based on this method. PA, mIoU, and FWIoU of SSDT on the Potsdam dataset are 0.10%, 0.86%, and 0.25% higher than the second-best model Swin Transformer, respectively. On the Vaihingen dataset, PA, MPA, mIoU, and FWIoU increased by 2.01%, 17.03%, 2.21%, and 3.62%, respectively. It proves the SSDT is fully effective. We also noticed that SSDT produced lower mean F1 scores, which we analyzed due to the imbalanced category in the test data. Therefore, when small objects like cars are accurately segmented, the SSDT model has higher mIoU values, whereas larger objects like bare ground with accurate segmentation have

TABLE II
COMPARATIVE EXPERIMENT RESULTS ON VAIHINGEN DATASET

Method	Impervious Surface	Building	Low Vegetation	Tree	Car	PA(%)	MPA (%)	mIoU (%)	FWIoU (%)	F1 score
FCN [6]	79.53	83.31	66.86	69.60	62.22	85.59	77.98	72.31	75.26	70.27
UNet [7]	81.41	84.69	68.94	68.37	61.11	86.76	72.31	72.90	77.16	84.03
SegNet [8]	80.68	85.11	69.78	64.75	61.70	86.08	71.30	72.40	76.10	83.68
PSPNet [10]	79.75	84.40	62.26	6409	4976	83.82	80.44	68.11	72.98	80.36
DeeplabV3+ [14]	85.32	84.13	72.05	69.25	68.19	88.42	83.88	76.65	79.76	75.46
LoG-CAN [31]	84.71	87.48	71.79	**72.60**	69.37	88.61	73.78	77.19	80.02	**86.93**
MSCG-Net [24]	81.04	85.70	67.02	68.25	63.94	86.25	72.28	73.19	76.34	70.76
DANet [18]	81.72	82.10	64.30	63.70	50.15	84.56	69.30	68.40	74.07	80.60
LANet [16]	84.97	86.20	68.48	67.41	65.68	87.20	72.03	74.55	77.85	85.12
CCANet [53]	81.12	**88.49**	70.36	70.55	51.86	87.22	71.72	72.48	78.06	83.42
ViT [25]	69.61	76.46	60.47	64.93	37.49	80.53	76.39	61.79	67.87	75.47
UNetformer [40]	85.00	88.10	71.33	67.44	70.04	88.00	73.49	76.38	79.18	86.35
Swin Transformer [29]	76.61	75.09	64.36	65.66	46.46	82.49	67.01	65.64	70.57	78.71
Swin-UNet [30]	84.24	88.03	69.09	70.83	68.73	87.71	72.87	76.18	78.67	86.24
SSDT(Ours)	**86.66**	87.70	**74.54**	72.18	**70.43**	**89.73**	**89.90**	**78.30**	**81.93**	75.13

The bold font represent the optimal values for the experiment.

TABLE III
COMPARATIVE EXPERIMENT RESULTS ON LOVEDA DATASET

Method	PA (%)	MPA (%)	mIoU (%)	FWIoU (%)	F1 score
FCN [6]	67.75	73.62	57.04	41.97	56.49
UNet [7]	62.32	67.40	47.61	39.12	50.59
SegNet [8]	61.36	64.55	48.80	42.07	47.20
PSPNet [10]	57.27	43.93	46.62	32.40	44.32
DeeplabV3+ [14]	66.59	63.67	55.98	43.47	59.92
LoG-CAN [31]	63.03	74.98	55.77	38.68	58.92
MSCG-Net [24]	60.28	52.03	48.38	37.17	49.74
DANet [18]	51.90	51.66	39.34	25.59	47.95
LANet [16]	62.11	74.09	51.05	44.18	51.38
CCANet [53]	63.63	67.07	55.63	31.40	59.38
ViT [25]	60.68	46.97	46.92	36.94	52.99
Swin Transformer [29]	51.41	55.48	42.74	20.93	48.08
SSDT(Ours)	**72.84**	**76.72**	**56.72**	**46.42**	**67.99**

higher F1 scores. These results also indicate that the proposed SSDT is superior in processing small categories of RS objects.

In addition, we also verify the proposed method on a large-scale dataset LoveDA, and the relevant results are shown in Table III. Compared with the best segmentation model LoG-CAN, the proposed SSDT achieved a 2.32% improvement in the MPA evaluation metric. It can also be noted that considering the Transformer-based method, our network exceeds ViT by 20.89% with a significant promotion. Thus, the above data sufficiently demonstrate the effectiveness of the SSDT mechanism.

