Swin Transformer Embedding Dual-Stream for Semantic Segmentation of Remote Sensing Imagery

Xuanyu Zhou, *Student Member, IEEE*, Lifan Zhou, *Member, IEEE*, Shengrong Gong, Shan Zhong, Wei Yan, and Yizhou Huang

***Abstract*—The acquisition of global context and boundary information is crucial for the semantic segmentation of remote sensing (RS) images. In contrast to convolutional neural networks (CNNs), transformers exhibit superior performance in global modeling and shape feature encoding, which provides a novel avenue for obtaining global context and boundary information. However, current methods fail to effectively leverage these distinctive advantages of transformers. To address this issue, we propose a novel single encoder and dual decoders architecture called STDSNet, which embeds the Swin transformer into the dual-stream network for semantic segmentation of RS imagery. The proposed STDSNet employs the Swin transformer as the network backbone in the encoder to address the limitations of CNNs in global modeling and encoding shape features. The dual decoder comprises two parallel streams, namely the global stream (GS) and the shape stream (SS). The GS utilizes the global context fusion module (GCFM) to address the loss of global context during upsampling. It further integrates GCFMs with skip connections and a multiscale fusion strategy to mitigate large-scale regional object classification errors resulting from similar features or shadow occlusion in RS images. The SS introduces the gate convolution module (GCM) to filter out irrelevant features, allowing it to focus on processing boundary information, which improves the semantic segmentation performance of small targets and their boundaries in RS images. Extensive experiments demonstrate that STDSNet outperforms other state-of-the-art methods on the ISPRS Vaihingen and Potsdam benchmarks.**

***Index Terms*—Dual-stream, remote sensing (RS), semantic segmentation, Swin transformer.**

I. Introduction

REMOTE sensing (RS) images have become a crucial source of data for surface information extraction. Advancements in aerospace and sensor technologies have enabled the capture of high-quality RS images, which can be applied in various scenarios [1]. Semantic segmentation is a pixel-level classification that predicts category labels for each pixel in an image. It has garnered widespread attention in RS-related fields. High-quality RS images can capture detailed information, such as texture, shape, and color features in large-scale regions. However, these features pose challenges for semantic segmentation in complex scenarios. Currently, the semantic segmentation of RS images serves as a crucial research area for RS image analysis and provides technical support for various significant RS applications, including urban planning [2], [3], [4], land cover survey [5], [6], environmental monitoring [7], and disaster assessment [8], [9].

In recent years, the rapid development of convolutional neural networks (CNNs) has played a crucial role in driving forward research on image semantic segmentation. The fully convolutional network (FCN) [10] with encoder–decoder architecture, specifically, has emerged as a popular configuration for semantic segmentation owing to its exceptional segmentation performance. Within the CNN-based encoder–decoder architecture, the encoder leverages successive convolution and pooling operations to reduce the spatial sizes of features, thereby enhancing their semantic representation and extracting features [11]. Conversely, the decoder restores the image resolution while simultaneously fusing high-level semantic and low-level spatial information [13].

While the FCN-based method has demonstrated promising results, it still faces significant challenges in the semantic segmentation of RS images. On the one hand, RS images typically encompass a wider spatial range and fewer object categories compared to natural images. Furthermore, it comprises a multitude of ground objects belonging to various categories that share similar spectral and material properties, as well as ground objects that are obscured by shadow. These factors make it more difficult to embed clear global scene information at a global level, leading to severe semantic confusion issues and ultimately segmentation errors in large-scale areas in RS images [14]. On the other hand, RS image-related applications require highly accurate ground object mapping, necessitating precise identification of diminutive objects and boundaries. However, CNN-based models often perform feature downsampling during feature extraction to reduce computational complexity, resulting in the loss of small-scale features and ultimately inadequate semantic segmentation of small targets and their boundaries. [15]. Consequently, more global context and boundary information are requested as clues for semantic reasoning in RS images [16], [17].

Manuscript received 8 June 2023; revised 27 August 2023; accepted 8 October 2023. Date of publication 24 October 2023; date of current version 23 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42071438 and Grant 61972059, and in part by the Natural Science Foundation of Jiangsu Province, Chinaunder Grant BK20221403. *(Corresponding authors: Lifan Zhou; Shengrong Gong.)*

Xuanyu Zhou is with the School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China (e-mail: zxy98_edu@163.com).

Lifan Zhou, Shengrong Gong, Shan Zhong, and Wei Yan are with the School of Computer Science and Engineering, Changshu Institute of Technology, Suzhou 215500, China (e-mail: zhoulifan_rs@163.com; shrgong@cslg.edu.cn; sunshine620@cslg.edu.cn; 120570473@qq.com).

Yizhou Huang is with the School of Electrical and Automation Engineering, Changshu Institute of Technology, Suzhou 215500, China (e-mail: 2024772881@qq.com).

Digital Object Identifier 10.1109/JSTARS.2023.3326967

Due to the locality of convolution, CNNs are limited in their ability to model global information. This limitation is attributable to the fact that each convolutional kernel focuses solely on the local pixels within its receptive field and, therefore, cannot model long-distance dependencies. To address this issue, DeepLabV3+ [18] employed pyramid-structured dilated convolutions [19] to expand the receptive field and improve global information extraction. UperNet [20] incorporated a pyramid of global pooling layers to effectively capture global context information. However, the grid effects of dilated convolutions and the semantic loss induced by the global pooling layer make it challenging to capture comprehensive global context information. In addition, although DANet [21] and DNLNet [22] leverage attention mechanisms to establish long-range dependencies, they are still not liberated from CNNs. In essence, CNNs aggregate global information from local features rather than encoding the global context directly. Therefore, using only CNNs makes it difficult to acquire clear global scene information from RS images with complex backgrounds [23].

Recently, transformer architectures based on self-attention mechanisms have demonstrated remarkable achievements in natural language processing (NLP) [25], offering novel solutions for modeling global relationships. Dosovitskiy et al. [26] proposed the vision transformer (ViT), which employs a pure transformer structure as a feature extractor for the first time to perform image recognition tasks. Zheng et al. [27] introduced the segmentation transformer (SETR), a first semantic segmentation model that builds upon ViT, thus representing the first attempt to apply transformers to semantic segmentation. To address the limitations of SETR, Xie et al. [28] proposed SegFormer, which reduces the number of parameters of the model, eliminates the position encoding, and designs a lightweight yet effective All-MLP decoder. In addition, Liu et al. [29] introduced the Swin transformer, which employs a shifting window strategy to constrain self-attention calculations within nonoverlapping windows while allowing for cross-window information exchange. The Swin transformer has demonstrated significant potential in certain dense prediction tasks and is progressively gaining popularity for semantic segmentation of RS images [16], [30].

In addition, some recent work investigating the properties of transformers indicates the superior encoding of shape features by transformers compared to CNNs. This discovery provides a novel avenue for extracting boundary information. Naseer et al. [31] demonstrated that ViTs can achieve comparable shape recognition capabilities to those of the human visual system, which surpass the performance of CNNs when properly trained to encode shape-based features. The authors also highlight that transformer models exhibit a higher shape bias compared to CNN models with similar parameter counts. Here, shape bias is defined as the fraction of correct decisions based on object shape. Similar conclusions have been drawn by Tuli et al. [32], who investigated the degree of correlation between various visual models and human vision from the perspective of error consistency and discovered that transformers have a higher shape bias than CNNs, which aligns more closely with human errors. In addition, Ghiasi et al. [33] visualized and compared the behavior of transformers and CNNs and observed that both architectures share a common feature where early layers learn low-level features while deeper layers learn high-level object features or abstract concepts. The distinction lies in the fact that transformers can more effectively utilize global context information. These studies effectively demonstrate that transformers are superior to CNNs in encoding shape features, corroborating the potential of extracting boundary information from transformers.

Despite the transformer's significant advantages over CNNs in global modeling and shape feature encoding, current semantic segmentation methods do not completely exploit these benefits. Most transformer-based methods empirically apply certain strategies that are applicable in CNNs to the decoder. However, these strategies were not originally devised considering the characteristics of transformers. This means that the full potential of the transformer cannot be realized. As a result, the decoding process suffers from the semantic loss of global context and boundary information. Specifically, some methods [16], [30] employ the traditional upsampling techniques commonly used in popular decoders, such as linear interpolation and transposed convolution. However, the feature representation for each location on these upsampled feature maps is restored from a restricted receptive field. This constraint limits the ability to restore pixel-wise predictions, resulting in the loss of global context during the upsampling process [34], [35]. In addition, certain methods [36], [37] migrate the decoder strategies that involve dense image representations, wherein shape, color, and texture features are simultaneously processed using CNNs. This simultaneous processing of diverse feature types hampers adequate decoding of shape features, leading to the loss of boundary information.

In this article, we propose a single encoder and dual decoders framework for RS imagery semantic segmentation called STDSNet, which employs the Swin transformer as the network encoder to mitigate the limitations of CNN in global modeling and shape feature encoding. The dual decoder consists of two parallel streams, namely the GS and the shape stream (SS). The GS aims to optimize the utilization of the global context captured by transformers, while the SS focuses on processing shape features to obtain more accurate boundary information. The primary contributions of this article are as follows.

1) We combine the Swin transformer backbone with a CNN-based two-stream network to construct a new single encoder and dual decoders semantic segmentation framework. The dual decoder fully exploits the advantages of the Swin transformer in encoding global context and shape features.
2) In the GS, we utilize the global context fusion module (GCFM) to recover the loss of global context during the upsampling process. We further integrate skip connections and a multiscale fusion strategy to optimize the utilization of global context, effectively ameliorating large-scale regional segmentation errors that arise due to similar features or shadow occlusion in RS images.
3) To filter out redundant feature information, we introduce the gate convolution module (GCM) in the SS. By integrating GCMs at various stages of the encoder, the network can selectively suppress noise layer by layer, allowing it

to focus solely on processing boundary information. This improves the semantic segmentation performance of small targets and their boundaries in RS images.

The rest of this article are organized as follows. Section II provides an overview of related work on STDSNet. Section III presents the comprehensive structure of STDSNet along with specific implementation details. Section IV outlines the experimental setup, shows the experiment results, and discusses them in detail. Finally, Section V concludes this article.

