HBSeNet: A Hybrid Bilateral Network for Accurate Semantic Segmentation of Remote Sensing Images

Thien Huynh-The, *Senior Member, IEEE*, Son Ngoc Truong, and Gia-Vuong Nguyen

***Abstract*—Semantic segmentation of aerial and satellite images plays a crucial role in a wide range of applications and services, catering to the increasing needs of environmental resource management, urban planning, and traffic safety. Many efficient semantic segmentation methods have been proposed by exploiting deep learning techniques with convolution neural networks (CNNs) architectures and self-attention mechanisms to achieve superior accuracy if compared with conventional machine learning-based approaches. In this article, we introduce a hybrid bilateral segmentation network (HBSeNet), a novel semantic segmentation architecture. Inspired by the success of dual-path network models that replace conventional single-branch encoder–decoder architectures, we construct a model with the core idea of combining the context path and the spatial path to optimize both the accuracy and the complexity of deep learning model in the field of remote sensing image segmentation. Moreover, HBSeNet innovates with auxiliary modules designed to enhance its performance, such as sequential atrous convolution, information synthesis module, and bridge for efficient multiscale feature extraction, fusion, and integration. In simulations, our model achieves a global accuracy of 92.04%, a mean intersection-over-union of 83.57%, and a mean boundary-F1-score of 90.23% when evaluated on the ISPRS Potsdam dataset, surpassing the state-of-the-art segmentation models, such as DeepLabV3+, SwinCNN, and ST-UNet.**

***Index Terms*—Artificial intelligence, bilateral network, image segmentation, object recognition, remote sensing.**

I. Introduction

SEMANTIC segmentation, or pixel-level classification, stands as a fundamental and computationally demanding task in image-based remote sensing applications and services. The objective is to assign an appropriate semantic label for each pixel within a remote-sensing image. Notably, semantic segmentation of very high-resolution and large-scale images has gained significant traction in various applications, including environmental monitoring, agricultural management, urban planning, and land cover classification. Remote sensing image (RSI) segmentation approaches traditionally fall into two broad categories: handcrafted feature-based and deep neural network (DNN)-based methods. Handcrafted methods rely on manually designed features and classifiers. The widely adopted simple linear iterative cluster algorithm typifies this approach, leveraging a k-means clustering scheme to efficiently produce superpixel features and partition RSIs into small neighboring pixel clusters. For instance, an innovative objective function and a novel graph construction strategy were introduced for superpixel segmentation [1], while the authors in [2] studied a robust SVM-HOG model, as the combination of histograms of oriented gradients (HOG) and support vector machine (SVM) for feature descriptor and classifier, respectively, employing GrowCut segmentation to capture handcrafted features. Subsequently, Markov models and region-based graph models are utilized to merge adjoining regions that have the highest similarity. However, very high-resolution and large-scale RSIs exhibit complex spectral characteristics, potentially leading to emphasized intraclass variance and a decline in interclass variance, thereby challenging the efficiency of such methods.

Inspired by the remarkable success across various computer vision tasks, semantic segmentation techniques based on deep learning (DL) have made substantial strides in both natural and remote sensing contexts. Nonetheless, unlike close-range natural images, the scale disparity inherent in RSIs presents a unique challenge. Specifically, small-scale land cover features can be lost with declining spatial resolution, ultimately compromising segmentation accuracy. Recognizing this characteristic, numerous research efforts have focused on constructing deep networks with different architectures adept at multiscale feature aggregation. For example, the work [3] implemented dilated convolution to enhance context information within feature aggregation, while the multiscale features extracted from various network layers (shallow and deep) are fused to enrich feature representation and learning efficiency. To further refine boundaries within RSIs, the authors [4] recently introduced a boundary attention module to capture land-cover boundary information from hierarchical feature aggregations. Overall, DL-based methods have demonstrably outperformed traditional approaches by a significant margin.

While several DL-based semantic segmentation methods have shown some applicable potential for RSIs, they face some substantial challenges. The large image sizes of RSIs often necessitate convolutional layers with expansive receptive fields to effectively seize object-specific information. This, however, can lead to a significant increase in both model complexity and the number of learnable parameters. Although atrous convolutions have been introduced to address this issue by expanding

Manuscript received 17 February 2024; revised 24 May 2024; accepted 21 July 2024. Date of publication 2 August 2024; date of current version 19 August 2024. This work was supported by Ho Chi Minh City University of Technology and Education (HCMUTE) under Grant T2024-150. *(Corresponding author: Gia-Vuong Nguyen.)*

The authors are with the Department of Computer and Communications Engineering, Ho Chi Minh City University of Education and Technology, Ho Chi Minh City 71307, Vietnam (e-mail: thienht@hcmute.edu.vn; sontn@hcmute.edu.vn; vuongng.cce@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3437737

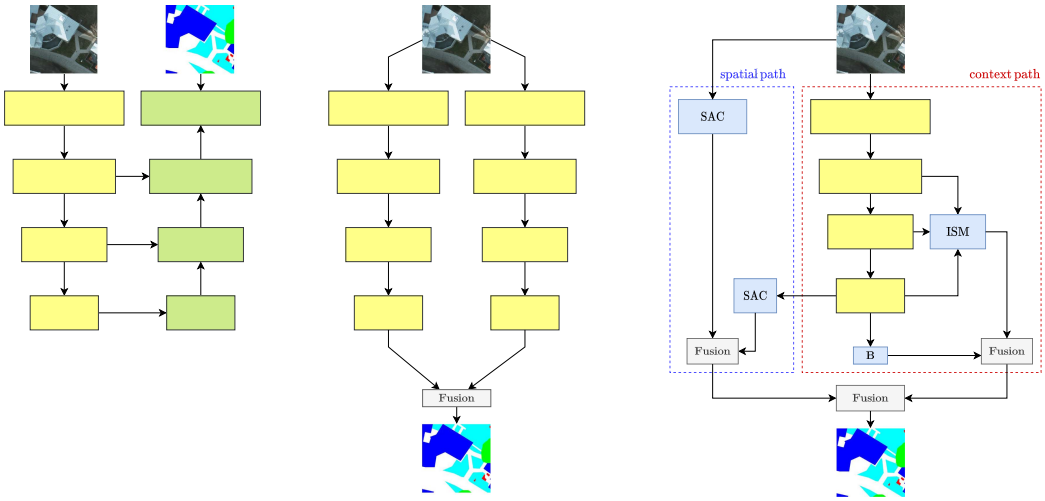


Fig. 1. Overview of different architectures are commonly used for image semantic segmentation. From left to right, it features U-shape, bilateral, and our proposed architectures. The yellow zone represents the down-sampling stage, while the green zone indicates the up-sampling stage.

receptive fields without increasing parameter count [5], their optimal adoption in network architectures remains a challenge. In addition, unlike natural images where target objects frequently occupy a large region in the image, RSIs often contain diverse object categories dispersed across varying regions. This dispersion can result in inaccurate boundary segmentation or, more problematically, pixel-level misclassification of object labels across extended areas.

To address the two abovementioned problems, we propose a semantic segmentation network model for RSIs, called the hybrid bilateral segmentation network (HBSeNet). Specifically, we design a novel densely connected sequential architecture, which establishes the foundation for building a spatial path to maximize the information gained while preserving complexity and processing time at levels deemed acceptable. In addition, inheriting theories from bilateral networks proposed in [6], [7], and [8], our HBSeNet is able to leverage several convolutional neural network (CNN) backbones to extract high-level abstraction features, along with clever enhancements of the conventional encoder–decoder architecture to construct the context path. This enhances the understanding capability of contextual information, especially when being combined with the spatial understanding of the spatial path to substantially boost the segmentation performance of the model.

In a nutshell, the key contributions of our work can be summarized as follows.

1) We propose a novel HBSeNet for RSI semantic segmentation, featuring a densely connected sequential architecture with the cooperation of spatial path and context path.
2) In the spatial path, we leverage atrous convolutions in an advanced structural connection to maximize information extraction for accurate segmentation while maintaining model efficiency.
3) We enhance the conventional encoder–decoder architecture by integrating effective auxiliary modules, establishing a well-organized context path, and strengthening its contextual learning capability.
4) Through diversified simulations, HBSeNet achieves superior performance on diverse RSI datasets, demonstrating the effectiveness of the proposed network architecture.

The rest of this article is organized as follows. Section II undertakes a comprehensive review of related works. Section III describes the details of our proposed method. Section IV conducts the experimental validation of the proposed method, along with a comparative assessment against other methodologies. Finally, Section V concludes this article.

II. Related Work

In recent times, semantic segmentation techniques for RSIs have emerged, offering effective tools for identifying objects from aerial views. These techniques can be categorized into various groups, including single-branch encoder–decoder models (denoted U-shape) and dual-path network models (denoted bilateral). Fig. 1 presents an overall comparison of these semantic segmentation architectures, including a depiction of our proposed model.