F. Performance of SFE, SPE, SD, and SDT

In this section, we conduct a set of experiments to verify the effectiveness of the proposed modules SFE, SPE, SD, and SDT, as shown in Table IV. It is worth noting that the first row of data in the table represents the initial semantic segmentation network without any module added, i.e., the baseline Swin Transformer. Based on Table IV, several sets of observations can be obtained.

By comparing the first three rows of Table IV, it can be observed that both the scale-separation modules SFE+SPE and the SD module contribute to the enhancement of model performance. The mIoU evaluation metric is improved by 1.84% and 2.00%, respectively, when compared with the Swin Transformer method. Besides, as shown in the last two rows, the SDT module enhances model accuracy by 1.39% for the mIoU evaluation metric, thus, proving the effectiveness of the SDT module. However, after comparing the results of SFE+SPE and SFE+SPE+SD, we surprisingly found that integrating SD resulted in a decrease in semantic segmentation accuracy. We found that the reason for the reduced accuracy in SFE+SPE+SD is due to considering both scale and SD only on the CNN architecture instead of the Transformer architecture, which causes cognitive confusion in the segmentation, eventually reducing the performance of the model. Besides, we believe that the lower performance of SD+SDT compared with SD is because the labels in SD cannot model different scale semantic information, resulting in poor robustness of the labels and, thus, failing to supervise SDT effectively. In summary, through analyzing the above data, it can be concluded that every module is indispensable, and as indicated in the last row of the table, the segmentation performance reaches its peak by integrating all modules.

G. Performance With Different Clustering Methods

To verify the effectiveness of the SD module, we conducted the following experiments, as shown in Table V. Here, we briefly explain the various abbreviated clustering methods in the table. The first row of AGNES is hierarchical clustering. The specific steps are as follows:

1) Each object is regarded as a class and the minimum distance between the two pairs is calculated;
2) Merge the two classes with the smallest distance into a new class;
3) Recalculate the distance between the new class and all classes;
4) Repeat (2) and (3) until all classes are finally merged into one class in which the number of output cluster partition $K = 7$.

The second row of MinBatch K-means is a variant of the K-means algorithm, which uses a small batch of data subsets to reduce the computation time while still trying to optimize the objective function. The specific steps are: 1) Randomly select some data from the data set to form a small batch and assign them to the nearest center of mass; and 2) Update the centroid, where the output number of cluster partition $K = 7$.

TABLE IV
PERFORMANCE OF DIFFERENT MODULE ON VAIHINGEN DATASETS

Method	Impervious Surface	Building	Low Vegetation	Tree	Car	PA (%)	MPA (%)	mIoU (%)	FWIoU (%)	F1 score
Swin Transformer	84.24	88.03	69.09	70.83	68.73	87.72	72.87	76.18	78.67	**86.24**
SFE+SPE	86.37	89.35	72.90	68.67	**70.62**	88.88	84.05	77.58	80.48	82.27
SD	86.22	**90.48**	73.67	68.98	69.14	89.33	73.80	77.70	81.25	87.17
SD+SDT	84.08	87.76	71.76	72.36	69.88	88.60	89.55	77.17	79.94	75.42
SFE+SPE+SD	84.46	87.65	70.97	**72.57**	70.50	88.68	80.30	77.23	80.19	72.47
SFE+SPE+SD+SDT(our SSDT)	**86.66**	87.70	**74.54**	72.18	70.43	**89.73**	**89.90**	**78.30**	**81.93**	75.13

The bold font represent the optimal values for the experiment.

TABLE V
PERFORMANCE OF DIFFERENT CLUSTERING METHODS ON VAIHINGEN AND POTSDAM DATASETS

Clustering Method	mIoU (%)	
	Vaihingen	Postdam
AGNES	77.71	82.57
MinBatch K-means	77.48	82.37
K-means(Ours)	**78.30**	**82.74**

TABLE VI
PERFORMANCE OF DIFFERENT FUSION METHODS ON VAIHINGEN AND POTSDAM DATASETS

Fusion Method	mIoU (%)	
	Vaihingen	Postdam
Element-wise **Add**ition	77.00	82.32
Matrix **Mul**tiplication	76.41	82.07
Concat(Ours)	**78.30**	**82.74**