II. Related Work

A. Semantic Segmentation of RS Images Based on CNN

In recent years, despite the success achieved by FCN-based methods in semantic segmentation tasks, their effectiveness in addressing challenging segmentation issues in RS imagery has remained limited. These challenges are posed by ground objects that are of small scale, have high similarity, or suffer from mutual occlusion. To mitigate these challenges, researchers have devised various solutions. Multiscale feature extraction has been employed by some researchers to enhance the network's performance. Zhang et al. [38] utilized a multibranch parallel convolution structure from HRNet [39] to generate feature maps at various scales, thereby strengthening the embedding of scale-related contextual information. Zhang et al. [40] improved the network's ability to express multiscale features by combining three different paths, namely pooling, transpose convolution, and dilated convolution, and learning the optimal combination of different scales of features. Hang et al. [41] proposed a multiscale progressive segmentation network, which gradually segments objects into small, large, and other scales by cascading three subnetworks. This method greatly improves the segmentation of large and small objects.

The attention mechanism is also a noteworthy aspect. Haut et al. [42] used a method that combines attention mechanisms with a residual-structured network in super-resolution RS images, which integrates the high and low-frequency characteristics of RS images and filters out low-frequency surface features that are not useful. Ding et al. [14] proposed a method that combines the patch attention module (PAM) and attention embedding module (AEM), which enhances the embedding of contextual information while enriching the semantic information of low-level features.

In addition, contextual information is equally important semantics for improving the accuracy of semantic segmentation. Zhou et al. [43] used convolutional features to capture spatial contextual information, encode category co-occurrence relationships into convolutional features, and thus decouple the most discriminative features. Zhou et al. [44] proposed a hierarchical context network to simultaneously explore pixel-to-pixel (P2P) and pixel-to-object (P2O) relationships, learn detail-granularity context, and semantic-grained context separately, thereby aggregating to obtain hierarchical context information. These CNN-based methods have made significant contributions to the development of semantic segmentation for RS images.

B. Semantic Segmentation of RS Images Based on Transformer

Transformer is built upon the self-attention mechanism, which demonstrates superior performance in global context modeling. In recent times, the application of transformer models in the realm of computer vision has yielded remarkable outcomes. In RS image semantic segmentation, transformer's exceptional capability in global modeling has garnered attention. Most existing transformer-based RS image segmentation methods employ a hybrid architecture that incorporates both CNN and transformer in an encoder–decoder design. ST-UNet [16] introduced a novel dual-encoder architecture by combining transformer and CNN. It utilized the Swin transformer to facilitate the acquisition of global semantic information by the network and to alleviate the limitations of CNN's global modeling. DC-Swin [36] employed transformers to extract multiscale global contextual feature information and integrated features through the densely connected feature aggregation module (DCFAM). Swin-CNN [30] adopted Swin transformer as the network backbone and integrated CNN's spatial pyramid pooling, channel attention, and edge detection strategies, thereby verifying that CNN's effective strategies are also applicable to transformers. CTMFNet [37] used HRNet and CrossFormer [45] as the backbone paths of CNN and transformer, respectively, to extract local and global contextual information. It further designed the dual-backbone attention fusion module (DAFM) to fuse the local and global contextual information of the two branches. These works achieved remarkable results, corroborating the suitability of transformer structure for RS image semantic segmentation.

C. Multitask Learning

Multitask learning is an effective approach for improving model generalization and robustness. This approach takes advantage of the complementary information between tasks and shares resources and parameters across tasks to improve the efficiency and prediction accuracy of the model. Several recent studies have used different strategies to combine complementary tasks and improve the performance of computer vision tasks. For example, Takikawa et al. [17] proposed a dual-stream network that explicitly injects boundary information into the segmented CNN and uses a dual-task loss function to refine the semantic mask and boundary prediction. Hang et al. [46] proposed an automatic identification model upon CNNs, which comprises two parallel decoders to learn multiscale features and edge information, respectively, effectively improving the accuracy of the model. Li et al. [47] introduced a boundary loss that adaptively fuses with the original semantic loss in RS image semantic segmentation tasks, enabling semantic and boundary information to complement and enhance each other. Bhattacharjee et al. [48] presented an end-to-end multitask transformer architecture that employs shared attention between transformers of multiple tasks to model the dependencies between the tasks. Zhang et al. [49] combined deformable CNN and query-based transformers to propose a novel MTL model for multitask learning of dense prediction.

Drawing inspiration from these excellent works, we have adopted the Swin Transformer backbone as the encoder of our

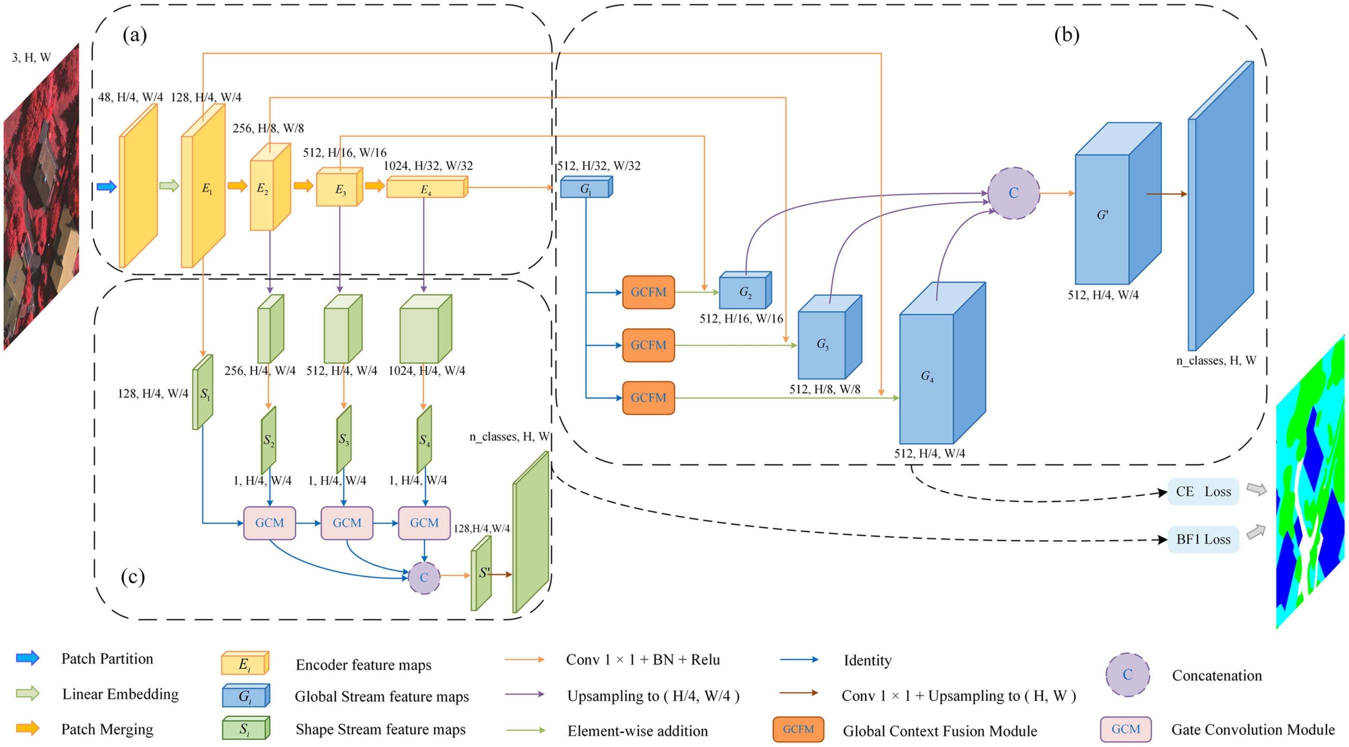

Fig. 1. Architecture of our proposed STDSNet. STDSNet comprises an encoder module that utilizes the Swin transformer backbone to extract features, and a two-branch decoder module consisting of the global stream and the SS. (a) Encoder. (b) Global Stream. (c) SS.

network to transmit the information stream to the dual-stream decoder based on CNNs. This hybrid framework enables the proposed network to harness the respective strengths of both the CNN and transformer architectures in semantic segmentation for RS images.

III. Method

In this section, we provide a comprehensive account of the architecture and component modules of STDSNet. We commence by delineating the overall network structure, which is followed by an introduction of the Swin transformer encoder, the global stream, and the SS.

A. Network Structure

As stated in Section I, transformer has demonstrated superiority over CNNs in encoding global context and shape features that are essential for accurate semantic segmentation of RS images. Our objective is to develop a CNN-based decoder that can effectively incorporate the characteristics of the transformer architecture. Towards this end, we propose the STDSNet, which synergistically combines the strengths of both CNNs and transformers.

The overall architecture of the proposed STDSNet, as depicted in Fig. 1, follows a single encoder and dual decoders framework. The robust capacity of the Swin transformer to encode global context and shape features is leveraged in STDSNet to augment the network's performance. In the dual-branch decoder component, we have devised two parallel streams, namely the global stream and the SS, for processing the global context and boundary information, respectively. To this end, the GCFM is introduced in the global stream to recover the loss of global context, while the GCM in the SS filters noise unrelated to boundary information. In the STDSNet, the outputs of the global stream and SS collaborate to produce the final prediction. This is achieved by coupling the respective loss functions of the dual streams.

B. Swin Transformer Encoder

The Swin transformer, which is based on ViT, invokes CNN's hierarchical structure and restricts self-attention to nonoverlapping windows at different layers, thereby significantly reducing processing complexity. It enables window-to-window information interaction through sliding window operations, thus giving the model the ability to model globally in disguise.