A. Single-Branch Encoder–Decoder Models

The semantic segmentation models have typically studied single-branch encoder–decoder architectures. These models with a single and unified pathway unveil their capability by distilling intricate spatial details and global context into accurate pixel-level predictions. In this section, we convey a brief review of several prominent single-branch models that have been introduced in the past. U-Net, SegNet, and fully convolutional network (FCN) are three well-known single-branch encoder–decoder architectures. U-Net [9] introduced the U-shaped architectural design, which uses skip connections to link the encoder and decoder. This allows the preservation of the fine-grained spatial information that could be lost during the downsampling process. U-Net achieved remarkable results in medical image segmentation with its simple and elegant design and later extended to various domains. However, its performance may suffer when dealing with complex scenes, due to its relatively shallow

encoders. In [10], SegNet prioritizes memory efficiency over U-Net by using pooling indices to perform upsampling without reconstructing the whole feature maps. With this beneficial feature, SegNet became a suitable solution for real-time applications, but it compromised the feature representation quality compared with other deeper architectures. FCN [11] was the pioneer of the encoder–decoder paradigm, in which it converted classification networks into fully convolutional ones. By using convolutions instead of fully connected layers, FCN enabled dense prediction for the whole image in semantic segmentation. However, its initial versions might not have cutting-edge context aggregation methods, thus being easily overcome by other recent models motivated by FCN.

Some deep models have been aptly designed to obtain high-resolution and context-aware semantic segmentation. For instance, DeepLabV3+ [12] leveraged atrous convolutions to analyze the scene at different scales, thus revealing the complex relationships between objects without losing resolution. As a result, DeepLabV3+ produced accurate segmentation maps with detail preservation, however, it may require more computational cost than other simpler architectures. Derived from accurately segmenting objects with varying sizes concept, the authors in [13] proposed an adaptive feature selection (AFS) module to address this issue. This module incorporates an attention mechanism to assign weights to different scales within the image, improving the segmentation of objects with uncertain scales. Moreover, SwinCNN [14] combines a Swin transformer for capturing long-range relationships between image features and CNN for handling local details. It utilizes various techniques to improve performance: atrous spatial pooling for capturing multiscale context, a U-shape decoder for progressively recovering image resolution, skip connections for preserving local details, and a channel attention block for feature enhancement. In the work [15], RefineNet was fabricated for accurate high-resolution segmentation by adopting a multipath refinement network that progressively enhances pixel-level classification. Although RefineNet improved the accuracy of semantic segmentation, especially for object boundaries, but its multipath structure may increase the overall network complexity and expose some difficulties in training and optimization. PSPNet [16] engaged in a pyramid pooling module that comprehensively aggregates context information at multiple levels of feature representation. Interestingly, PSPNet can handle distant and diverse elements in the scene, consequently generating a comprehensive view of the image. Indeed, these three models demonstrate how context aggregation and high-resolution segmentation can be achieved by different methods.

To optimize the performance tradeoff between accuracy and complexity, Unet++ [17] took advantage of residual connections by revolutionizing them to advanced structural architectures to handle learnable information more efficiently. Rather than combining neighboring levels, Unet++ convolved several dense connections in the decoder to gather multiscale features extracted at different presentation levels. With an advanced architecture, Unet++ prevented the details of small objects in a scene from attaining high-accuracy semantic segmentation, even if the computational expense may be burdensome for resource-constrained computing platforms. By exploiting a multiscale feature pyramid for semantic segmentation, feature pyramid networks [18] studied a selective attention mechanism in the decoder by extracting and learning salient features at multiple appropriate scales. Concretely, rich contextual information helps to localize objects in a scene more accurately and improve the overall segmentation accuracy. However, it is important to acknowledge that the inherent computational overhead relating to building and navigating this pyramidal structure increases significantly and potentially raises some implementation concerns on Internet-of-Things devices. Recently, ESPNet [19] introduced a spatial pyramid of atrous convolutions with different dilation factors by cleverly replacing a regular pyramid pooling with strategically dilated convolutions, thus enhancing learning efficiency and presenting a lightweight structure. This allows ESPNet to operate smoothly on edge devices besides delivering an improvement of segmentation accuracy without computational intensification. However, its structure mostly relying on dilated convolutions can limit its capability of coarse-to-fine feature presentations if compared with other regular pooling-aided pyramid mechanisms. Despite inherent technical constraints, the noteworthy achievements of these semantic segmentation models promote and emphasize the potential for advancing the overall system performance.

B. *Dual-Path Network Models*

The urgent requirements for accurate and efficient semantic segmentation for diverse application models have promoted the development of numerous cutting-edge architectures. Among these, dual-path networks have emerged as a potential solution by leveraging the learning capability of two distinct pathways: a spatial path focuses on local details and a semantic path (a.k.a., context path) plays the role of learning global context. As pioneering the dual-path paradigm, BiSeNet [6] was launched as a lightweight yet powerful architecture. Its spatial path enables to extraction of fine-grained information through convolutions, while the semantic path deploys dilated convolutions to accumulate globally contextual information. Moreover, the effective spatial attention mechanism deployed in BiSeNet fuses the features from both paths to improve segmentation accuracy and fulfill real-time processing. However, BiSeNet may be limited by intricate object boundaries due to its weak learning of global dependencies. Achieving a balance between detail and global comprehension in semantic segmentation models is a challenging issue. From this perspective, ContextNet [20] was proposed with a deep network architecture with two branches of feature learning: one aims to capture the global context and the remainder is to maintain fine-grained specifics. This innovative integration enables the model to comprehend both the macro and micro details of a scene for semantic segmentation. However, the dual-branch configuration of ContextNet may require a higher computing cost if compared with other straightforward and single-branch structures. LinkNet [21], on the other hand, leveraged the advancements of residual connections throughout its encoder–decoder structural architecture. These processes are able to prevent downsampling operations

and enable the preservation of fine-grained details, accordingly boosting segmentation accuracy. However, LinkNet's dependency on residual connections increased computational cost as well as reduced the processing speed if compared with less complicated architectural models. In [22], ICNet, building upon BiSeNet, introduced some intermediate channels to facilitate the information exchange issue among different paths. This enables bidirectional communication to enrich feature representations in both paths. In addition, channel-wise attention in ICNet refines feature aggregation, thus being able to yield superior segmentation performance for complex scenes. However, the increasing complexity of ICNet, compared with BiSeNet, may result in higher computational costs, which becomes an essential drawback when being implemented for resource-limited computing platforms. By embracing the ability of dense connections in feature emphasis, DANet [23] embedded them into both the spatial and semantic paths. This promotes information flow and feature reuse to consequently enhance learning efficiency. Remarkably, dual attention mechanisms further strengthen the interaction between paths and then allow the network to capture fine-grained dependencies and global context comprehensively. Although DANet presented impressive accuracy, its intricate architecture may demote its real-time applicability. The refinement continues with U^2-Net [24] by merging the U-Net architecture with squeeze-and-excitation blocks to advance the competence of relevant feature extraction and learning. These blocks dynamically adjust channel weights as well as highlight informative features and suppress irrelevant ones. Besides, residual connections organized in the U-Net structure facilitate information flow to generate accurate segmentation. However, U^2-Net's reliance on pretrained encoders like ResNet usually limits its flexibility in adapting to diverse datasets.

Several deep models have exploited attention mechanisms in the architecture of spatial and context paths. The authors in [25] proposed a novel approach called the BANet, which has two pathways: a dependence path to capture long-range relationships between objects and a texture path to capture fine-grained details within objects. These pathways are then effectively merged using a feature aggregation module with linear attention. Moreover, in [26], pixel-level attention is the center stage of a deep network, namely PANet, providing adaptive context aggregation for each pixel. This empowers the model to selectively focus on relevant contextual information based on its individual needs. Spatial path attention further enriches local details, leading to accurate boundary delineation. While PANet demonstrates impressive results, its reliance on multiple attention mechanisms might inflate its computational demands. In [27], BASNet employs backward attention which guides the spatial path by propagating semantic information from higher levels back to lower levels, effectively refining local features with global context. In addition, feature channel scaling adjusts the information volume exchanged between paths, ensuring efficient fusion. However, BASNet's backward attention mechanism might present challenges in optimization and training compared to standard forward flow approaches.

The field of semantic segmentation has witnessed an intensive manipulation of attention mechanisms to culminate in ever-refined feature representations. AUNet [28] aptly integrated the channel and spatial attention mechanism to attain a selective refinement of features. In particular, channel attention illuminates informative channels across the entire feature map, while spatial attention focuses on crucial regions within each channel. This dual-structural style amplifies feature saliency, which leads to an improvement in segmentation accuracy. However, AUNet's multiattentional ensemble necessitates careful balancing to avoid redundancy and computational strain [29]. To attain high-resolution image segmentation, HRNet [30] utilizes stage-wise residual connections for the progressive refinement of features. It leverages parallel aggregation of multiscale features to seize useful information at different granularities. In addition, channel attention modules dynamically scale feature channels, thus emphasizing salient information in scenes. Despite the fact that HRNet excels at semantic segmentation tasks for high-resolution RSIs, its reliance on multiple processing stages might increase its inference time. ST-UNet [31] ascends to the stage, effectively merging the transformative power of transformers with the U-Net architecture. Its encoder has a long-range dependence on Swin transformers and tightly works with the U-Net decoder, a guardian of spatial information. Moreover, hybrid attention mechanisms expedite context fusion across scales thoroughly. While ST-UNet's performance overwhelms other existing segmentation models, its computational demands might outshine those of purely convolutional networks. Notably, to combine the strengths of the transformer and the CNN, STDSNet [32] uses a Swin transformer as its backbone to overcome CNN limitations and incorporates a dual-decoder design with separate global and shape streams. The global stream tackles context loss during upsampling with a specific module, while the shape stream employs another module to focus on boundary information. This combination improves segmentation accuracy, particularly for small targets and their boundaries.