The data is updated on every small sample set compared with the K-means algorithm. For each small batch, the updated centroid is obtained by calculating the average value, and the data in the small batch is allocated to the centroid. With the increase in the number of iterations, the change of this centroid is gradually reduced until the centroid is stable or the specified number of iterations is reached, and the calculation is stopped. As shown in Table V, the K-means clustering method we used performed best in both datasets for the mIoU metric. Compared with AGNES and MinBatch K-means, K-means is 0.59% and 0.82% higher on the Vaihingen dataset and 0.17% and 0.37% higher on the Potsdam dataset. This is attributable to the following advantages of the K-means algorithm: First, it can determine the classification of some samples based on the categories of fewer known clustered samples; Second, to overcome the inaccuracy of clustering a small number of samples, the algorithm itself has an optimization iteration function, which iterates again on the clusters already obtained to determine the clusters of some samples, optimizing the initial supervised learning of the unreasonable classification of samples. Therefore, we used the K-means clustering method to effectively extract rich semantic clustering information, and the experimental results proved the effectiveness of the method.

H. Performance With Different Fusion Methods

In this section, we conducted three experiments on the Potsdam dataset and the Vaihingen dataset to verify the effectiveness of the fusion methods used within the MFFD module. In this experiment, three different fusion methods, elementwise addition (Add), matrix multiplication (Mul), and concatenation (Concat), were used to fuse the output attentional features F^Z of the SDT module, the output multiscale features F^X of the SFE module and the SD module's output multisemantic features F^S are fused.

As shown in Table VI, the first row is elementwise addition, which performs direct element-by-element addition of features. The second row is matrix multiplication, which performs the interaction of information between elements by matrix multiplication. Since the above methods fuse in a too crude way, the performance is not satisfactory in both datasets. The Concat fusion method we used performed best in mIoU metrics on both datasets. Compared with Add and Mul, Concat is 1.30% and 1.89% higher on the Vaihingen dataset and 0.42% and 0.67% higher on the Potsdam dataset, respectively. Therefore, we used the Concat method to effectively fuse the multiview features, and the experimental results proved the effectiveness of the method.

I. Boxplot Analysis of Scale-Separated Results

To verify the necessity of scale-separation within SD, we randomly selected 20 images within the Vahingen dataset for input to the two models, and the resulting mIoU data were generated as boxplots. The difference between the two models is whether the scale-separation operation is added or not. As shown in Fig. 4 boxplot, the vertical axis is the mIoU metric and the horizontal axis is the individual semantics within the dataset. Fig. 4(a) is our proposed SSDT method, which performs the scale-separation operation within the SD and avoids the influence of scale information on the SD; Fig. 4(b) is the method without scale information, which ignores all scale information, does not perform scale-separation, and only performs SD. Boxplot is a statistical graph used as a display of information about the dispersion of a set of data, which can reflect the characteristics of the data distribution, and also allows comparison of the characteristics of the distribution of multiple sets of data. The boxes plotted for each semantic in the figure include the upper edge, lower edge, median, and two quartiles of a set of data; the box connects the two quartiles; the upper and lower edges are connected to the box, and the median is in the middle of the box. The analysis follows.

First, it is obvious from observing the two boxplots that (b) has four more small circles, or outliers, than (a). Outliers in a batch of data deserve attention, and it is very dangerous to ignore the existence of outliers. Including outliers in the process of calculating and analyzing data without eliminating them can