1) Swin Transformer Backbone: As shown in Fig. 2, the Swin transformer backbone consists of four stages. Specifically, the patch partition module divides the input RGB image into nonoverlapping patches with a size of 4 × 4, and the number of channels is $4 \times 4 \times 3 = 48$. Stage 1 consists of a linear embedding layer and two successive Swin transformer blocks, which transform the blocks from two-dimensional to one-dimensional sequence features through the linear layer and project the number of channels to $\mathbf{C}$. In stages 2 through 4, the patch merging modules halve the height and width of the feature map and double the number of channels during the downsampling process. The data sequence through the Swin

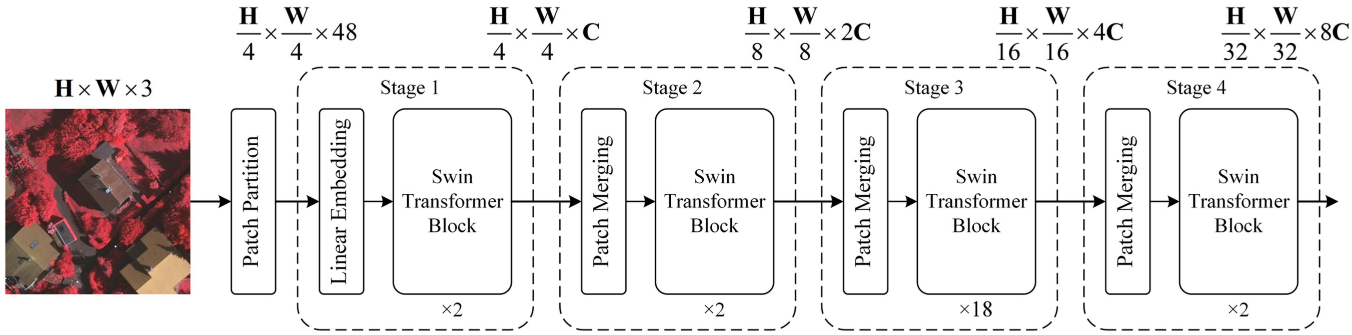

Fig. 2. Structure of the Swin transformer backbone, consisting of four stages.

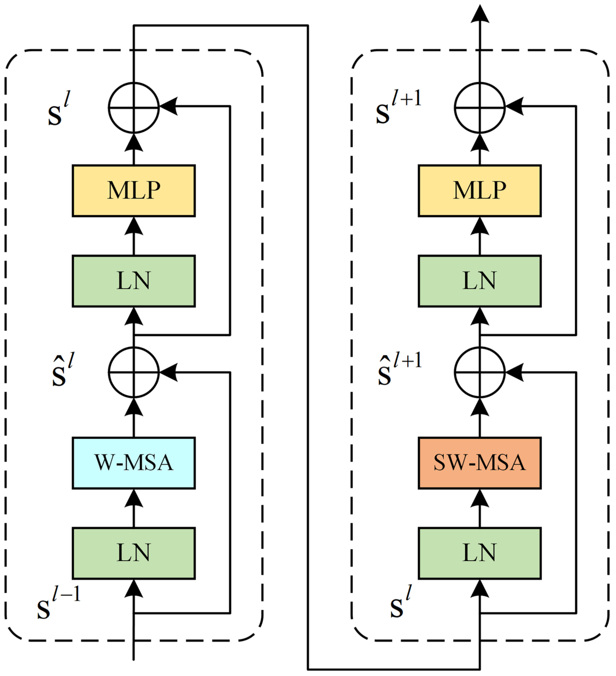

Fig. 3. Two successive Swin transformer blocks: the former is based on W-MSA, and the latter is based on SW-MSA.

transformer block does not alter the size of the image or the number of channels, but it increases the attention weights on its sequence.

In Fig. 2, the output of the ith stage is marked as E_i ($i = 1, 2, 3, 4$), which corresponds to Fig. 1(a). In our STEDNet, we adopt the Swin-base (Swin-B) backbone configuration (i.e., channel $\mathbf{C} = 128$, a window size of 7, channel numbers of $\{128, 256, 512, 1024\}$ per stage, layer numbers of $\{2, 2, 18, 2\}$ per stage, and head numbers of $\{4, 8, 16, 32\}$ per layer).

2) Swin Transformer Block: Swin transformer blocks appear in pairs in the Swin transformer backbone, as depicted in Fig. 3. The former Swin transformer block is based on the multihead self-attention mechanism of the window (W-MSA), whereas the latter is based on the multihead self-attention mechanism of the shifting window (SW-MSA). Each Swin transformer block contains a multilayer perceptron (MLP), a LayerNorm (LN) layer prior to each self-attention mechanism and MLP, and a residual connection afterward. They can be expressed as follows:

$$\begin{aligned}\hat{\mathrm{s}}^{l} &= \text{W-MSA}(\text{LN}(\mathrm{s}^{l-1})) + \mathrm{s}^{l-1}\\ \mathrm{s}^{l} &= \text{MLP}(\text{LN}(\hat{\mathrm{s}}^{l})) + \hat{\mathrm{s}}^{l}\\ \hat{\mathrm{s}}^{l+1} &= \text{SW-MSA}(\text{LN}(\mathrm{s}^{l})) + \mathrm{s}^{l}\\ \mathrm{s}^{l+1} &= \text{MLP}(\text{LN}(\hat{\mathrm{s}}^{l+1})) + \hat{\mathrm{s}}^{l+1}\end{aligned} \tag{1}$$

where $\hat{\mathrm{s}}^{l}$ and $\hat{\mathrm{s}}^{l+1}$ represent the outputs of the W-MSA and SW-MSA modules, respectively, and s^{l} and s^{l+1} represent the outputs of the MLP in the former and latter Swin transformer blocks, respectively.

C. Global Stream

Global context encoding is critical for large-scale region categorization in RS images [16], [50]. The absence of global context significantly degrades the precision of semantic segmentation, which becomes particularly pronounced when dealing with high-resolution RS images [30], [37]. The Swin transformer's robust global modeling abilities allow for the inclusion of higher-level global contextual information from the encoder's deeper layers.

However, popular decoders typically employ conventional upsampling techniques such as linear interpolation, which result in the loss of global context during upsampling. Consequently, these decoders fail to fully utilize the global context obtained from the Swin transformer. To address this limitation, we designed the global stream, which uses the GCFM to compensate for the loss of global context during upsampling. Moreover, we establish three skip connections to interconnect feature maps of corresponding sizes between the encoder and decoder, followed by fusing feature maps of different scales, thereby optimizing the utilization of global context.

1) Structure of the Global Stream: The global stream architecture is illustrated in Fig. 1(b). Initially, the channel dimension of the output tensor $E_4 \in \mathbb{R}^{1024\times(\mathrm{H}/32)\times(\mathrm{W}/32)}$ from the final stage of the encoder is converted to 512 by applying a 1×1 convolution operation. The resulting output is denoted as $G_1 \in \mathbb{R}^{512\times(\mathrm{H}/32)\times(\mathrm{W}/32)}$. This conversion process is performed to prepare for subsequent processing in the GCFM. Similarly, the channel dimensions of the output tensors (E_1–E_3) from stages 1–3 of the encoder are also converted to 512 by a 1×1

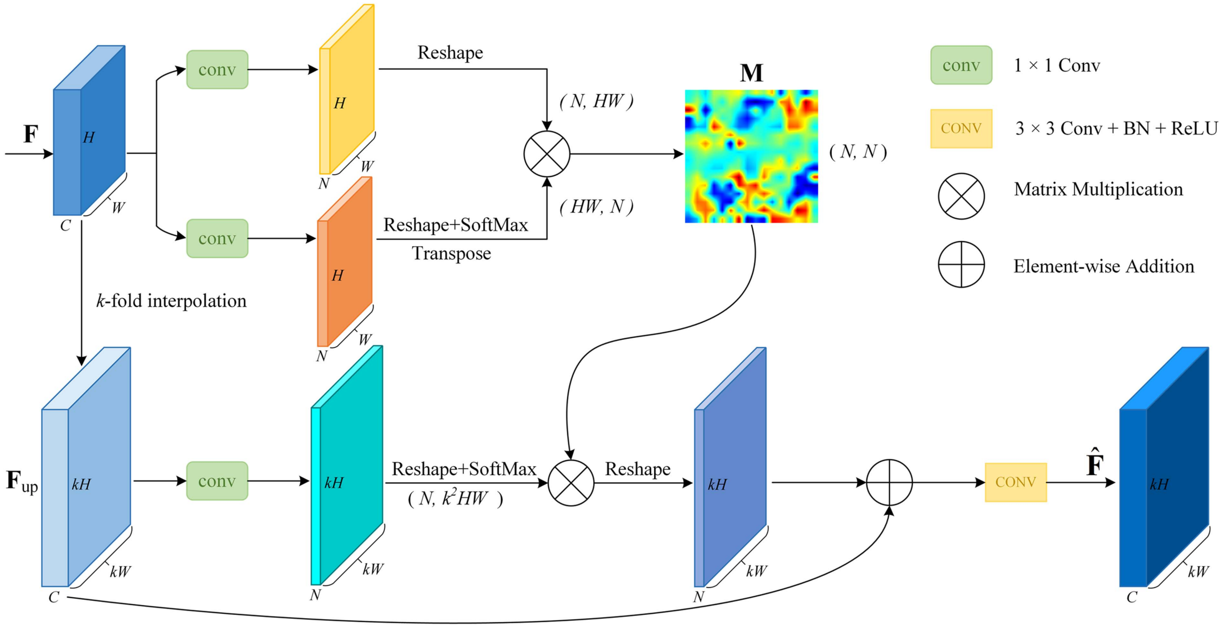

Fig. 4. Detailed design of the GCFM. It employs a parameter k to regulate the up-sampling multiplier and integrates the correlation matrix to recover the loss of global context during the upsampling process.

convolution operation to prepare them for feature fusion. Three GCFMs are employed in the global stream. The size of each feature map inputted to the GCFM is upsampled to k times its original size by tuning the parameter k. Subsequently, G_1 is inputted into the first GCFM (k = 2), and the resulting output is fused with the feature map output from E_3 via element-wise summation, yielding $G_2 \in \mathbb{R}^{512\times(\mathrm{H}/16)\times(\mathrm{W}/16)}$. Then, G_1 is inputted into the second GCFM (k = 4), and the output is fused with the feature map output from E_2, producing $G_3 \in \mathbb{R}^{512\times(\mathrm{H}/8)\times(\mathrm{W}/8)}$. Finally, G_1 is inputted into the third GCFM (k = 8) and further fused with the feature map output from E_1, and the resulting tensor is denoted as $G_4 \in \mathbb{R}^{512\times(\mathrm{H}/4)\times(\mathrm{W}/4)}$.