This intricate landscape of dual-path network models highlights the constant evolution in semantic segmentation. While each network possesses unique strengths and weaknesses, the unifying theme of exploiting the duality between spatial and semantic information remains a key driver of progress. Future research in this domain might focus on further optimizing attention mechanisms, reducing computational complexity, and adapting these architectures to handle diverse conditions.

III. Methodology

In the domain of semantic segmentation, recent advancements suggest that utilizing bilateral networks [6], [7], [33], [34] leads to superior segmentation performance compared to conventional single-branch networks [5], [12], [17]. Indeed, several research trends have focused on building models that integrate spatial and context paths, achieving significant success in enhancing semantic segmentation accuracy [6], [7], [8]. Building upon these concepts, our work introduces a novel network model that segregates the spatial and context paths with the incorporation of innovative enhancement methods to achieve superior image segmentation performance in the domain of remote sensing. In this section, we will detail the architecture of our proposed

Algorithm 1: The Implementation of SAC Module.

Require: A set of dilation rates for N elements $a_1, a_2, \ldots, a_N$.
Ensure: $a_1 < a_2 < \ldots < a_N$.
  $i \leftarrow 1$,
  $\mathbf{T}_1 \leftarrow \mathbf{SAC}_{in}$,
  $\mathbf{K} \leftarrow \mathbf{T}_1$,
  for $i < N$ **do**
    $\mathbf{T}_{i+1} = \mathcal{A}^{3\times3}_{1,a_i}(\mathbf{T}_i)$,
    $\mathbf{K} \leftarrow \langle \mathbf{K}, \mathbf{T}_{i+1} \rangle$,
    $i \leftarrow i + 1$,
  end for
  $\mathbf{SAC}_{out} \leftarrow \mathbf{K}$.

spatial path, context path, and the comprehensive model, referred to as the HBSeNet.

A. Spatial Path

With the introduction of atrous convolutions [5], which facilitate the expansion of kernel receptive fields while preserving the number of parameters, more and more studies are leveraging them to enhance model performance [6], [12], [35]. Primarily, these investigations concentrate on integrating atrous convolutions in the last layers of the backbone to enlarge the receptive field for feature maps at higher abstraction levels [5], [12]. However, challenges arise when adopting this approach in real-time applications with low-resolution input images, where the feature map size at high abstraction levels becomes smaller than the receptive field of the kernels, thus leading to potential unnecessary weight redundancy [36]. An alternative utilization of atrous convolution involves the implementation of atrous spatial pyramid pooling (ASPP) modules for extracting multiscale feature information [12], [37]. The ASPP employs parallel atrous convolutional layers with varying reception fields on the same input. Consequently, kernels with large receptive fields may not fully capitalize on the relevant information obtained through filters with smaller receptive fields, and vice versa. Addressing this problem requires a substantial number of filters with the same receptive field to ensure comprehensive information, hence resulting in a surge in parameter count with just one ASPP module.

Sequential atrous convolution (SAC): To overcome these limitations, we propose a new approach via the architecture named SAC, transforming conventional parallel connections into a dense sequential connection. Each SAC module comprises a set of N atrous convolutional layers with corresponding a set of dilation rates $a_1, a_2, \ldots, a_N$. The pseudocode of the whole proposed SAC module is presented by Algorithm 1, where $\langle \cdot \rangle$ is a depth-wise concatenation, $\mathcal{A}^{3\times3}_{1,a_i}$ denotes a sequential operation, including an atrous convolution (specified by the filter size 3×3, the stride 1, and the identical dilation rate $\mathbf{a}_i$), a batch normalization (bn), and a leaky rectified linear unit (leaky ReLU) activation. It is noted that $\mathbf{T}$ and $\mathbf{K}$ are variables that store temporary features used for data processing in SAC modules.

Sequential deployment of atrous convolutional layers enables us to leverage information extracted from previous layers. As a result, this leads to a reduction in the number of kernels needed in a layer, thereby significantly reducing the number of parameters in the SAC module regardless of its placement within the model. It is worth noting that, to preserve important feature information with such a small number of parameters, we adopt to use the leaky ReLU function [38] with a scale of 0.01, rather than the standard ReLU function, within the SAC module.

Based on the SAC module, we propose the establishment of a spatial path by incorporating two distinct SAC modules: one at the input (denoted as $\mathbf{S}_{\text{in1}}$) and another at the final layer in the backbone (denoted as $\mathbf{S}_{\text{in2}}$). It is realized that there exists a difference in the level of abstraction and resolution of the inputs of the two SAC modules. For details, $\mathbf{S}_{\text{in1}}$ has a large resolution but a low abstraction level, and $\mathbf{S}_{\text{in2}}$ has a relatively small resolution but a higher abstraction level. For that reason, $\mathbf{S}_{\text{in2}}$ may require more learnable parameters as well as guarantee that the receptive field should be less than or equal to its spatial size. The following describes the detailed architecture of modules to process $\mathbf{S}_{\text{in1}}$ and $\mathbf{S}_{\text{in2}}$. In particular, for $\mathbf{S}_{\text{in1}}$, we design an SAC module comprising six atrous convolutional layers with varying dilation rates, including three layers with small rates ($[1, 2, 3]$) and three layers with large rates ($[5, 7, 9]$), strategically focusing on spatial features of the input at multiscale of reception fields. For $\mathbf{S}_{\text{in2}}$, the module SAC must be specified for a notably reduced size, which means, we should fabricate four atrous convolutional layers with relatively small dilation rates to address the weight redundancy issues as studied in [36]. The detailed architecture of the spatial path can be referred to in Fig. 2. The output of the spatial path comprises a concatenation of spatial information at both low and high abstraction levels.

B. Context Path

While the spatial path captures spatial information within the image, the context path is a powerful tool for gathering contextual information across various abstraction levels. Similar to encoder–decode architectures [5], [7], [17], where encoded information (obtained through a DNN, or backbone) is passed through up-sampling steps to generate the output segmentation map, we propose an innovative context path built on ResNet50 [39] as the backbone. However, in contrast to standard ResNet50, we exclude the last stage to maintain feature map resolution and reduce the number of trainable parameters in our proposed backbone. This section will offer a comprehensive overview of our context path, including technical details of the information synthesis module (ISM) and the bridge to the context path, both designed to improve feature integration for more accurate segmentation.

ISM: In the context path, besides constructing a model entirely based on typical encoder–decoder architectures, we recognized the importance of introducing some enhancements. In some recent works, researchers have highlighted the importance of constructing information transmission from encoder to decoder to help the up-sampling stage [6], [10], [40]. Inheriting those ideas, we propose to build an ISM. This module serves to amalgamate information from various stages in the backbone, thereby facilitating the upsampling process with a diverse array

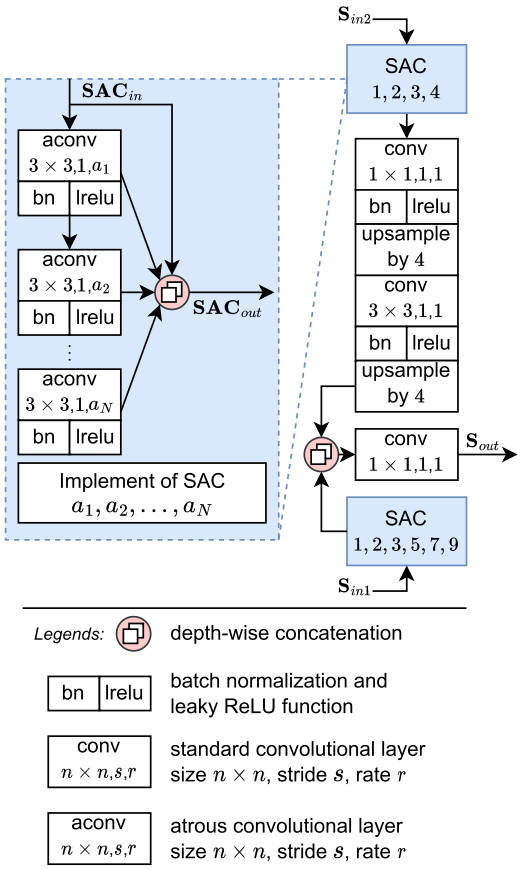


Fig. 2. Our proposed spatial path consists of 2 SAC modules. The detailed implementation of the SAC module is highlighted in the blue zone. For instance, the SAC for $\mathbf{S}_{\text{in1}}$ has 6 atrous convolutions with the set of dilation rates is $[1, 2, 3, 5, 7, 9]$.

of information from different contexts. The ISM module can be succinctly expressed through the following mathematical equations:

$$\mathbf{ISM}_{\text{out}} = \mathbf{C} \odot \mathbf{s} \tag{1}$$

with $\mathbf{ISM}_{\text{out}}$ denotes the output of the ISM module and $\mathbf{C}$ are the feature maps that contain information from various stages referred in the algorithm

$$\begin{aligned} \Theta_i &= \mathcal{C}_{1,1}^{1\times1}\left(\mathbf{X}_i\right) \\ \mathbf{C} &= \langle \Theta_1, \Theta_2, \Theta_3 \rangle \end{aligned} \tag{2}$$

and $\mathbf{s}$ is the channel-wise attention scores [41] calculated by the following equations:

$$\begin{aligned} \mathbf{z}(c) &= \frac{1}{\mathrm{H} \times \mathrm{W}} \sum_{h=1}^{\mathrm{H}} \sum_{w=1}^{\mathrm{W}} \mathbf{C}(h, w, c) \\ \mathbf{s} &= \sigma\left(\mathbf{W}_2 \delta\left(\mathbf{W}_1 \mathbf{z}\right)\right). \end{aligned} \tag{3}$$

The detailed descriptions of the meaning of mathematical symbols used in (2) and (3) can be referred to Table I.