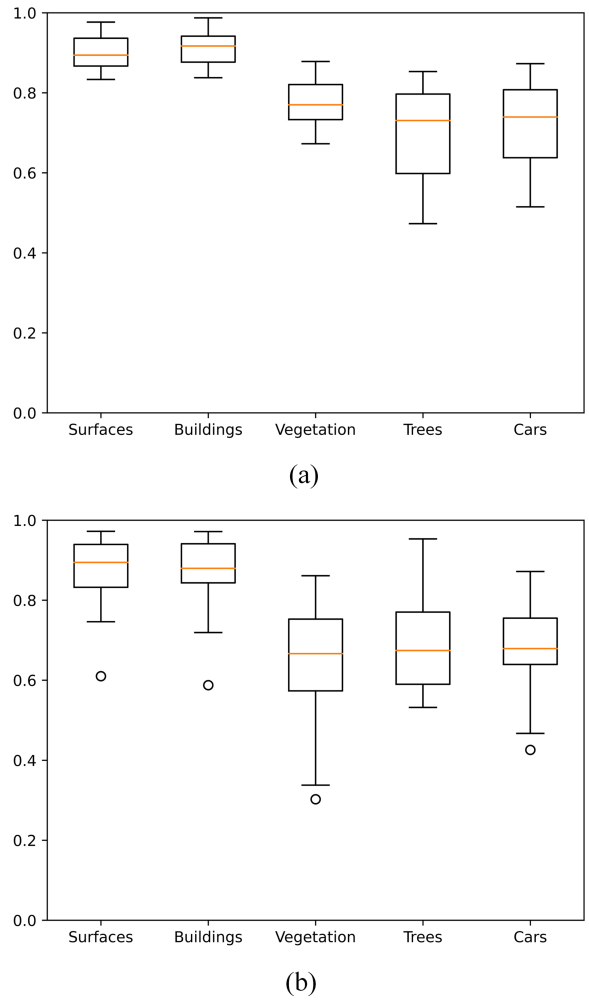

Fig. 4. Boxplot of scale-separated results. (a) Our proposed SSDT method. (b) Method without scale information.

have adverse effects on the results. Second, since the SSDT method incorporates a scale-separation operation, the obtained scale information-rich semantic features are used to guide the Transformer model to extract features, resulting in improved intra-class compactness and smaller intra-class variance, such as the quartile size of (a), which is a shorter box length. Meanwhile, the median of each semantic in (a) is significantly higher than that in (b), indicating that the SSDT method is more stable than the method without scale information, and the overall mIoU index is also higher. In addition, the length between the upper and lower edges of each semantic in (a) is significantly smaller than that of the box in (b), indicating that the distribution of the normal values of mIoU is more concentrated in the SSDT method with the addition of the scale-separation operation. Finally, the experimental results demonstrate the necessity of scale-separation within the semantic decoupling.

J. Qualitative Analysis of the Semantic Segmentation Results

Here, we give the visualizations of the predicted semantic segmentation maps for the Potsdam and Vaihingen datasets. As shown in Figs. 5 and 6, the first column is input raw image, the second column is ground truth, the remaining columns in the middle are the viewable views of each baseline, and the last column presents the viewable views of SSDT in this article.

First, Fig. 5. is the visualization of the Potsdam dataset: the third, fourth, or fifth columns are for the FCN-based approach. For example, due to the simplistic and crude way of UNet fusion, the semantics of "Building" and "Tree" in the black box in the fourth column are not recognized as they should be; the SegNet extraction of features ignores the shallow semantic detail information and spatial information, resulting in severe blurring in the black box in the fifth column and a large number of pixel missegmented within the image. The sixth, seventh, or eighth columns are scale-specific approaches, such as DeepLabv3+ for multiscale feature extraction using null convolution, PSPNet using pyramid pooling, and LoG-CAN using GCA modules and local class-aware modules. However, the scale information extracted by the above methods is limited, for example, the sixth column should have identified only "Low Vegetation," but PSPNet classified both "Tree" and "Low Vegetation" semantics; the two red boxes in the seventh column where "Clutter" should have appeared have been misclassified as another semantic; in the eighth column, the semantics of "Building" in the red box are missing and not fully segmented. Therefore, graph convolution-based methods, such as MSCG-Net (ninth column) are proposed, which is based on GCN to interact pixels with each other by constructing nodes and edges to establish connections and fuse contextual information to obtain better performance. But it cannot optimally solve the semantic coupling problem, as shown in the green box in the figure, there is a messy split in what should be a clear diagram. In addition, the tenth, eleventh, and twelfth columns are the attention mechanism-based methods, DANet, LANet, and CCANet. DANet improves the segmentation accuracy based on the self-attention mechanism to obtain the dependence of features in spatial dimension and channel dimension. LANet calculates the similarity between pixels to obtain the dependence degree between locations so that the fusion of global information is not limited by image distance. The above method achieves specific results, but does not consider the influence of scale information on semantic information, and ignores important contextual information, such as some confusion areas appearing in the yellow box in the figure. Moreover, among several ViT-based models, Swin-UNet performs the best, in which UNetFormer uses only the UNet architecture and simply splices the CNN with the Transformer-based decoder, with unsatisfactory results such as missegmentation in the purple box; nearly half of the pixels in the Swin Transformer's visualization are misclassified as "Impervious Surface." It is worth noting that our proposed SSDT based on Swin-UNet shows extremely high intraclass compactness and interclass relaxation for each semantic class, and also achieves pleasing details in spatial consistency.