After acquiring feature maps G_i $(i = 2, 3, 4)$, we upsample G_2 and G_3 to the same resolution as G_4. Next, the feature maps G_2–G_4 are concatenated along the channel dimension for further feature fusion. Subsequently, a 1 × 1 convolution is applied for dimensionality reduction, and the resulting output is denoted as $G' \in \mathbb{R}^{512\times(\mathrm{H}/4)\times(\mathrm{W}/4)}$. Finally, a 1 × 1 convolution layer is used, followed by linear interpolation upsampling on the feature map G' to generate a prediction mask for the global stream.

2) GCFM: The GCFM acquires the correlation matrix by integrating visual primitives extracted from high-level feature maps. The correlation matrix captures the dense spatial mapping relationships between all feature pixels, preserving the long-range context dependencies of spatial features and supplying denser global contextual information. Subsequently, the correlation matrix is fused with the upsampled decoder feature map, enabling the recovery of global context during the upsampling process.

As depicted in Fig. 4, the GCFM accepts the output of the final layer of the encoder backbone network, denoted as $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, as its input. Initially, two 1 × 1 convolutional layers are applied to the feature map $\mathbf{F}$ to decrease its channel count, resulting in the acquisition of two feature maps, each containing N channels. Then, the visual primitives are acquired adaptively from these two feature maps and fused to produce the correlation matrix, denoted as $\mathbf{M} \in \mathbb{R}^{N\times N}$. The correlation matrix encodes a comprehensive depiction of the relationship among distinct feature space positions and can be computed as follows:

$$\mathbf{M} = \delta(\mathbf{F}) \otimes \mathrm{SoftMax}^{\top}(\mu(\mathbf{F})). \tag{2}$$

where δ and μ stand for two 1 × 1 distinct convolutional layers, while $\otimes$ signifies the matrix multiplication operation.

To address the loss of global semantics caused by upsampling, the correlation matrix $\mathbf{M}$ is utilized during the bilinear interpolation upsampling of the global stream. To begin with, we obtain $\mathbf{F}_{\mathrm{up}} \in \mathbb{R}^{C\times kH\times kW}$ by bilinear interpolation upsampling the feature map $\mathbf{F}$ with a factor of k $(k = 2, 4, 8)$. Following this, we convert the feature map $\mathbf{F}_{\mathrm{up}}$ into an N-dimensional attention vector using a 1 × 1 convolutional layer and a softmax function. Subsequently, the correlation matrix is fused with the N-dimensional attention vector, and the resulting output establishes a skip connection with $\mathbf{F}_{\mathrm{up}}$. Finally, we obtain the result $\hat{\mathbf{F}} \in \mathbb{R}^{C\times kH\times kW}$ through a 3 × 3 convolution, BN, and ReLU layer, which is calculated as

$$\hat{\mathbf{F}} = \xi(\mathbf{M} \otimes \mathrm{SoftMax}(\eta(\mathbf{F}_{\mathrm{up}})) \oplus \mathbf{F}_{\mathrm{up}}) \tag{3}$$

where ξ denotes the 3 × 3 convolutional layer together with the BN and ReLU layers, while η represents the 1 × 1 convolutional layer. In addition, we use $\oplus$ to indicate the operation of element-wise addition.

D. Shape Stream

Both the encoders employed in the transformer and CNN architectures tend to capture low-level features at shallow layers [33]. However, the transformer architecture has superior encoding of shape features and a stronger shape bias compared to CNNs [31], [32], which provides a novel approach for extracting boundary information.

However, the present state-of-the-art methods utilized for image semantic segmentation involve dense image representation, wherein CNNs are employed to simultaneously process shape, color, and texture information [28], [51], [52]. These methods may not be optimal as they comprise disparate types of information that are crucial for recognition. On the other hand, the majority of current multitask learning methods rely on CNN-based encoders [17], [47], [53], such as ResNet [11] and ResNeSt [54], which are unable to overcome the disadvantages of CNNs in encoding shape features during feature extraction. As a result, these methods do not transfer well to encoder architectures that use transformers. To overcome these limitations, we propose a novel single encoder and dual decoders architecture that maximizes the shape features captured by transformers in order to extract boundary information more effectively. This is achieved by incorporating shape features into a separate processing branch (called the SS) to acquire boundary information. Our dual-stream network architecture enables the parallel processing of information in both the global stream and the SS.

1) Structure of the SS: The diagram in Fig. 1(c) depicts the structure of the SS. Initially, a 1 × 1 convolution is applied to the feature map E_1, which maintains its size and number of channels, resulting in the feature map $S_1 \in \mathbb{R}^{128\times(\mathrm{H}/4)\times(\mathrm{W}/4)}$. The feature maps E_i ($i = 2, 3, 4$) are upsampled to match the size of S_1, and a channel reduction to 1 is performed through a 1 × 1 convolution, respectively, producing the feature maps $S_i \in \mathbb{R}^{1\times(\mathrm{H}/4)\times(\mathrm{W}/4)}$ ($i = 2, 3, 4$). After the completion of these operations, the preparation of the GCM is finalized.

We employ three GCMs to selectively gate the boundary-related information within the SS. Specifically, we integrate three GCMs at distinct stages of the encoder to enable selective suppression of noise layer by layer. As shown in Fig. 1(c), the feature maps S_1 and S_2 are simultaneously inputted into the first GCM. The output from this step is then fed into the second GCM along with a feature map S_3 to obtain the second GCM's output. Furthermore, the output of the second GCM undergoes further processing in a similar manner with feature map S_4 to produce the output of the third GCM. Notably, the output feature map size and channels of all three GCMs remain consistent with S_1. Following this, we perform channel-wise concatenation of the outputs from the three GCMs and apply a 1 × 1 convolution layer to adjust their channel dimensions, resulting in the feature map $S' \in \mathbb{R}^{128\times(\mathrm{H}/4)\times(\mathrm{W}/4)}$. Finally, we apply a 1 × 1 convolution layer and linear interpolation upsampling to S' to generate the ultimate prediction mask for the SS.

2) GCM: In the shallow layers of the encoder, a multitude of low-level features are captured, causing ambiguity between the object and background boundaries. These superfluous low-level features are perceived as noise during context fusion and can negatively impact the precision of object boundary segmentation. To address this, we propose the GCM, which leverages high-level semantics within the SS to gate low-level semantics, where high-level semantics possess a higher-level semantic comprehension of the scene and guide low-level semantics in selectively focusing on the pertinent boundary-related information. In this way, extraneous noise can be effectively eliminated.

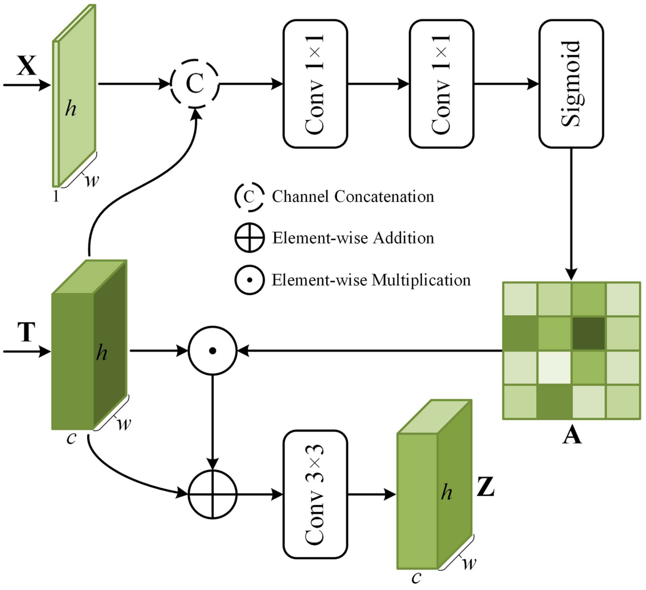

Fig. 5. Detailed design of the GCM. It utilizes high-level semantics to gate low-level semantics, thereby eliminating extraneous feature information.

As shown in Fig. 5, the GCM's operation commences with the concatenation of the high-level semantic $\mathbf{X} \in \mathbb{R}^{c\times h\times w}$ and the low-level semantic $\mathbf{T} \in \mathbb{R}^{1\times h\times w}$, which are then subjected to a pair of successive 1 × 1 convolutional layers to obtain the attention map $\mathbf{A} \in \mathbb{R}^{c\times h\times w}$. The computational process of the attention map can be represented as follows:

$$\mathbf{A} = \sigma(\phi(\lambda(\mathbf{X} \cup \mathbf{T}))) \tag{4}$$

where $\cup$ denotes the channel concatenation operation. λ and ϕ represent a convolutional layer and a normalized convolutional layer, respectively, where both operations employ a convolution kernel size of 1 × 1, followed by a sigmoid function σ. Then, GCM involves the element-wise multiplication of the low-level semantic $\mathbf{T}$ with an attention map $\mathbf{A}$, which is subsequently followed by the inclusion of a residual connection and a convolutional layer. This process can be expressed by the following equation:

$$\mathbf{Z} = \psi(\mathbf{T} \odot \mathbf{A} \oplus \mathbf{T}) \tag{5}$$

where ψ denotes a 3 × 3 convolution operation and $\odot$ signifies the operation of element-wise multiplication. $\mathbf{Z} \in \mathbb{R}^{c\times h\times w}$ represents the output of the GCM, which is subsequently transmitted to the subsequent layer of the SS for further processing. More specifically, as shown in Fig. 1(c), the low-level semantic of the previous GCM output will be employed as input to the following GCM, together with the high-level semantic from deeper features. This multistage approach facilitates the gating of the low-level semantic through the high-level semantic, thereby effectively filtering out redundant low-level features.

E. Loss Function

In our experiment, we employ the cross-entropy loss as the loss function for the global stream. The global stream loss (L_{GS}) is denoted as

$$L_{\text{GS}} = -\frac{1}{N}\sum_{j}^{N}\sum_{i}^{C}[\, y_{ji}\log(\hat{y}_{ji}) + (1 - y_{ji})\log(1 - \hat{y}_{ji})] \quad (6)$$

where N and C refer to the number of samples and categories, respectively. In addition, the symbols y and $\hat{y}$ signify the ground truth of the segmentation and the corresponding predictions, respectively.