In our context path implementation, we utilize the ReLU activation function due to its remarkable success in constructing deep image segmentation networks [6], [10], [17]. Conversely, owing to the nature of the context path, a substantial number of parameters are required to attain high-level abstract information. Hence, employing leaky ReLU to prevent information loss, as in the spatial path, becomes unnecessary. In addition, we adopt the output information from stages 2, 3, and 4 of the backbone as input to the ISM module to maximize contextual information. We omit the information from the first stage because it has a low level of abstraction and can be effectively retained by the spatial path.

TABLE I
DESCRIPTIONS OF THE MEANING OF MATHEMATICAL SYMBOLS

Notation	Description
$\mathbf{X}_i$	Output feature maps of singular stage $i+1$.
$\mathcal{C}_{1,1}^{1\times1}$	A sequential operation, including a standard convolution (with filter size 3×3, stride 1, and dilation rate 1), bn, and ReLU.
Θ_i	Extracted information from $\mathbf{X}_i$.
$\mathbf{C}$	Synthesized information from various stages.
H and W	Height and width of $\mathbf{C}$, respectively.
$\mathbf{z}(c)$	Vector containing average value of channel c.
σ	ReLU activation function.
δ	Sigmoid activation function.
$\mathbf{W}_1$ and $\mathbf{W}_2$	Weight matrices of two fully connected layers.

Bridge: The concept of a bridge, as defined in previous works [42], is a module that facilitates the connection between the last layer of the encoder and the first layer of the decoder. Its purpose is to enhance information synthesis and improve segmentation capabilities. In our model, we employ a simple bridge architecture consisting of just two convolutional layers with distinct receptive fields, applied to the same feature maps. To safeguard against any information loss, the output of the bridge module undergoes a concatenation layer. A detailed illustration of the bridge module can be referred to the following equation:

$$\mathbf{B}_{\text{out}} = \left\langle \mathcal{C}_{1,1}^{1\times1}\left(\mathbf{X_3}\right), {}^{\delta}\mathcal{A}_{1,4}^{3\times3}\left(\mathbf{X_3}\right) \right\rangle \tag{4}$$

where $\mathbf{B}_{\text{out}}$ denotes the output feature map of bridge and ${}^{\delta}\mathcal{A}_{1,4}^{3\times3}$ is sequential operation similar to $\mathcal{A}_{1,4}^{3\times3}$ but using ReLU activation instead.

C. Network Architecture

To extract contextual information from RSIs, we construct a context path using a truncated ResNet50 backbone, an ISM module, and a bridge. By integrating this context path with a spatial path that leverages the coarse information of the input image as $\mathbf{S}_{\text{in1}}$ and the fine features extracted at the stage 4 as $\mathbf{S}_{\text{in2}}$, we introduce a novel semantic segmentation model named HBSeNet. This model stands out as a high-performing solution for remote sensing object recognition. A detailed depiction of the entire architecture is provided in Fig. 3, where the spatial path has been abbreviated. The amalgamation of spatial and contextual information occurs through an additional layer, and a concluding convolution layer is employed to generate the output with the specified number of categories.

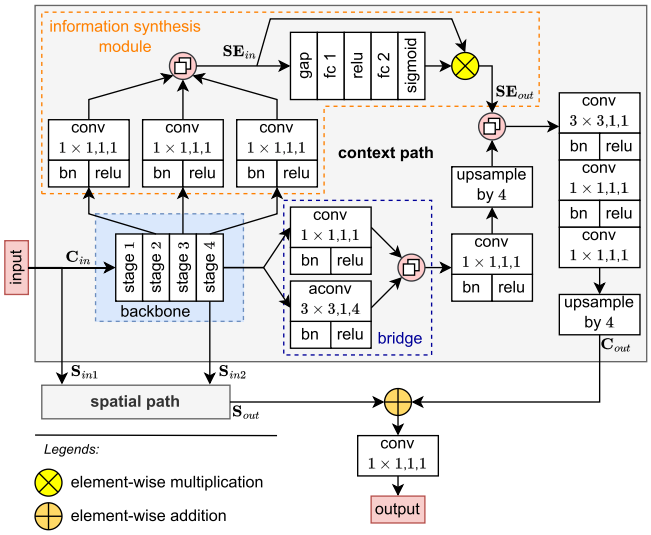


Fig. 3. Our proposed comprehensive model, information will be synthesized based on a layer combining outputs obtained from both the spatial path and the context path. Subsequently, it will be passed through a convolutional layer to produce the satellite image segmentation mask. The gray zone depicts the context path with truncated ResNet50 backbone, modules ISM, and bridge.

Loss function: In this article, we adopt a loss function to supervise the output of the HBSeNet model. Specifically, during the training process, we utilize the cross-entropy loss function to assess the discrepancy between the predicted results of the proposed model and the actual ground truth, enabling the update of the learning weights to construct the model. The mathematical representation of the cross-entropy loss function can be referred to the following equation:

$$\mathcal{L} = -\frac{1}{N_p}\sum_{i=1}^{N_p}\sum_{c=1}^{N_c} y_{i,c}\log(p_{i,c}) \tag{5}$$

where $\mathcal{L}$ denotes the cross-entropy loss, N_p denotes the number of pixels needed to classify in a minibatch, N_c is the number of categories, $y_{i,c}$ is a binary indicator, with a value of 1 if the class label c is the correct classification for the observed pixel i and 0 otherwise, and $p_{i,c}$ is the predicted probability observed pixel i is of class c.

IV. Experimental Results

We conduct our evaluation sequentially on common and widely used datasets, such as UAVid [43], ISPRS Vaihingen [44], and ISPRS Potsdam [45]. First, we provide a brief introduction to the datasets and a detailed description of our implementation. Then, we investigate the ablation studies, where we successively remove the improvement components from the model and compare the impact of each component on the model's performance. The ablation study is implemented and evaluated on the UAVid dataset. Finally, we present a comparison in terms of both accuracy and processing speed between HBSeNet and other state-of-the-art models on the UAVid, ISPRS Vaihingen, and ISPRS Potsdam datasets to provide an overall evaluation of our proposed model.

A. Datasets

UAVid dataset:[1] The UAVid dataset is a valuable contribution to the field of remote sensing, specializing in semantic segmentation of urban scenes captured by unmanned aerial vehicles. This study focuses on the UAVid dataset's image component, which comprises 42 sequences distributed across a training set with 20 sequences, a validation set with 7 sequences, and a test set with 15 sequences. Each sequence consists of a series of consecutive images and their corresponding pixel-perfect semantic segmentation labels (segmentation masks), originally captured in 4 K resolution and downscaled to 540 × 960 pixels. Notably, the dataset encompasses eight distinct object categories: *building*, *road*, *static car*, *tree*, *low vegetation*, *human*, *moving car*, and *background clutter*.

ISPRS Vaihingen dataset:[2] The ISPRS Vaihingen dataset is a well-known benchmark for remote sensing applications, particularly in semantic segmentation and classification. It consists of high-resolution aerial images of Vaihingen, Germany, captured in RGB channels. Our study focuses specifically on the image segmentation aspect of the dataset. Comprising 33 individual image patches of varying sizes, each patch incorporates a sequence of images along with corresponding semantic segmentation labels. Originally provided with an average resolution of approximately 2500 × 2000 pixels, these images and masks are cropped to 512 × 512 pixels. The dataset covers land cover categories with five object classes: *impervious surfaces*, *buildings*, *low vegetation*, *trees*, and *cars* besides *background clutters*. We solely utilize the IRRG bands in our experiment, excluding the digital surface model information. For this work, we use 11 patches of images $(1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37)$ for training, and five remaining patches $(11, 15, 28, 30, 40)$ for testing as following the previous works [46], [47].

ISPRS Potsdam dataset:[3] Similarly to the ISPRS Vaihingen dataset, the ISPRS Potsdam dataset is a valuable resource for remote sensing research. In our work, this dataset is utilized for its image segmentation. The original high resolution with 6000 × 6000, is cropped to 512 × 512. The dataset encompasses various urban object categories, including *impervious surface*, *building*, *low vegetation*, *tree*, *car*, and *background clutter*. Following the settings of most research [48], [49], we use 17 tiles for training, 7 tiles for validating, and the remaining 14 tiles for testing. Specifically, all 38 tiles are divided into training set (17 images, IDs: 2_10, 3_10, 3_11, 3_12, 4_11, 4_12, 5_10, 5_12, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_9, 7_11, 7_12); validation set (7 images, IDs: 2_11, 2_12, 4_10, 5_11, 6_7, 7_8, 7_10); and test set (the remaining images). We only utilize the RGB bands of TOP mentioned above without ground reference data lacking eroded boundaries, and the evaluation results are, therefore, not as high as reported in some examples in the literature.