Second, Fig. 6 is a visualization of the Vaihingen dataset: Observing the first three rows, it can be seen that some small scale target objects in the red box are ignored. This phenomenon is not only present in the CNN-based network (UNet) but also in the Transformer-based method. This is because the model is

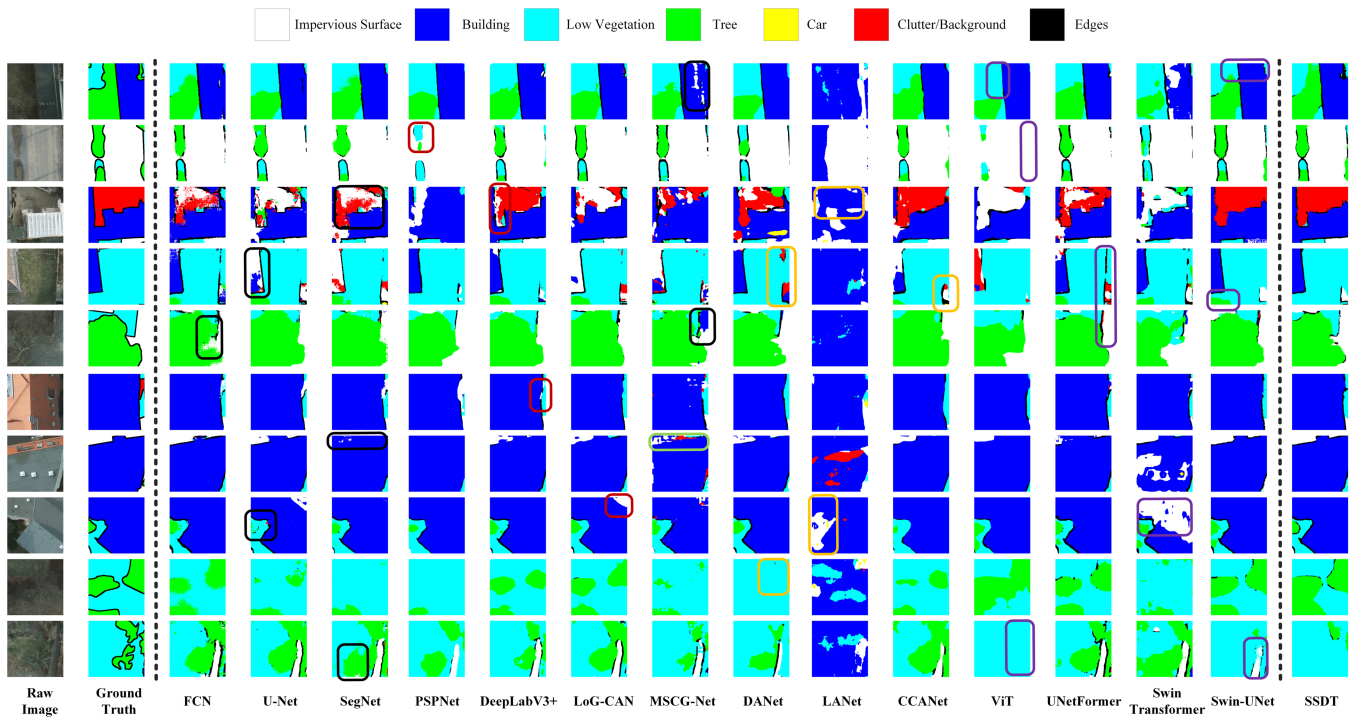

Fig. 5. Examples of semantic segmentation results on the Potsdam dataset.