Furthermore, to enhance the network's ability to extract boundary information, we utilize the boundary loss (BF1) [55] as the SS's loss function, which has demonstrated its effectiveness in RS image semantic segmentation tasks [30], [53], [56]. The SS loss (L_{SS}) consists of the following calculation steps:

$$y_{gt}^{b} = \text{pool}\,(1 - y_{gt}\,, \theta_0) - (1 - y_{gt})$$

$$y_{pd}^{b} = \text{pool}\,(1 - y_{pd}, \theta_0) - (1 - y_{pd}) \quad (7)$$

$$y_{gt}^{b,\text{ext}} = \text{pool}\,(y_{gt}^{b}, \theta),\ y_{pd}^{b,\text{ext}} = \text{pool}\,(y_{pd}^{b}, \theta) \quad (8)$$

$$\text{P}^{c} = \frac{\text{sum}\,(y_{pd}^{b} \circ y_{gt}^{b,\text{ext}})}{\text{sum}\,(y_{pd}^{b})},\ \text{R}^{c} = \frac{\text{sum}\,(y_{gt}^{b} \circ y_{pd}^{b,\text{ext}})}{\text{sum}\,(y_{gt}^{b})} \quad (9)$$

$$L_{\text{SS}} = 1 - \frac{1}{C+1}\sum_{c=0}^{C}\frac{2 \times \text{P}^{c} \times \text{R}^{c}}{\text{P}^{c} + \text{R}^{c}} \quad (10)$$

where y_{gt} and y_{pd} represent the binary maps corresponding to the ground truth and predicted labels for an arbitrary class c within an image, respectively. The max-pooling operation is denoted by pool $(\cdot)$, and the sliding window size, i.e., the convolution kernel size, is denoted by θ_0, which is typically set to 3. The extended boundaries of the true map and predicted map are represented by $y_{gt}^{b,\text{ext}}$ and $y_{pd}^{b,\text{ext}}$, respectively, which are obtained through maximum pooling. Here, θ represents the size of the padding operation, and it is set to 5 in this article. We use P^{c} and R^{c} to represent precision and recall, respectively. The sum operation is denoted by sum $(\cdot)$, and pixel-wise multiplication is represented by $\circ$.

By coupling their respective loss functions, the output feature maps from the global stream and the SS culminate in the final prediction of the STDSNet. The total loss (L_{total}) of the model is calculated by aggregating the global stream loss L_{GS} and the SS loss L_{SS} through a weighted sum, which can be expressed as

$$L_{\text{total}} = \alpha L_{\text{GS}} + \beta L_{\text{SS}} \quad (11)$$

where α and β are the weights of the global stream and the SS, respectively. We conduct a systematic analysis by varying the hyperparameter values of α and β to investigate their impact on the overall performance of the model. Subsequently, we perform a meticulous evaluation of the experimental outcomes and ultimately select the values $\alpha = 1$ and $\beta = 0.2$ as optimal. The comprehensive details of our experimental methodology and findings are expounded in Section IV-C.

IV. Experiments and Results

A. Datasets

The effectiveness of the proposed network has been evaluated through experiments conducted on the International Society of Photogrammetry and Remote Sensing (ISPRS) Vaihingen and Potsdam datasets.

1) Vaihingen Dataset: Vaihingen is a small German village with numerous adjacent multistory buildings in Germany. The Vaihingen dataset comprises true orthophoto (TOP) tiles, each containing three spectral bands of red, green, and near-infrared, along with the digital surface model (DSM) and the normalized DSM (NDSM). Note that we have not used DSM and NDSM in our experiments to reduce computation. Based on studies [14] and [58], we have used 16 TOP images for training, while the remaining 17 have been used for testing. The average size of the images is 2064 × 2494 pixels, and the ground sampling distance (GSD) is 9 cm. The dataset is tagged with six categories to facilitate study on semantic segmentation, namely impervious surface (white), building (blue), low vegetation (cyan), tree (green), car (yellow), and clutter (red). Notably, we have ignored the category "clutter" during the quantitative evaluation of the dataset, in line with previous studies [57] and [58].

2) Potsdam Dataset: Potsdam is a quintessential German city with a compact urban structure and an intricate network of buildings. The Potsdam dataset comprises 38 TOP tiles of identical dimensions (6000 × 6000 pixels), with a GSD of 5 cm. Following studies [14] and [58], 24 images are utilized for training and the remaining 14 for testing. Each TOP image in the Potsdam dataset comprises near-infrared, red, green, and blue channels, as well as DSM and NDSM. Similar to the Vaihingen dataset, the Potsdam dataset assigns six categories for semantic segmentation of the images, with the category "clutter" being excluded during quantitative evaluation.

B. Evaluation Metrics

Mean intersection over union (mIoU) and mean F1 score (mF1) are employed to evaluate the results of semantic segmentation in our experiments. The mIoU is calculated as the mean of the IoU for all categories, whereas the mF1 is calculated as the mean of the F1 score for all categories. IoU is defined as the ratio of the intersection and union of the true value and the projected value for each category, whereas the F1 score is a thorough evaluation of precision and recall for each category. The calculation formulas for these two evaluation metrics, which are based on the confusion matrix, are given as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (12)$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (13)$$

TABLE I
ABLATION EXPERIMENT OF THE PROPOSED MODULES ON THE VAIHINGEN AND POTSDAM DATASETS

Dataset	Method	mIoU(%)	mF1(%)
Vaihingen	STDSNet-GCFM	81.09	89.39
	STDSNet-GCM	81.45	89.61
	STDSNet	81.77	89.81
Potsdam	STDSNet-GCFM	81.46	89.63
	STDSNet-GCM	81.81	89.86
	STDSNet	82.33	90.17

where precision and recall are calculated as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

where true positive (TP) is the number of pixels where both the predicted and actual conditions are true. False positive (FP) indicates the number of pixels for which the predicted condition is true but the actual condition is false. False negative (FN) represents the number of pixels for which the predicted condition is false but the actual condition is true.

C. Implementation Details

Both our method and other state-of-the-art methods are trained and tested using a single NVIDIA Tesla P100 with 16 GB of memory utilizing the PyTorch framework. To ensure fairness in the experiments, all participating methods execute training and testing in the same experimental environment. Specifically, the maximum number of epochs is set to 100, and the batch size is set to 4. All of our transformer-based models are trained using adaptive moment estimation (AdamW) [59], with the initial learning rate set to 6×10^{-5} and a weight decay of 0.01. We enable the warm-up strategy in the initial 1500 iterations to slow down the initial overfitting of the model to the minibatch and maintain a smooth distribution. Each raw image in the training and testing sets is divided into 512 × 512 pixel patches. During the data preprocessing stage, various standard data augmentation techniques are implemented, including resizing, random cropping, random flipping, photometric distortion, and normalization. In our experiment, all methods are initiated with pretrained weights to expedite the convergence of the model.

D. Ablation Study

In the proposed STDSNet, the GCFM and the GCM emerge as two pivotal components. In order to evaluate the specific contribution of these two modules to STDSNet, we performed ablation experiments on the ISPRS-Vaihingen and Potsdam datasets. Table I shows the results of the experiments with and without the integration of the two components. For the sake of simplicity, we refer to STDSNet-GCFM and STDSNet-GCM, which denote the removal of the GCFM and GCM from STDSNet, respectively.

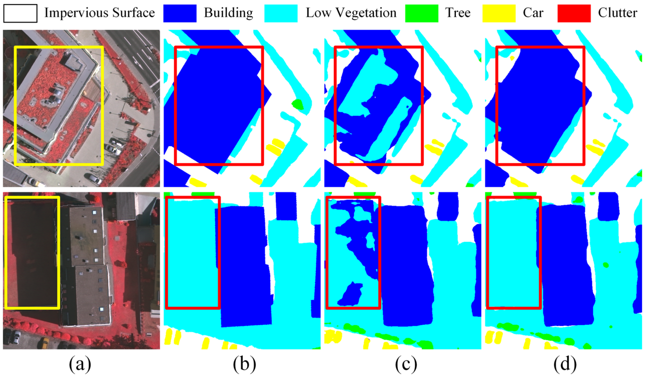

Fig. 6. Comparison of segmentation results without and with using GCFM in the STDSNet. (a) image. (b) ground truth. (c) STDSNet-GCFM (without GCFM). (d) STDSNet (with GCFM).

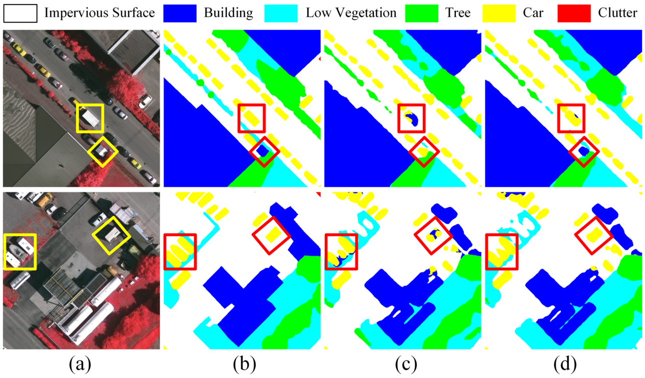

Fig. 7. Comparison of segmentation results without and with using GCM in the STDSNet. (a) image. (b) ground truth. (c) STDSNet-GCM (without GCM). (d) STDSNet (with GCM).

1) Effect of GCFM: As presented in Table I, STDSNet-GCFM reduces mIoU from 81.77% to 81.09% and mF1 from 89.81% to 89.39% on the Vaihingen dataset, which is a drop of 0.68% and 0.42%, respectively, compared to STDSNet. Similar trends can be observed on the Potsdam dataset, where the removal of GCFM in STDSNet lowers mIoU and mF1 by 0.87% and 0.54%, respectively. Fig. 6 illustrates the visualized comparison results. In the first row, the top of the "Building," which is extremely similar to the "Low Vegetation," can be effectively distinguished from them after using GCFM. In the second row, segmentation errors caused by the large shadow occlusion are prevented after using GCFM. This demonstrates that the loss of global context during upsampling is reduced after using GCFM to mitigate large-scale regional object classification errors resulting from similar features or shadow occlusion.