[1][Online]. Available: https://uavid.nl/

[2] [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[3][Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

B. Implementation Details

Training configurations: For training configuration, we employ weight initialization according to the *Glorot* standard [50]. Specifically, the convolutional layers in the backbone utilize weights pretrained on the ImageNet dataset [51]. In the training phase, all weights (a.k.a., trainable parameters) are iteratively updated using the stochastic gradient descent with momentum with the momentum parameter $\beta = 0.9$. The initial learning rate is set to 2.5^{-2} and decayed by a factor of 0.3 after every 10 epochs. In addition, $L2$ regularization with a coefficient of 0.0001 is applied to mitigate model overfitting. Notably, all network models are trained for 60 epochs to ensure model convergence, in which the mini-batch size is 8 for the UAVid dataset and 12 for the remaining datasets. Both the training and the evaluation phase are implemented in MATLAB R2023b and are trained on one RTX2080 GPU.

Data augmentation: As mentioned earlier, images are first resized to a predetermined resolution, 540×960 for the UAViD dataset and 512×512 for the ISPRS Vaihingen and Postdam datasets, to ensure compatibility with the neural network architecture. This preprocessing allows for simplifying feature extraction and reducing computational complexity. For data augmentation, we apply random left-right reflection to randomly flip the image horizontally along its vertical axis with a range $[-10, 10]$ pixel. This augmentation allows the model to learn more diversified features that are independent of the object's orientation in the image [48], [49]. In real-world scenarios, objects can be viewed from various angles, and flipping helps the model generalize better to unseen directions.

Evaluation metrics: The segmentation results of all models are evaluated using metrics that are common and widely adopted in the field of semantic segmentation [52]. Specifically, we employ global accuracy, mean intersection-over-union (IoU), and mean boundary-F1-score (BFScore) to assess model accuracy. For comparisons of processing speed, we recommend using frames per second (fps). All results are the average of each metric value over 3 executions. Due to the unavailability of segmentation masks in the testing set, our complete segmentation results are based on the metric values obtained during the evaluation process on the validation set.

C. Ablation Study

In this section, we conduct a comprehensive ablation study to investigate the impact of each component within our proposed HBSeNet architecture on both accuracy and processing speed. The ablation is performed in a step-by-step manner, with evaluations based on the UAVid dataset.

Ablation for bridge: As mentioned in Section II, we design a simple bridge module to connect the downsampling (as encoder) and the upsampling (as decoder) stages through two convolutional layers and a concatenation layer. This module aids in better apprehending abstract information, thereby enhancing the model's learning capabilities and increasing segmentation accuracy compared with the model with a sole context path, however, the detailed improvements presented in Table II are not impressive. The results of the model constructed by context path with only bridge module, denoted as CP (B), in the global accuracy, mean intersection over union (mean IoU), and mean boundary-F1-score (mean BFScore) of 82.97%, 54.36%, and 69.50%, respectively. Compared with the naïve architecture involving only a clean context path (i.e., without improvements of ISM and bridge modules), the bridge presents some slight improvements in global accuracy, mean IoU, and mean BFScore with less than 0.7%. The effectiveness of the bridge module is also proven through its higher performance when combined with the CP (ISM) model, showing that our improvement strategy is truly effective and highly flexible. Meanwhile, the decrease in the fps metric indicates that accuracy enhancement through high-level abstraction features analysis comes with slowing down processing speed.

TABLE II
DETAILED PERFORMANCE COMPARISON OF COMPONENTS IN OUR PROPOSED HBSENET

Method	Global Acc	Mean IoU	Mean BFScore	Speed (fps)	Params (M)
CP	82.59	53.99	69.05	15.46	14.2
CP (B)	82.97	54.36	69.50	14.79	14.6
CP (ISM)	84.41	58.89	71.37	13.36	16.6
CP (ISM + B)	84.98	59.36	71.41	12.93	17.1
CP (ISM + B) + SP	**85.11**	**60.04**	**71.73**	**11.55**	**17.5**

CP: context path; SP: spatial path; ISM:Information synthesis module; B: bridge. The evaluation is implemented with 540 × 960 resolution input images and an RTX2080 GPU.
The values are bolded to emphasize their superiority in the measurements.

Ablation for ISM: With ISM, the crucial encoding information of an image is preserved and emphasized at multiple deeper analysis stages. This abundance of necessary information refines the up-sampling stage, thus resulting in an accuracy increase within the output segmentation map. Notably, ISM demonstrates a considerable improvement over using solely the context path, as it increases accuracy metrics by approximately 2.21%, average IoU by 9.09%, and average BFScore by 3.35% (as shown in Table II). This substantial enhancement underscores the importance of enriching multiscale relevant information acquired during downsampling and subsequently conveyed to the upsampling process. However, it is worth noting that the computation complexity of the model increases with the significant increase in a number of learnable parameters [approximately 2.4 millions (M)] and the processing speed being slower than the non-ISM model (a.k.a., only the context path in consideration). Furthermore, ISM remarkably delivers a significant improvement in model accuracy even when the contextual path includes an additional bridge module. In this scenario, the global accuracy, mean IoU, and mean BFScore exhibit increases of 2.43%, 9.18%, and 2.74%, respectively. These results convincingly demonstrate the adaptability and compatibility of ISM within our proposed models.

Ablation for spatial path: The spatial path bolsters the model's ability to capture spatial information, with the SAC architecture augmenting class recognition capabilities. This significantly enhances accuracy compared to versions without the spatial path. By effectively extracting information from preceding convolutional layers using stacked atrous convolutions, we minimize the

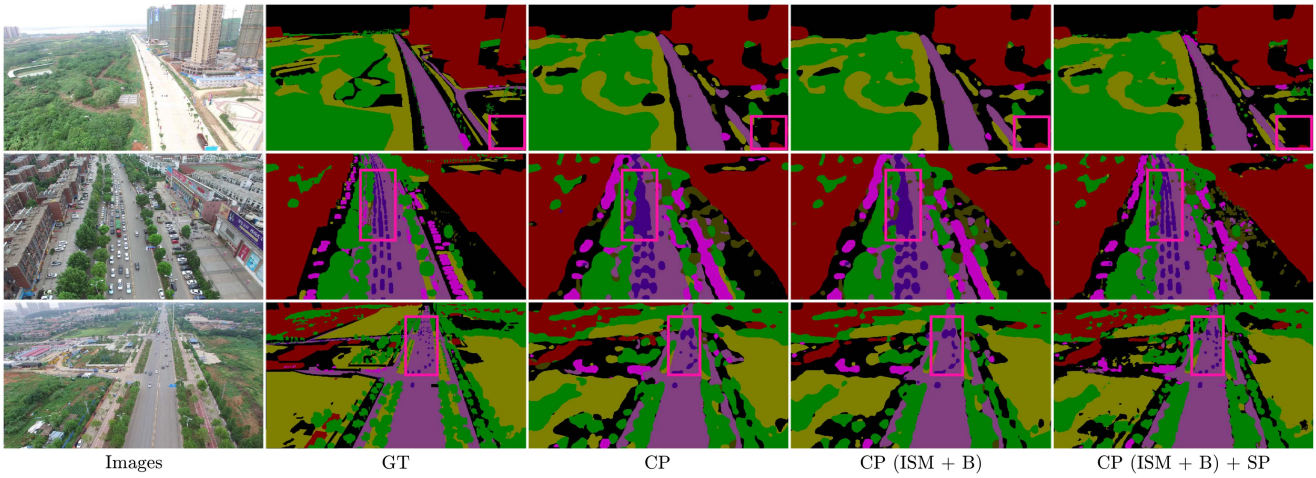


Fig. 4. Visualization results (as segmentation masks) for ablation study. Here are 3 examples of the results of the output before and after adding our proposal methods. The pink rectangles are used to emphasize the regions that show clear improvement in segmentation accuracy. GT: Ground truth; CP: Context path; SP: Spatial path; ISM: Information synthesis module; B: Bridge.

number of required weights. To prevent information loss when the number of learning weights in the spatial path is notably small (approximately 400 K), we strategically employ the leaky ReLU activation function [38]. Incorporating both the completed context path and spatial path yields the highest segmentation accuracy across all metrics: 85.11% global accuracy, 60.04% mean IoU, and 71.73% mean BFScore. However, this comes with increased complexity (approximately 17.5 M parameters) and a decrease in model speed to approximately 11.55 fps when evaluated on the UAVid dataset [43] at a resolution of 540×960. A comprehensive presentation of the evaluation results with both paths is provided in Table II.