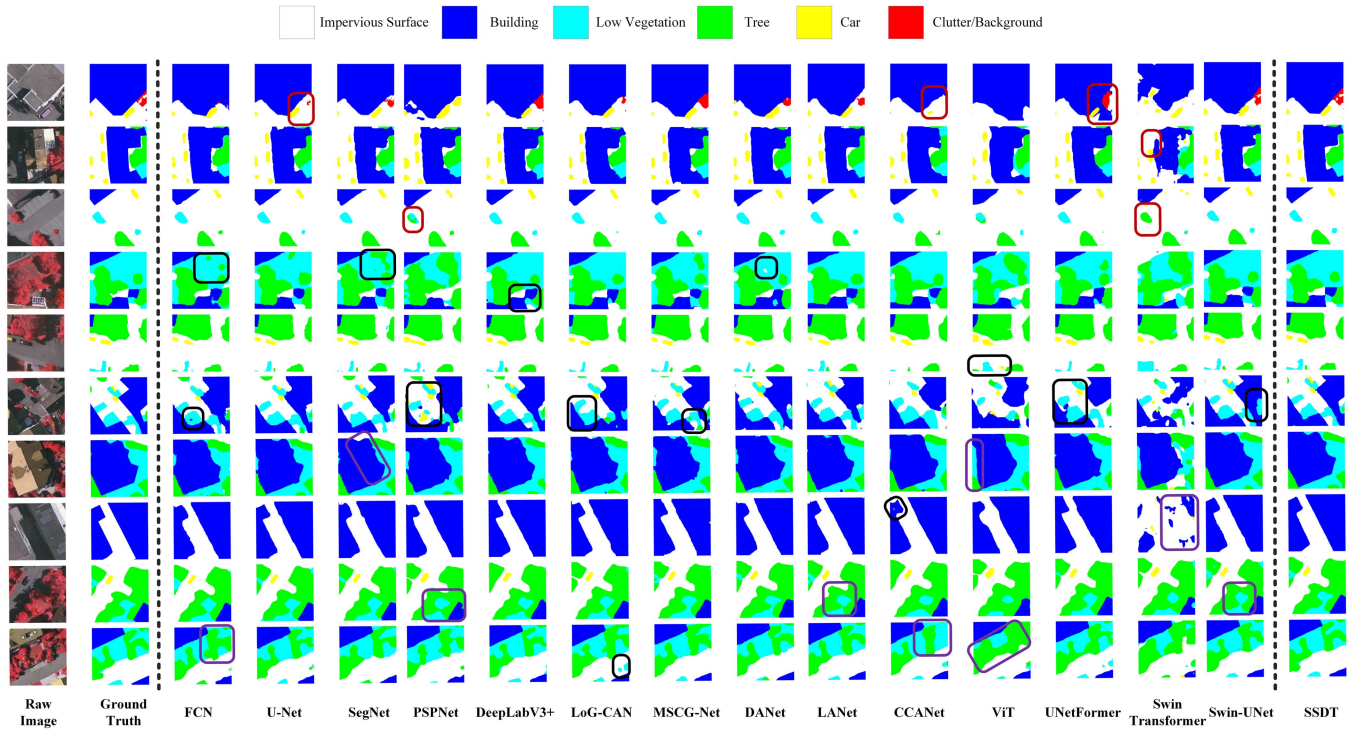

Fig. 6. Examples of semantic segmentation results on the Vaihingen dataset.

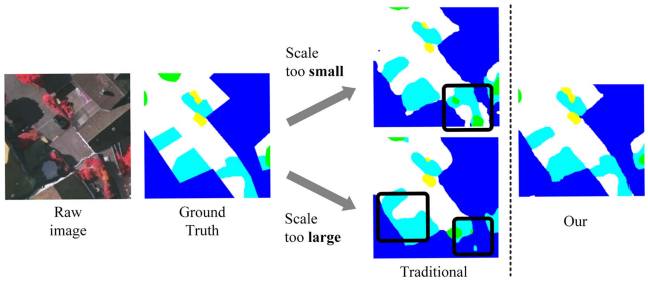

Fig. 7. Examples of visualizations with significant semantic interweaving on the Vaihingen dataset.

modeled for big scale target objects without the scale-separation operation in the SD process, and thus the small scale targets are overwhelmed. Semantic entanglement occurs in the black boxes in the middle three lines, either other semantics are newly recognized, or new semantic connections appear on the original semantics with unclear boundary segmentation. The purple boxes in the last four rows show some confusing results, such as the unclear edge segmentation of the "Tree" semantics, which is a consequence of too little similarity within the same semantics and too much similarity between different semantics in the modeling process. It is worth noting that the method proposed in this article (SSDT) has the most convincing visualization both in terms of SD for each scale size and in terms of boundary detail regions.

Specifically, as shown in Fig. 7, we identified some visualizations with significant semantic interweaving in the Vaihingen dataset to highlight the advantages of SSDT effectively. On the one hand, when the scale is too small, as shown in the above figure, a "Tree" that should not appear appears. On the other hand, when the scale is too large, it causes semantic space entanglement, and the edges that should have been mixed, resulting in a decrease in semantic segmentation results. On the contrary, our proposed method SSDT has significant advantages over the traditional methods mentioned above. Not only does it avoid semantic entanglement that should not occur, but it also has clear edges and achieves better segmentation results.