2) Effect of GCM: In comparison to STDSNet, STDSNet-GCM lowers mIoU from 81.77% to 81.45% and mF1 from 89.81% to 89.61% on the Vaihingen dataset, a decrease of 0.32% and 0.2%, respectively. Similarly, the removal of GCM in STDSNet lowers mIoU and mF1 by 0.52% and 0.31% on the Potsdam dataset, respectively. A more visual comparison of the segmentation results is shown in Fig. 7. In the first row,

TABLE II
Ablation Experiments of the Proposed Modules on Vaihingen Dataset

Model Name	Decoder Modules			Iou(%)					Evaluation Index	
	All-MLP	GS	SS	Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)	mF1(%)
Swin-B+All-MLP	✓			85.97	91.20	70.73	80.01	73.56	80.29	88.85
Swin-B+GS		✓		86.10	92.12	72.15	80.67	75.52	81.31	89.53
Swin-B+GS+SS		✓	✓	86.19	92.23	72.67	81.17	76.57	81.77	89.81

the small target “Car” and “Building” can be segmented more accurately after using GCM. In the second row, the use of GCM improves the segmentation of small target “Car,” particularly the boundaries of small targets, which can be segmented more meticulously. This shows that using GCM improves the performance of semantic segmentation of small targets and their boundaries by filtering out irrelevant features and allowing it to concentrate on processing boundary information.

3) Ablation Study for the Dual-Decoder Structure: To validate the effectiveness of the dual decoder structure proposed in STDSNet, we perform a decomposition of STDSNet to evaluate the impact of each stream on semantic segmentation tasks for RS images. Specifically, we investigate the contributions of the global and SSs in the dual decoder to enhance the network’s performance. To establish a baseline for the network, we adopt the ALL–MLP decoder [28], which has been proven to achieve satisfactory segmentation outcomes using the Swin transformer in RS images [60]. Through experimentation on the Vaihingen dataset, we analyze the effects of integrating various components.

The experimental results are presented in Table II. The quantitative results in the table indicate that our dual-stream network yields a 1.48% significant increase in mIoU and a 0.96% increase in mF1 compared to the baseline. Specifically, when employing the global stream (GS) as the network’s decoder instead of ALL-MLP, both mIoU and mF1 exhibit improvements of 1.02% and 0.68%, respectively. This demonstrates that the GS can effectively address the segmentation challenges posed by similar feature objects and the objects obstructed by shadows in RS imagery, thereby significantly enhancing the quality of segmentation results obtained for large-scale regions in RS images. Furthermore, the introduction of SS to the GS further improves mIoU by 0.46% and mF1 by 0.28%, with the most substantial improvements observed in the small target “Car” category, enhancing IoU by 1.05%. These results indicate that the SS can effectively improve the segmentation results of small targets and their boundaries.

To visually assess the efficacy of the dual-stream architecture, we present visual examples of semantic segmentation results for the Vaihingen dataset in Fig. 8. The first row of the figure displays notable examples from the entire scene, which have been labeled as colored squares. Rows 2 to 4 provide zoomed-in versions of these examples to showcase greater detail. It is evident from the second and third rows of Fig. 8(d) that the use of the GS as the decoder can effectively ameliorate large-scale regional object classification errors that arise due to similar features or shadow occlusion in RS images. Furthermore, incorporating the SS onto the GS, as demonstrated in the fourth row of Fig. 8(e), further improves the segmentation accuracy of small targets such as “Car” and their contours.

4) Ablation Study for Hyperparameter Setting: The hyperparameter setting in (11) is not robust to the segmentation performance of the network, as the multitask loss is influenced by various factors such as correlation, sample distribution, and task complexity. To determine the optimal weight ratio between GS loss and SS loss in the total loss, we conduct ablation experiments on hyperparameters. Specifically, we fix the value of α at 1 and vary the value of β. Fig. 9 demonstrates the efficacy of STDSNet for semantic segmentation of the ISPRS Vaihingen and Potsdam benchmarks with different hyperparameter β thresholds.

Upon observing Fig. 9, it can be noted that both curves exhibit similar trends. Initially, the STDSNet’s segmentation accuracy demonstrates a gradual improvement on both datasets with an increasing value of β. Notably, both curves attain their highest mF1 at $\beta = 0.2$. However, as β continues to increase, a decreasing trend in the network’s segmentation accuracy is observed. This is due to the different scale losses associated with the global and SSs. When the weight loss ratios of the two streams are close, the network tends to excessively focus on the SS, neglecting the dominant role of the GS, ultimately leading to a decline in overall segmentation accuracy. Consequently, α and β are ultimately set to 1 and 0.2, respectively.

E. Comparison With Existing State-of-the-Art Methods

We conduct a comparative analysis of the proposed STDSNet with other state-of-the-art methods, including FCN [10], DeepLabV3+ [18], DANet [21], SegNeXt [51], MAResU-Net [61], SegFormer [28], UPerNet [20], ST-UNet [16], and Swin-CNN [30]. The first five comparative methods are CNN-based, while the remaining four employ a hybrid architecture of CNN and transformer. To ensure experimental fairness, we utilize backbone versions that exhibit comparable parameter counts and computational operations for comparative analysis. Specifically, ResNet101 (R-101) serves as the backbone for FCN, DeepLabV3+, DANet, and MAResU-Net, while MSCAN-L is used for SegNeXt, MIT-B5 for SegFormer, and Swin-B for UPerNet and Swin-CNN. In particular, ST-UNet uses the R-101 and Swin-B dual branches as the backbone of the network.

1) Results on Vaihingen Dataset: We present the quantitative results of our proposed method and other state-of-the-art methods on the ISPRS Vaihingen dataset in Table III. The proposed STDSNet achieves the highest mIoU and mF1 of 81.77% and 89.81%, respectively, and outperforms all other methods in terms of IoU for all categories. Compared to the suboptimal

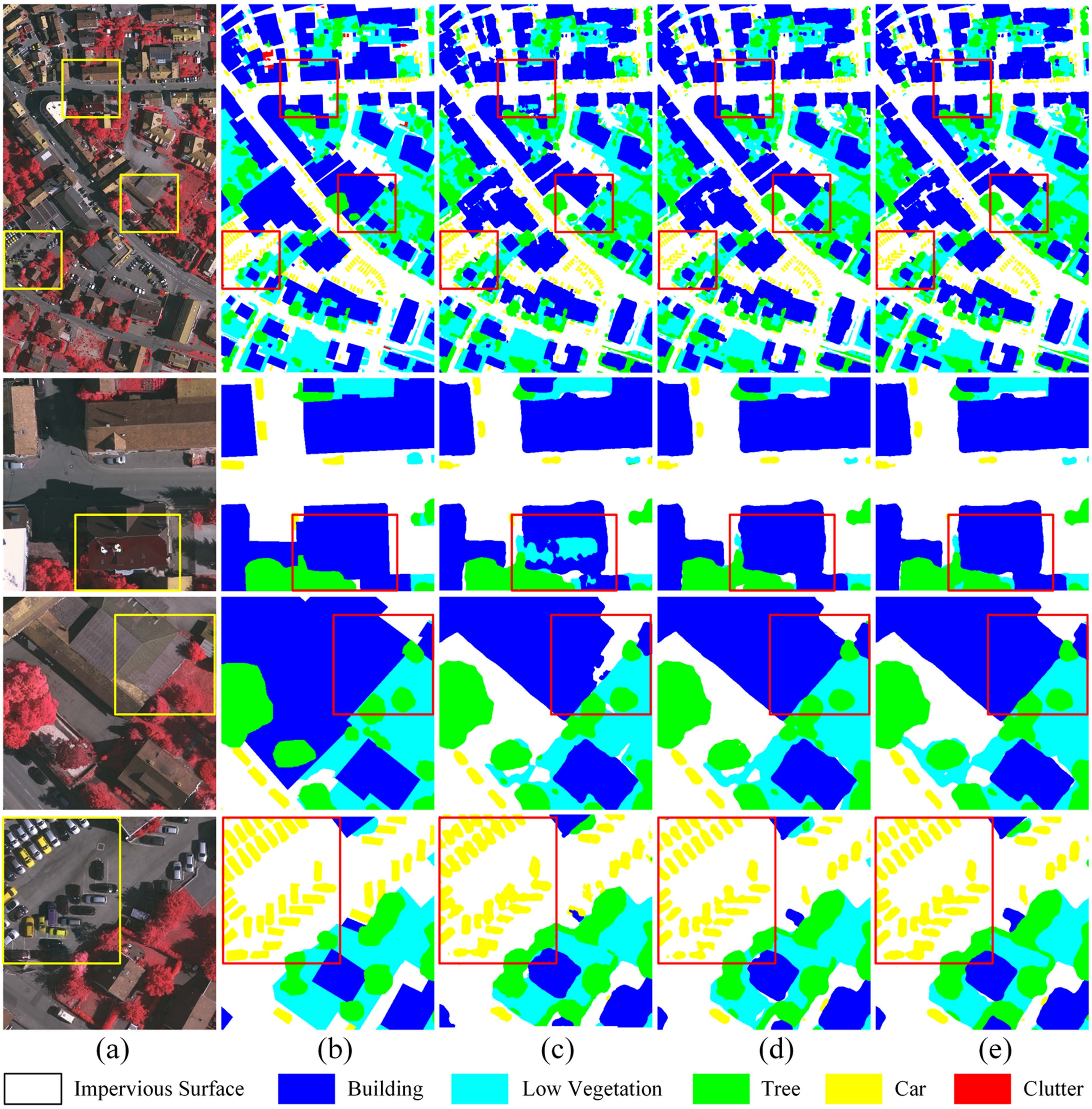

Fig. 8. Comparison of segmentation results using different components in the single encoder and dual decoders framework. (a) image. (b) ground truth. (c) Swin-B + All-MLP. (d) Swin-B + global stream. (e) Swin-B + global stream + shape stream.

method, STDSNet improved by 0.66% in mIoU and 0.4% in mF1. Particularly for the challenging "Car" category, which is a relatively small target object in RS images, our method achieved an IoU of 76.57%, surpassing the suboptimal method by 1.39%.

To visually compare the proposed method with several semantic segmentation methods in Table III, we have visualized the prediction results of each method in Fig. 10. For ground objects exhibiting high similarity or occlusion due to shadows, the STDSNet demonstrates superior capability in making accurate class predictions for large-scale regions. In the first to third rows of the figure, it is evident that certain methods exhibit poor segmentation performance when applied to ground objects with high similarity, as well as those obscured by shadows, due to a severe semantic confusion issue. This can be attributed to these methods' insufficient or inappropriate utilization of finer-grained contextual information obtained at the global level, ultimately leading to object misclassification in large-scale regions within RS images. In comparison, STDSNet still exhibits relatively accurate predictions, which can be attributed to its GS's capacity to optimize the utilization of the global context acquired by the transformer. Furthermore, as demonstrated in the fourth to sixth rows of the figure, STDSNet significantly improves the segmentation of small targets and their edges compared to other methods when faced with small and dense targets such as "Car." This is consistent with our expectation and indicates that the SS is capable of filtering the redundant noise and making reasonable use of the shape features captured by the transformer.