In summary, the ablation study provides a comprehensive overview of the importance and effects of each module on the model's performance. Indeed, all modules proposed in HBSeNet contribute to improved accuracy, with ISM and the spatial path having the most significant impact. However, this enhanced accuracy comes with increased model complexity and higher execution time, as reflected in the number of parameters and fps. To facilitate a visual evaluation of the segmentation output accuracy, we present the segmentation results as image overlays in Fig. 4, in which the results are produced from various configurations of model architecture, including the naive context path, the full context path with ISM and bridge, and the complete HBSeNet with spatial path and full context path. Our analysis highlights the superior segmentation accuracy of the complete HBSeNet, further underscoring its robust recognition capabilities. This is most evident in the pink rectangle of visualization, where the complete version demonstrably outperforms its incomplete counterparts. In the first case, while confusion persists in classifying the former incomplete versions, our comprehensive model exhibits a superior classification probability, this advantage is exemplified by the reduction of the number of misclassified pixels observed in the lower right corner region. The second and third cases reveal that the complete model significantly enhances boundary recognition for moving car objects across various scenes. This improvement is particularly evident in its superior discrimination of singular moving cars on the road, where the incomplete version fails to establish clear distinctions.

D. Comparison Results

The HBSeNet's performance is obtained comprehensively through our comprehensive evaluation, where we not only dissect its components but also compare it to cutting-edge image segmentation models in terms of accuracy and complexity on the UAVid, ISPRS Vaihingen, and ISPRS Potsdam datasets, using standard quantitative metrics, such as global accuracy, mean IoU, mean BFScore, and model size (a.k.a., the total number of trainable parameters). In addition, intensive discussions are provided regarding experimental results for better insights and analysis of our work.

Overall performance analysis: In our implementation, both compared models experience training convergence to ensure the reliability of our method. In addition, to prove the model flexibility across different datasets, we demonstrate accuracy based on global accuracy, mean IoU, and mean BFScore measurements on UAVid, ISPRS Vaihingen, and ISPRS Potsdam datasets with various image sizes. Based on the results presented in Table III, our HBSeNet model achieves superior segmentation results with higher accuracy compared to other existing models. This analysis delves into these results, highlighting HBSeNet's effectiveness in segmenting objects within aerial and satellite imagery.

First, on the UAVid dataset results, HBSeNet emerges as the leader in all three key performance metrics: global accuracy of 85.11%, mean IoU of 60.04%, and mean BFScore of 71.73%. HBSeNet outperforms all the compared models, including DeepLabV3+ (with 84.65% global accuracy, 59.26% mean IoU, and 69.11% mean BFScore), ShelfNet (with 84.25% global accuracy, 57.61% mean IoU, and 70.92% mean BFScore), and others. This achievement underscores HBSeNet's capability in accurately segmenting diverse objects, such as buildings, road,

TABLE III
METHOD COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART MODELS IN TERMS OF SEGMENTATION PERFORMANCE AND MODEL SIZE

UAVid dataset at resolution of 540×960

Method	Backbone	Global Acc	Mean IoU	Mean BFScore	Params (M)
DeepLabV3-AFS [13]	R50	77.01	51.77	61.90	27.8
PSPNet-AFS [13]	R50	76.88	51.65	61.65	34.9
BANet [25]	ResT-Lite	81.99	55.19	65.86	**10.5**
CG-Swin [52]	Swin-S	83.71	57.19	66.28	N/A
LWN-A-F [46]	EffNet-B1	82.68	56.27	65.11	15.0
BiSeNet [6]	R50	83.06	54.41	68.70	49.0
ShelfNet [53]	R50	84.25	57.61	70.92	35.6
DeepLabV3+ [12]	R50	84.65	59.26	69.11	43.9
HBSeNet (Our)	R50	**85.11**	**60.04**	**71.73**	17.5

ISPRS Vaihingen dataset at resolution of 512×512

Method	Backbone	Global Acc	Mean IoU	Mean BFScore	Params (M)
DeepLabV3+ [12]	R101	89.62	80.51	88.43	62.4
DANet [23]	R101	90.32	81.14	89.22	68.5
GFFNet [54]	HRV2-W48	91.27	81.84	89.95	74.3
RSSFormer [55]	HRV2-W48	90.93	81.58	89.41	72.9
STDSNet [31]	Swin-B	90.25	81.77	89.81	130.1
SwinCNN [14]	Swin-B	90.01	81.11	89.41	235.8
ST-UNet [30]	R101	89.96	80.89	89.29	208.4
HBSeNet (Our)	R101	**91.30**	**82.35**	**89.99**	**36.5**

ISPRS Potsdam dataset at resolution of 512×512

Method	Backbone	Global Acc	Mean IoU	Mean BFScore	Params (M)
DeepLabV3+ [12]	R101	90.12	80.53	89.34	62.4
DANet [23]	R101	90.72	81.41	89.63	68.5
GFFNet [54]	HRV2-W48	91.27	81.22	89.95	74.3
RSSFormer [55]	HRV2-W48	90.80	81.81	89.47	72.9
STDSNet [31]	Swin-B	91.39	82.33	90.17	130.1
SwinCNN [14]	Swin-B	90.94	81.79	89.86	235.8
ST-UNet [30]	R101	90.58	81.64	89.74	208.4
HBSeNet (Our)	R101	**92.04**	**83.57**	**90.23**	**36.5**

The values are bolded to emphasize their superiority in the measurements.

vehicles, and human within UAV imagery, while also excelling in capturing accurate object boundaries. It is important to note that most methods in the UAVid dataset comparison and HBSeNet itself utilize the ResNet-50 (R50) backbone. This suggests that R50 might be well-suited for semantic segmentation tasks involving aerial imagery.

The second experiment is on ISPRS Vaihingen and Potsdam datasets, in this situation, HBSeNet also maintains the leading performance. The trends observed on the UAVid dataset extend to the ISPRS Vaihingen and Potsdam datasets as well. HBSeNet consistently achieves the highest global accuracy (in particular, 91.30% for Vaihingen and 92.04% for Potsdam), mean IoU (82.35% for Vaihingen and 83.57% for Potsdam), and mean BFScore (89.99% for Vaihingen and 90.23% for Potsdam). This further solidifies HBSeNet's position as a leading semantic segmentation method across diverse aerial and satellite imagery scenarios. Furthermore, for the ISPRS Vaihingen and Potsdam datasets, the table showcases a wider variety of backbones, including ResNet-101 (R101), Swin-B, and others. HBSeNet maintains its leading position even when employing R101, demonstrating its effectiveness with different backbones.

Detailed accuracy analysis: To demonstrate more detail about our proposed model capacity in individual class recognition, we provide comprehensive results about the comparison of mean IoU between our method against other state-of-the-art methods.

UAVid dataset: In Table IV, a comprehensive assessment of accuracy across distinct categories is facilitated through the mean IoU benchmark. Overall, HBSeNet emerges as the leader in mean IoU across all classes on the UAVid dataset, it is true that individual class mean IoU reveals that HBSeNet excels in segmenting several crucial classes. It also presents the highest mean IoU for *background clutter* with 57.32%, *road* with 74.27%, *trees* with 75.68%, and *moving cars* with 63.96%, and *static cars* with 46.56%. The only exception is *low vegetation* when our HBSeNet performance is slightly lower than DeepLabV3+. Notably, HBSeNet archives the highest mean IoU on *buildings* with 89.67%, followed by DeepLabV3+ and ShelfNet, and the remaining methods, which show a significantly lower than our results (around 81.11%–87.52%). This achievement highlights HBSeNet's capability in accurately segmenting diverse objects within the UAVid dataset's aerial imagery. On the other hand, our proposed model yields only 5.69% mean IoU for *humans* class. Although it is the best performance compared with other state-of-the-art methods, our model still does not have significant improvements in segmenting small objects, which forms a critical weakness of HBSeNet when facing small objects that require high-precision recognition.

ISPRS Vaihingen: Table V compares the performance of various semantic segmentation methods on the ISPRS Vaihingen dataset. The proposed method, HBSeNet, reports as the leader in mean IoU of *Impervious Surfaces* with 87.76% on the Vaihingen dataset. This outperforms all the compared methods, including DeepLabV3+ with 84.82%, DANet 86.32%, GFFNet 87.38%, RSSFormer 86.74%, STDSNet 86.09%, SwinCNN 85.45%, and ST-UNet 85.24%. This achievement demonstrates HBSeNet's capability in precisely segmenting large objects within the Vaihingen dataset. Notably, HBSeNet exhibits competitive results in other individual classes as well. For example, HBSeNet reaches the highest mean IoU for different classes, such as *low vegetation* with 73.23%, *trees* with 81.18%, and *cars* with 77.62%, to demonstrate its proficiency in segmenting these crucial elements. For the *buildings* class, HBSeNet obtains the second-best mean IoU with 91.98%, slightly lower than STDSNet. The marginal difference for *buildings* suggests comparable performance, while HBSeNet remains highly competitive for it. Overall, HBSeNet delivers an exceptional performance across all Vaihingen dataset classes. It is important to acknowledge that STDSNet achieves a slightly higher mean IoU for *Buildings*. This might indicate a specialization in building segmentation for STDSNet and the weakness of HBSeNet when having a simple decoder with continuously upsampling layers, making it difficult for our proposed model to recognize square objects, such as buildings. However, HBSeNet maintains strong building segmentation while achieving the best overall mean IoU, signifying a more balanced performance across all classes.