K. Complexity Analysis

We also performed complexity analysis experiments and show the results in Table VII. The "Time Cost" in the table represents the time for the model to segment an image. As can be seen from the table, FCN has the lowest time complexity with 0.08 due to its simple network architecture. Compared with our model, the training time cost has slightly increased and we analyze that is because we have built a two-layer architecture of CNN and Transformer and achieved joint decoupling of scale and semantics. In addition, we also provide the parameters of the baselines and our SSDT model with ten million measurement units. As can be seen from the table, MSCG-Net has the smallest parameters since the applied GCN involves fewer parameters. In addition, because our model considers scale-separation in both CNN and Transformer architectures, it has relatively large training parameters. In the future, we will continue to work on reducing the computational cost.

TABLE VII
COMPLEXITY RESULTS

Method	Time Cost(s)	Parameters (10 m)
FCN [6]	**0.08**	0.19
UNet [7]	0.12	0.17
SegNet [8]	0.09	0.29
PSPNet [10]	0.11	0.28
DeeplabV3+ [14]	0.12	0.59
LoG-CAN [31]	0.10	0.31
MSCG-Net [24]	0.11	**0.10**
DANet [18]	0.11	0.50
LANet [16]	0.09	0.24
CCANet [53]	0.10	0.59
ViT [25]	0.14	0.89
UNetformer [40]	0.09	0.12
Swin Transformer [29]	0.11	0.88
Swin-UNet [30]	0.08	0.41
SSDT(Ours)	0.18	1.86

The bold font represent the optimal values for the experiment.

V. CONCLUSION

To solve the serious entanglement of semantic information scale in RS images, a new idea is proposed in this article: scale decoupling in the process of SD can effectively avoid the impact of scale coupling on semantic judgment. Meanwhile, the SD module is implemented in the Swin Transformer. This will not only help the Swin Transformer provide effective semantic features but also help make up for the lack of spatial location in the Swin Transformer. We creatively came up with SSDT, which consists of five modules, SPE, SDT, SFE SD, and MFFD, using scale information to divide pixels into different semantic clusters by clustering method, semantic features rich in scale information can be obtained, and semantic features can be used as guidance, and the similarity between image features can be calculated to mining the interdependency of features within the class, and the correlation between features between the classes can be reduced. The problem of serious coupling of semantic information scales in RS images is solved.

We conduct multiple sets of comparison experiments and ablation experiments on the Potsdam dataset and the Vaihingen dataset to verify the effectiveness of the proposed method. Qualitative and quantitative results demonstrate the effectiveness of the SD Transformer framework with scale-separation in RS image segmentation tasks. Specifically, the proposed SSDT outperformed the state-of-the-art Swin Transformer by 0.86% and 2.12% on the Potsdam and the Vaihingen datasets for mIoU evaluation, respectively.

REFERENCES

[1] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 8, pp. 582–596, 2019.
[2] N. Keiller et al., "Exploiting ConvNet diversity for flooding identification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, no. 9, pp. 1446–1450, Sep. 2018.
[3] M. Siam, S. Elkerdawy, M. Jagersand, and S. Yogamani, "Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst.*, 2017, pp. 1–8.

[4] M. Leena and K. Kirsi, "Segment-based land cover mapping of a suburban area—comparison of high-resolution remotely sensed datasets using classification trees and test field points," *Remote Sens.*, vol. 3, no. 8, pp. 1777–1804, 2011.

[5] A. G.-Garcia, S. O.-Escolano, S. Oprea, V. V.-Martinez, and J. G.-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[9] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 6230–6239.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFS," 2014, *arXiv:1412.7062*.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[13] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[15] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[16] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021, doi: 10.1109/TGRS.2020.2994150.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[18] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[19] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[20] H. Hu, J. Cui, and H. Zha, "Boundary-aware graph convolution for semantic segmentation," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 1828–1835.

[21] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8950–8959.

[22] S.-Y. Pan, C.-Y. Lu, S.-P. Lee, and W.-H. Peng, "Weakly-supervised image semantic segmentation using graph convolutional networks," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[23] H. Huang et al., "Graph-BAS3Net: Boundary-aware semi-supervised segmentation network with bilateral graph convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7386–7395.

[24] Q. Liu, M. C. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 44–45.

[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[26] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[27] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[29] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[30] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.

[31] X. Ma et al., "LoG-CAN: Local-global class-aware network for semantic segmentation of remote sensing images," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[32] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012.