TABLE III
COMPARISON OF SEGMENTATION RESULTS ON THE VIHINGEN DATASET WITH STATE-OF-THE-ART METHODS

Method	Backbone	Iou(%)					Evaluation Index	
		Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)	mF1(%)
FCN	R-101	85.16	90.76	70.92	80.26	70.28	79.48	88.34
DeepLabV3+	R-101	85.01	91.15	70.29	79.73	71.96	79.63	88.45
DANet	R-101	84.11	90.99	70.76	80.11	73.14	79.82	88.59
SegFormer	MIT-B5	85.51	91.71	71.68	80.83	74.47	80.84	89.25
UPerNet	Swin-B	85.86	91.51	72.46	80.84	73.40	80.81	89.23
SegNeXt	MSCAN-L	85.19	91.39	72.69	81.03	74.32	80.92	89.30
MAResU-Net	R-101	85.04	90.87	70.54	80.43	73.47	80.07	88.71
ST-UNet	R-101+Swin-B	85.24	91.56	72.33	80.92	74.38	80.89	89.29
Swin-CNN	Swin-B	85.45	91.55	72.64	80.71	75.18	81.11	89.41
STDSNet	Swin-B	**86.09**	**92.18**	**72.82**	**81.17**	**76.57**	**81.77**	**89.81**

The best results are in bold.

TABLE IV
COMPARISON OF SEGMENTATION RESULTS ON THE POTSDAM DATASET WITH STATE-OF-THE-ART METHODS

Method	Backbone	Iou(%)					Evaluation Index	
		Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)	mF1(%)
FCN	R-101	82.56	89.09	71.78	74.53	81.34	79.86	88.67
DeepLabV3+	R-101	82.69	90.37	71.54	73.93	82.25	80.16	88.83
DANet	R-101	83.40	90.96	72.51	74.48	82.72	80.81	89.24
SegFormer	MIT-B5	84.14	91.69	73.92	75.04	83.11	81.58	89.71
UPerNet	Swin-B	84.12	91.96	73.58	75.18	83.27	81.62	89.73
SegNeXt	MSCAN-L	84.03	91.78	73.94	75.54	83.24	81.71	89.79
MAResU-Net	R-101	83.49	91.02	72.66	74.17	83.21	80.91	89.29
ST-UNet	R-101+Swin-B	84.09	91.73	73.82	75.11	83.45	81.64	89.74
Swin-CNN	Swin-B	84.15	91.89	**74.21**	75.40	83.32	81.79	89.86
STDSNet	Swin-B	**84.90**	**92.23**	74.09	**76.22**	**84.21**	**82.33**	**90.17**

The best results are in bold.

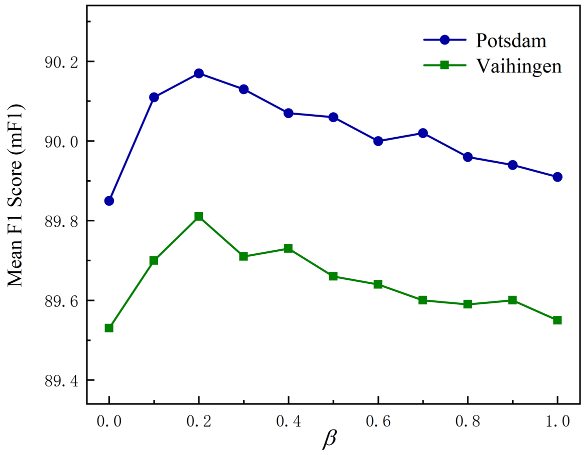

Fig. 9. Curve of effectiveness on STDSNet at different β thresholds.

2) Results on Potsdam Dataset: We present the quantitative results of each method on the ISPRS Potsdam dataset in Table IV to provide further evidence of the effectiveness of STDSNet. The results in the table indicate that our method surpasses the other methods in terms of mIoU and mF1, achieving 82.33% and 90.17%, respectively, while also achieving the highest IoU in most categories. Compared to the second-best method, STDSNet achieves an improvement of 0.54% in mIoU and 0.31% in mF1. Notably, for the "Car" and "Tree" categories, which are relatively small in the Potsdam dataset, STDSNet also outperforms the suboptimal method by 0.89% and 0.82% in IoU, respectively.

Similarly, we visualize the qualitative comparison between STDSNet and alternative methods in Fig. 11 in order to further demonstrate the efficacy of our STDSNet. From the first to fourth rows of the figure, we observe that STDSNet outperforms other methods in accurately identifying ground object classes in large-scale areas, particularly in achieving accurate segmentation of objects with similar appearances and those occluded by shadows. Furthermore, from the fifth to the sixth row, it is evident that STDSNet effectively addresses the problem of missing and incorrect detection of small targets and generates more precise outlines of such targets. These outcomes align with our initial expectations and provide strong validation of the contribution

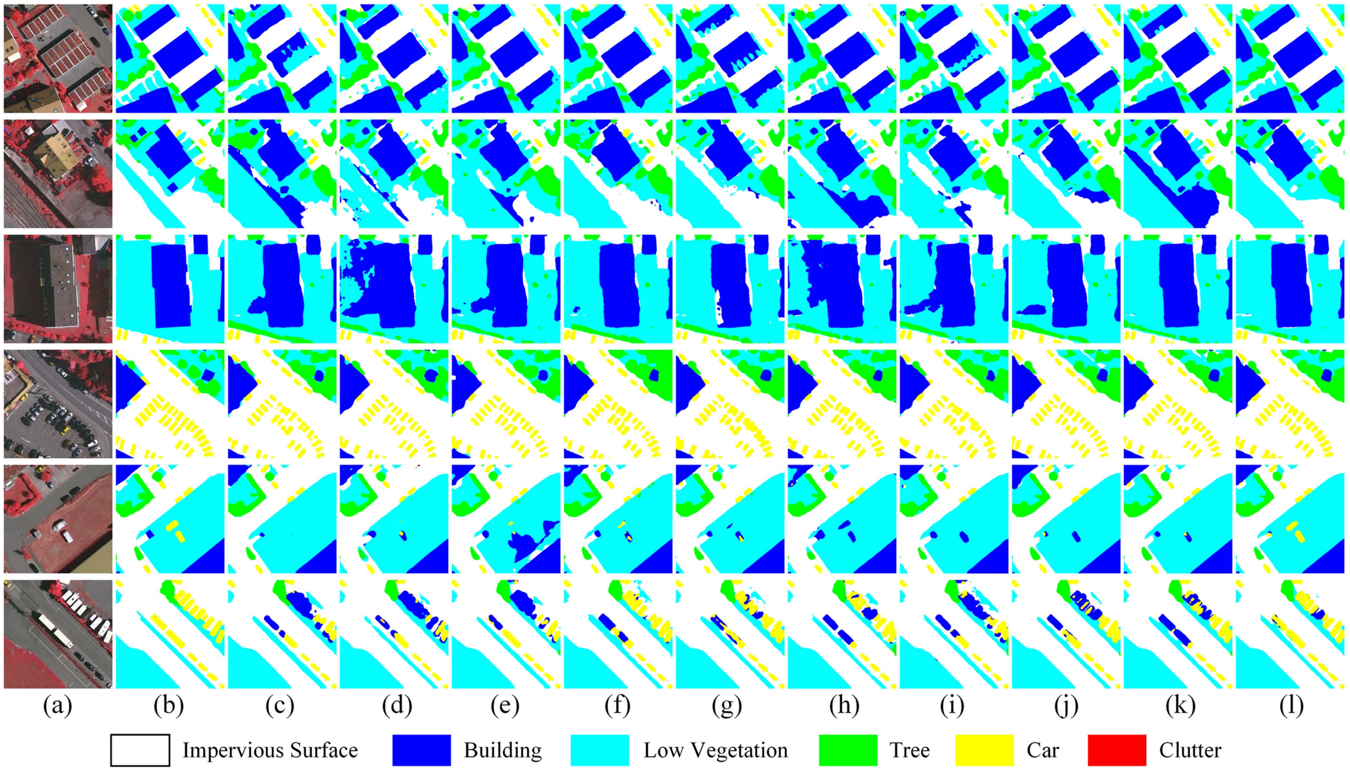

Fig. 10. Comparison between the proposed STDSNet and other state-of-the-art methods on the Vaihingen dataset. (a) image. (b) ground truth. (c) FCN. (d) DeepLabV3+. (e) DANet. (f) SegFormer. (g) UPerNet. (h) SegNeXt. (i) MAResU-Net. (j) ST-UNet. (k) Swin-CNN. (l) STDSNet.