TABLE IV
CLASS-WISE PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART MODELS ON THE UAVID

Method	Backbone	Bac. Clutter	Building	Road	Tree	Low Veg.	Mov. Car	Sta. Car	Human
DeepLabV3-AFS [13]	R50	47.54	82.94	64.69	59.01	57.13	56.78	42.73	3.37
PSPNet-AFS [13]	R50	48.66	81.11	63.73	59.67	55.86	58.65	41.74	3.76
BANet [25]	ResT-Lite	52.87	86.06	70.11	62.72	60.32	61.19	43.55	4.68
CG-Swin [52]	Swin-S	53.65	87.52	72.91	66.33	64.37	62.91	45.53	4.31
LWN-A-F [46]	EffNet-B1	52.31	87.39	71.72	65.44	62.24	61.72	44.41	4.93
BiSeNet [6]	R50	50.31	84.39	69.32	63.01	61.72	60.16	43.22	3.13
ShelfNet [53]	R50	56.08	88.18	70.17	72.55	65.25	61.17	43.51	4.01
DeepLabV3+ [12]	R50	56.96	89.44	71.22	75.52	**67.91**	62.46	45.47	5.10
HBSeNet (Our)	R50	**57.32**	**89.67**	**74.27**	**75.68**	67.18	**63.96**	**46.56**	**5.69**

The values are bolded to emphasize their superiority in the measurements.

TABLE V
CLASS-WISE PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART MODELS ON THE ISPRS VAIHINGEN

Method	Backbone	Imp. Surface	Building	Low Veg.	Tree	Car
DeepLabV3+ [12]	R101	84.82	90.71	71.32	78.41	77.12
DANet [23]	R101	86.32	91.21	71.62	79.92	76.74
GFFNet [54]	HRNetV2-W48	87.38	92.11	71.72	80.67	77.24
RSSFormer [55]	HRNetV2-W48	86.74	91.82	71.19	80.41	77.53
STDSNet [31]	Swin-B	86.09	**92.18**	72.82	81.17	76.57
SwinCNN [14]	Swin-B	85.45	91.55	72.64	80.71	75.18
ST-UNet [30]	R101 + Swin-B	85.24	91.56	72.33	80.92	74.38
HBSeNet (Our)	R101	**87.76**	91.98	**73.23**	**81.18**	**77.62**

The values are bolded to emphasize their superiority in the measurements.

TABLE VI
CLASS-WISE PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART MODELS ON THE ISPRS POTSDAM

Method	Backbone	Imp. Surface	Building	Low Veg.	Tree	Car
DeepLabV3+ [12]	R101	83.12	91.91	72.24	73.52	81.77
DANet [23]	R101	84.11	90.99	70.76	80.11	73.14
GFFNet [54]	HRNetV2-W48	84.71	90.10	72.92	76.63	84.72
RSSFormer [55]	HRNetV2-W48	83.33	91.61	73.24	78.31	82.55
STDSNet [31]	Swin-B	84.90	**92.23**	74.09	76.22	84.21
SwinCNN [14]	Swin-B	84.15	91.89	74.21	75.11	83.45
ST-UNet [30]	R101 + Swin-B	84.09	91.73	73.82	75.11	83.45
HBSeNet (Our)	R101	**85.51**	91.92	**74.25**	**80.98**	**85.22**

The values are bolded to emphasize their superiority in the measurements.

ISPRS Potsdam: Table VI compares the performance of various semantic segmentation methods on the ISPRS Potsdam dataset via individual mean IoU class metrics. Similar to the Vaihingen dataset analysis, we can gain valuable insights into the effectiveness of HBSeNet by dissecting these results. The proposed model, HBSeNet, achieves the highest mean IoU of 85.51% on *impervious surfaces* among all the compared methods. This suggests that HBSeNet is able to effectively segment surfaces in the Potsdam dataset. It outperforms STDSNet, as the second-best mean IoU model, by a small margin. Moreover, HBSeNet also achieves the best mean IoU for the other three classes, including *trees* with 80.98%, *cars* with 85.22%, and *low vegetation* 74.25%. However, for the *buildings* class, HBSeNet performs competitively with other methods, just achieving the second-best result with 91.92% mean IoU. Overall, the comparison results show that HBSeNet is a competitive semantic segmentation method that achieves the state-of-the-art performance on the ISPRS Potsdam dataset. It achieves the highest overall mean IoU and the best mean IoU for four out of the five classes. These results suggest that HBSeNet is a promising method for semantic segmentation tasks in aerial imagery.

Visualization Analysis: Having established HBSeNet's strong performance through quantitative metrics, we delve into a visual exploration of its segmentation capabilities. We compare the segmentation outputs of HBSeNet with state-of-the-art models that share similar complexities, as reflected in the previous complexity analysis. For the UAVid dataset, we compare HBSeNet with BiSeNet and DeepLabV3+ as utilizing the ResNet-50 backbone. Similarly, for the remaining two datasets, we present segmentation masks generated by ST-UNet and SwinCNN compared with our ResNet-101-based HBSeNet. These comparisons are visually depicted in Fig. 5. Each dataset showcases two images: the ground truth and the segmentation masks produced by the compared models. While existing methods like BiSeNet and DeepLabV3+ achieve object segmentation, they face challenges in accurately delineating boundaries, especially

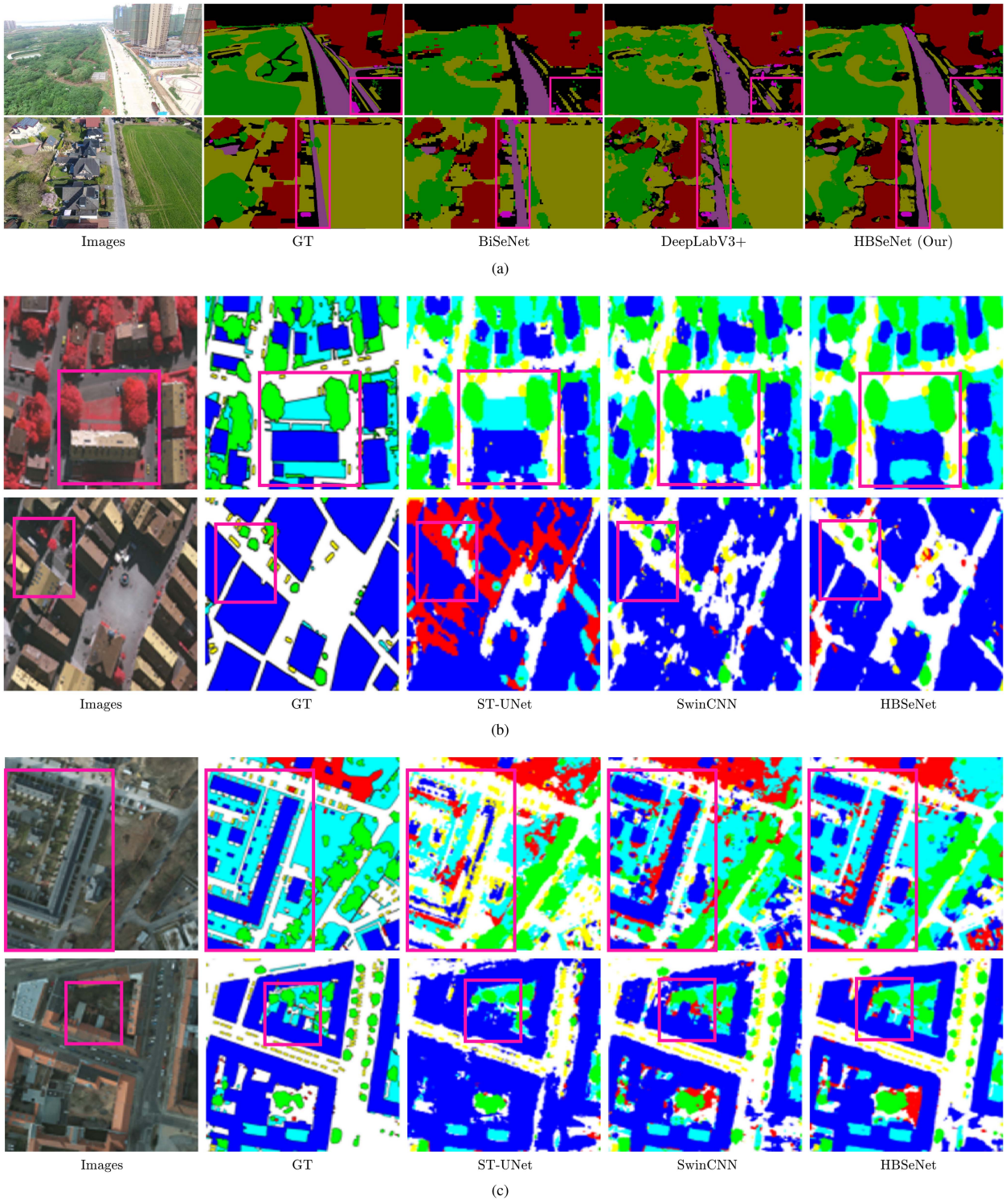


Fig. 5. Visualization results (as segmentation masks) of HBSeNet and other state-of-the-art models based on different datasets. (a) UAVid. (b) ISPRS Vaihingen. (c) ISPRS Potsdam. The pink rectangles are used to emphasize the regions that show a clear improvement in segmentation accuracy. GT: ground truth.

for closely positioned objects. BiSeNet prioritizes speed over accuracy, leading to less detailed segmentation masks on the UAVid dataset, often resulting in misclassifications of small regions. DeepLabV3+ exhibits better segmentation on UAVid but struggles to maintain object coherence and capture finer boundary details in other datasets. Even cutting-edge models like SwinCNN and ST-UNet, which combine convolutional layers with transformer self-attention, encounter difficulties with fine-grained boundaries between high-resolution objects. This is evident in the pink rectangular areas in Fig. 5, where these models exhibit lower accuracy. Interestingly, HBSeNet not only outperforms the compared models in segmentation accuracy but

also demonstrates superior boundary delineation, particularly in the regions highlighted by pink rectangles. However, there is still room for improvement. In some specific cases with small-sized object classes, further refinement is needed to achieve even sharper segmentation mask borders.