[33] S. Wang, Z. Wang, H. Li, J. Chang, W. Ouyang, and Q. Tian, "Accurate fine-grained object recognition with structure-driven relation graph networks," *Int. J. Comput. Vis.*, vol. 132, pp. 137–160, 2023.

[34] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.

[35] Z. Li, H. Guo, Y. Chen, C. Liu, Q. Du, and Z. Fang, "Few-shot hyperspectral image classification with self-supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5517917.

[36] J. Yang, J. Sun, Y. Ren, S. Li, S. Ding, and J. Hu, "GACP: Graph neural networks with ARMA filters and a parallel CNN for hyperspectral image classification," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 1770–1800, 2023.

[37] Y. Zhang, M. Zhang, W. Li, S. Wang, and R. Tao, "Language-aware domain generalization network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501312.

[38] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7287–7296.

[39] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023.

[40] L. Wang et al., "UNetformer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.

[41] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[42] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[43] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Gool, "2021 IEEE/CVF international conference on computer vision (ICCV)," 2021, pp. 367–376.

[44] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2021, pp. 36–46.

[45] F. Zhang et al., "ACFNet: Attentional class feature network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6798–6807.

[46] H. Hu, D. Ji, W. Gan, S. Bai, W. Wu, and J. Yan, "Class-wise dynamic graph convolution for semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.

[47] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4401120.

[48] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2020.

[49] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[50] U. Michieli and P. Zanuttigh, "Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1114–1124.

[51] X. Li et al., "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 435–452.
[52] J. Nie et al., "Scale–relation joint decoupling network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412812.
[53] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 4401120, doi: 10.1109/TGRS.2021.3055950.
[54] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, "ISPRS semantic labeling contest," *ISPRS: Leopoldshöhe*, Germany, vol. 1, no. 4, 2014.
[55] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2021, *arXiv:2110.08733*.
[56] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020, Art. no. 125.

Chengyu Zheng is currently working toward the Ph.D. degree with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China.

Her research interests focus on the processing and analysis of remote-sensing data.

Yanru Jiang is currently working toward the master's degree in engineering with the School of Information Science and Engineering, Ocean University of China, Qingdao, China.

Her main research direction is the semantic segmentation of remote-sensing images.

Xiaowei Lv is currently working toward the master's degree in the discipline of master of engineering, specializing in computer software and theory with the School of information, Renmin University of China, Beijing, China.

His main research interests include remote sensing image semantic segmentation and graph structure mining.

Jie Nie (Member, IEEE) received the B.S. and Ph.D. degrees from Ocean University of China, Qingdao, China, in 2002 and 2011, respectively, both in computer science.

From 2009 to 2010, she was a Visiting Scholar with the University of Pittsburgh, Pittsburgh, PA, USA. . From 2015 to 2017, she was a Postdoctoral Fellow with Tsinghua University, Beijing, China. She is currently a Professor with Ocean University of China. During recent five years, she has authored or coauthored more than 60 papers in international leading and key journals and conferences in the area of artificial intelligence and Big Data analysis such as IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, ACM SIGMM, ACM SIGIR, etc. Her current research interests include the fields of artificial intelligence and visual analysis of marine Big Data.

Dr. Nie was the Guest Editor and Area Chair of many journals and conferences.

Xinyue Liang (Member, IEEE) received the B.S. degree in communication engineering from Beijing Jiaotong University, Beijing, China, in 2014, the M.S. degree in network services and systems in 2016 and the Ph.D. degree in electrical engineering in 2021 from KTH Royal Institute of Technology, Stockholm, Sweden.

She is currently a Lecturer with the Ocean University of China, Qingdao, China, in 2022. Her research interests include distributed machine learning, marine Big Data analytics, and distributed signal analysis.

Dr. Liang was the recipient of the China National Scholarship Fund to study at KTH as a Ph.D. student in 2016.

Zhiqiang Wei (Member, IEEE) received the B.S. degree in mechanical engineering from Shandong University, Jinan, China, in 1992, the M.S. degree in electromechanics from the Harbin Institute of Technology, Harbin, China, in 1995, and the Ph.D. degree in precision instruments from Tsinghua University, Beijing, China, in 2001.

He is currently a Professor with the Ocean University of China, Qingdao, China. His current research interests include the fields of intelligent information processing, intelligent computing of ocean Big Data, and multimedia content analysis and understanding.