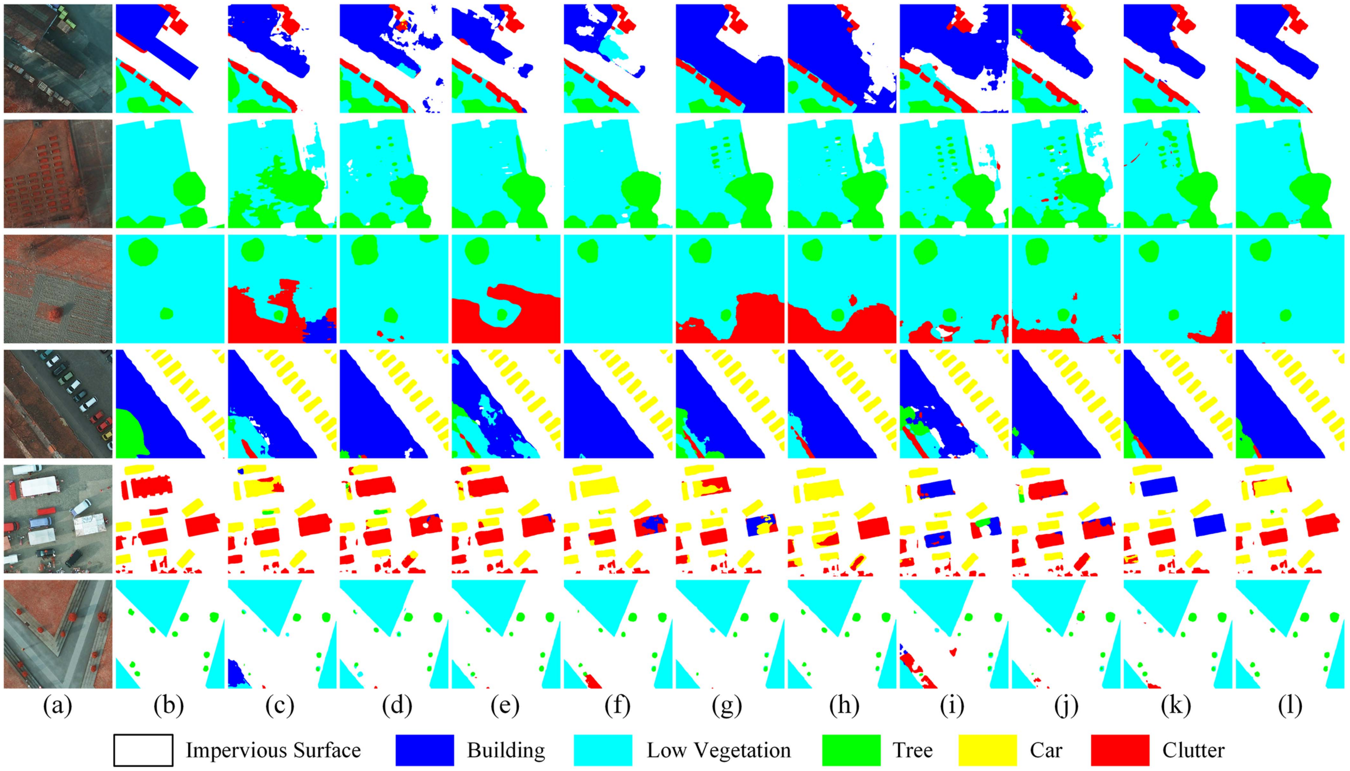

Fig. 11. Comparison between the proposed STDSNet and other state-of-the-art methods on the Potsdam dataset. (a) image. (b) ground truth. (c) FCN. (d) DeepLabV3+. (e) DANet. (f) SegFormer. (g) UPerNet. (h) SegNeXt. (i) MAResU-Net. (j) ST-UNet. (k) Swin-CNN. (l) STDSNet.

TABLE V
COMPARISON OF MODEL PARAMETERS AND FLOPS

Method	Parameters (M)	FLOPS (G)
FCN	68.48	275.88
DeepLabV3+	60.21	254.55
DANet	66.45	277.25
SegFormer	81.97	51.86
UPerNet	119.99	296.10
SegNeXt	48.78	64.65
MAResU-Net	139.83	317.98
ST-UNet	208.45	479.11
Swin-CNN	235.77	688.23
STDSNet	130.13	327.66

of GCFM and GCM to the proposed dual-stream network. Furthermore, they also demonstrate that the dual-stream network can reasonably leverage the global context captured by the transformer as well as the shape features, thereby enhancing the overall segmentation performance of the network.

However, the problems with STDSNet should also be emphasized. From the visualization results in the second row of Fig. 11, it can be seen that the bottom-right portion of the result in STDSNet fails to effectively recognize the low vegetation category. This suggests that although our work has improved the semantic segmentation of RS images in general, there is still much room for improvement. In the future, we will study this issue in depth in our subsequent work.

3) Evaluation in Efficiency: For comprehensive comparisons, Table V lists the performance metrics of all models in the same experimental environment. These metrics comprise the number of floating-point operations per second (FLOPS), measured in gigaflops (G), and the number of parameters, measured in millions (M). For an input size of 3 × 512 × 512, most hybrid models consisting of CNNs and transformers obviously exceed the models solely based on CNNs in terms of the aforementioned metrics. In contrast to these methods based on pure CNN, it is evident that our proposed method entails a tradeoff between efficiency and accuracy, favoring the latter. Nevertheless, it is noteworthy to highlight that our STDSNet demonstrates favorable performance compared to a network with a transformer structure. Particularly when utilizing the same backbone in the encoder, STDSNet reduces the parameters by approximately half when compared to Swin-CNN, accompanied by a reduction in FLOPS of over 50%. While the parameters and FLOPS may limit STDSNet's applicability in real-time and lightweight scenarios, it still holds a promising future in the semantic segmentation of RS images.

V. CONCLUSION

In this article, we integrate the dual-stream network with the Swin transformer to create a novel network named STDSNet, which achieves more accurate semantic segmentation of RS images. The network follows a single encoder and dual decoders architecture. In the encoder, we employ the Swin transformer as the network backbone to address the limitations of CNN in capturing global context information and shape features. The dual decoder consists of two CNN-based parallel streams, namely the GS and the SS. The GS optimizes the utilization of global context, while the SS focuses on processing boundary information. We construct the GCFM in the GS to recover the loss of global context during the upsampling process. In addition, we design the GCM in the SS to filter out redundant noise that is not related to boundary information. As demonstrated by the experiments, the STDSNet outperforms other state-of-the-art methods in terms of mIoU and mF1 metrics on the ISPRS Vaihingen and Potsdam datasets. The result confirms the efficacy of the proposed method in effectively reducing large-scale regional object classification errors caused by similar features or shadow occlusion in RS images and improving the segmentation performance for tiny targets and their boundaries.

Although the dual-stream network has been designed to leverage the transformer's strengths in feature extraction and has achieved remarkable semantic segmentation performance, there is still room for further improvement. Specifically, the dual decoder architecture, which improves accuracy, also contributes to increased model parameters and computational complexity. In light of these limitations, we plan to devote our future research efforts toward streamlining the model architecture and optimizing the network to better accommodate various types of RS tasks.

REFERENCES

[1] S. Jiang, C. Jiang, and W. Jiang, "Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 230–251, Sep. 2020.

[2] W. Zhou et al., "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[3] J. Chen, Z. Liu, D. Jin, Y. Wang, F. Yang, and X. Bai, "Light transport induced domain adaptation for semantic segmentation in thermal infrared urban scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23194–23211, Dec. 2022.

[4] C. Liu et al., "Context-aware network for semantic segmentation toward large-scale point clouds in urban environments," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5703915.

[5] X. Mao et al., "PolSAR data-based land cover classification using dual-channel watershed region-merging segmentation and bagging-ELM," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 4000905.

[6] Z. Lv, X. Yang, X. Zhang, and J. A. Benediktsson, "Object-based sorted-histogram similarity measurement for detecting land cover change with VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2504405.

[7] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, Art. no. 111716, May 2020.

[8] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5215512.

[9] W. Bo, J. Liu, X. Fan, T. Tjahjadi, Q. Ye, and L. Fu, "BASNet: Burned area segmentation network for real-time detection of damage maps in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627913.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, LasVegas, NV, USA, 2016, pp. 770–778.

[12] 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.

[13] 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.

[14] 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.

[15] X. Chen et al., "Adaptive effective receptive field convolution for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3532–3546, Apr. 2021.

[16] X. He et al., "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[17] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5229–5238.

[18] 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.

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

[20] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany: Springer, 2018, pp. 418–434.

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

[22] M. Yin et al., "Disentangled non-local neural networks," in *Proc. 16th Eur. Conf.*, U.K.:Springer, 2020, pp. 191–207.

[23] L. Mou, Y. Hua, and X.-X. Zhu, "Relation matters: Relational contextaware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

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

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

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

[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] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[31] M.-M. Naseer et al., "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 23296–23308.

[32] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision?," 2021 , *arXiv:2105.07197*.

[33] A. Ghiasi et al., "What do vision transformers learn? A visual exploration," 2022, *arXiv:2212.06727*.

[34] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3126–3135.

[35] J. Wang et al., "CARAFE: Content-aware ReAssembly of FEatures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3007–3016.

[36] 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 Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506105.

[37] P. Song, J. Li, Z. An, H. Fan, and L. Fan, "CTMFNet: CNN and transformer multi-scale fusion network of remote sensing urban scene imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5900314.

[38] X. Zhang, Z. Xiao, D. Li, M. Fan, and L. Zhao, "Semantic segmentation of remote sensing images using multiscale decoding network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1492–1496, Sep. 2019.

[39] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[40] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 12, no. 4, Art. no. 701, Feb. 2020.

[41] 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.

[42] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019.

[43] 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. 2021.

[44] F. Zhou, R. Hang, H. Shuai, and Q. Liu, "Hierarchical context network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4407612.

[45] W. Wang, L. Yao, L. Chen, D. Cai, X. He, and W. Liu, "CrossFormer: A versatile vision transformer based on cross-scale attention," 2021, *arXiv:2108.00154*.

[46] R. Hang, G. Li, M. Xue, C. Dong, and J. Wei, "Identifying oceanic eddy with an edge-enhanced multiscale convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9198–9207, 2022.

[47] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5400314.

[48] D. Bhattacharjee, T. Zhang, S. Süsstrunk, and M. Salzmann, "Mult: An end-to-end multitask learning transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12031–12041.

[49] Y. Xu, Y. Yang and L. Zhang, "DeMT: Deformable mixer transformer for multi-task learning of dense prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 3072–3080.

[50] H. Zhang et al., "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.

[51] M. Guo et al., "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 1140–1156.

[52] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020 s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.

[53] C. Zhang, W. Jiang, and Q. Zhao, "Semantic segmentation of aerial imagery via split-attention networks with disentangled nonlocal and edge supervision," *Remote Sens.*, vol. 13, no. 6, Art. no. 1176, Mar. 2021.

[54] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746.

[55] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *Proc. 16th Int. Symp. Neural Netw.*, Russia: Springer, 2019, pp. 388–401.

[56] B. Wang, Z. Chen, L. Wu, X. Yang, and Y. Zhou, "SADA-Net: A shape feature optimization and multiscale context information-based water body extraction method for high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1744–1759, 2022.

[57] X. Li et al., "PointFlow: Flowing semantics through points for aerial image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4217–4226.

[58] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607713.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[60] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3585.

[61] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205.