Complexity analysis: Another crucial aspect is the number of parameters. On the one hand, for the UAVid dataset, HBSeNet achieves this superior efficiency with a relatively small model size with approximately 17.5 M trainable parameters. This is significantly smaller if compared to other deep models like BiSeNet with 49.0 M and DeepLabV3+ with 43.9 M parameters. This translates to opportunities for faster processing speed, making HBSeNet more suitable for real-time applications in domains like autonomous vehicles or drone-based monitoring where fast and accurate image understanding is essential. On the other hand, for the ISPRS Vaihingen dataset, compared to powerful models like STDSNet and SwinCNN having 130.13 M and 235.77 M parameters, respectively, HBSeNet with 36.5 M parameters can maintain a superior efficiency with around 3 times smaller of model size while achieving exceptional accuracy. Similar to Vaihingen, HBSeNet outperforms other methods in all metrics while presenting a smaller model size if compared to STDSNet and SwinCNN having 130.13 M and 235.77 M parameters. This reinforces HBSeNet's capability to achieve high accuracy with a cost-efficient architecture. A key factor contributing to HBSeNet's remarkable ability to achieve high accuracy with a small number of weights lies in our optimization of the ResNet backbone. Traditional ResNet architectures often include a final stage with several convolutional layers and a global average pooling layer. However, through careful analysis, we identified that this final stage, while contributing to the overall depth of the network, might not be strictly necessary for effective semantic segmentation on the remote sensing datasets.

Insightful comparative analysis: Compared with the proposed HBSeNet model, existing works have several pros and cons. DeepLabV3-AFS [13] and PSPNet-AFS [13] utilize AFS mechanisms to enhance feature selection at specific layers. However, they often fail to capture long-range dependencies and global context, which are crucial for accurate semantic segmentation in RSIs. The CG-Swin model [53], which employs Swin transformers, is good at capturing local features but struggles to integrate global contextual information as effectively as models like HBSeNet that incorporate dedicated context paths. GFFNet [55], focusing on global feature fusion, may lead to the loss of local details necessary for fine-grained segmentation tasks and can be computationally intensive. In [56], RSSFormer emphasizes foreground saliency enhancement, potentially resulting in suboptimal performance in accurately segmenting less prominent but equally important background features. The dual-stream Swin transformer in STDSNet [32] requires high resources, and managing dual streams effectively is challenging, possibly leading to comprehensive performance if not properly tuned. Based on the combination of transformers and CNNs, SwinCNN [14] significantly increases model complexity and computational requirements, making it less efficient compared to more streamlined architectures like HBSeNet. In addition, the hybrid nature of SwinCNN poses challenges in balancing different types of layers and their respective hyperparameters. Although Swin transformers in ST-UNet [31] improve local feature extraction, they may not effectively capture broader contextual information without additional mechanisms, and integrating Swin transformers into the UNet architecture considerably increases computational cost and memory, thus limiting its scalability. In contrast, the densely connected sequential architecture of HBSeNet ensures efficient information flow between layers, enhancing feature reuse and improving segmentation accuracy. By integrating a spatial path with atrous convolutions and a well-organized context path, HBSeNet achieves a balance between local detail preservation and global context understanding. The streamlined architecture of HBSeNet maximizes information extraction while maintaining computational efficiency, making it suitable for real-time applications. Furthermore, the exploitation of effective auxiliary modules strengthens its contextual learning capability, addressing the shortcomings of existing models that may not adequately capture context.

V. Conclusion

In this work, we presented a novel semantic segmentation model with outstanding performance in the field of remote sensing. To achieve this, we proposed several enhancements to the model: constructing a bilateral network by combining a context path and a spatial path to capture both contextual and spatial features of the images simultaneously; introducing performance-improving modules, such as an SAC module for the spatial path, an ISM, and a bridge for the context path to minimize information loss during propagation within the network as well as optimize the number of model parameters. To demonstrate the robustness and reliability of the model, we conducted extensive simulation evaluations using different widely adopted datasets, including UAVid, ISPRS Vaihingen, and ISPRS Potsdam, with varying image resolutions and hardware deployments. The results reveal that our model achieved a moderate segmentation speed, but appreciably excels in accuracy when compared to other state-of-the-art models.

However, it is important to acknowledge some drawbacks of HBSeNet. The accuracy of HBSeNet for small objects remains lower than desired, potentially due to limitations in capturing sufficient meaningful information of small objects besides the receptive field may not be large enough to capture all relevant features. In addition, while the auxiliary modules contribute to performance, they introduce computational complexity, resulting in slower training and processing speeds to be inappropriate for real-time applications. Future research directions could explore techniques for small object segmentation (such as multiscale attention mechanism and data augmentation) and investigate methods (such as network pruning, knowledge distillation, and depthwise separable convolution) to optimize the network architecture for faster processing while maintaining accuracy.

REFERENCES

[1] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 2097–2104.

[2] A. Radman, N. Zainal, and S. A. Suandi, "Automated segmentation of iris images acquired in an unconstrained environment using HOG-SVM and growcut," *Digit. Signal Prog.*, vol. 64, pp. 60–70, Feb. 2017.

[3] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.

[4] 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, Feb. 2022, Art. no. 5400314.

[5] 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.

[6] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[7] P. Nie, X. Cheng, Z. Song, M. Mao, T. Wang, and L. Meng, "Rethinking BiSeNet: A lightweight network for urban water extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 4203910.

[8] L. Teng et al., "FLPK-BiSeNet: Federated learning based on priori knowledge and bilateral segmentation network for image edge extraction," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1529–1542, Jun. 2023.

[9] 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*, Munich, Germany, 2015, pp. 234–241.

[10] 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.

[11] 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.

[12] 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.*, Munich, Germany, 2018, pp. 801–818.

[13] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2021, Art. no. 8006705.

[14] 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, Jan. 2022, Art. no. 4408820.

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

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6230–6239.

[17] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet: A nested U-Net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Dec. Support*, Granada, Spain, 2018, pp. 3–11.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2117–2125.

[19] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 552–568.

[20] R. P. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: Exploring context and detail for semantic segmentation in real-time," in *Proc. BMVC*, 2018, pp. 1–11.

[21] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, St Petersburg, FL, USA, 2017, pp. 1–4.

[22] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 405–420.

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

[24] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107404.

[25] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, pp. 1–20, Aug. 2021.

[26] T. Xie, K. Wang, R. Li, X. Tang, and L. Zhao, "PANet: A pixel-level attention network for 6D pose estimation with embedding vector features," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1840–1847, Apr. 2022.

[27] H. Shuai, X. Xu, and Q. Liu, "Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4973–4984, Apr. 2021.

[28] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[29] O. Oktay et al., "RAU-Net: U-Net Model Based on Residual and Attention for Kidney and Kidney Tumor Segmentation," *IEEE Int. Conf. Consum. Electron. Comput. Eng.*, 2021, pp. 353–356, doi: 10.1109/ICCECE51280.2021.9342530.

[30] S. Seong and J. Choi, "Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3087.

[31] 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, Jan. 2022, Art. no. 4408715.

[32] X. Zhou, L. Zhou, S. Gong, S. Zhong, W. Yan, and Y. Huang, "Swin transformer embedding dual-stream for semantic segmentation of remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 175–189, Oct. 2023.

[33] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3065. [Online]. Available: https://www.mdpi.com/2072-4292/13/16/3065

[34] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS-J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021.

[35] Z. Li et al., "Cascaded multiscale structure with self-smoothing atrous convolution for semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2021, Art. no. 5605713.

[36] G.-V. Nguyen, C. V. Phan, and T. Huynh-The, "Accurate spectrum sensing with improved DeepLabV3 for 5G-LTE signals identification," in *Proc. 12th Int. Symp. Inf. Commun. Technol.*, Ho Chi Minh, 2023, pp. 221–227.

[37] Y. Qiu, Y. Liu, Y. Chen, J. Zhang, J. Zhu, and J. Xu, "A2SPPNet: Attentive atrous spatial pyramid pooling network for salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 1991–2006, Jan. 2022.

[38] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," May 2015, *arXiv:1505.00853*. [Online]. Available: https://api.semanticscholar.org/CorpusID:14083350

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

[40] S.-T. Tran, M.-H. Nguyen, H.-P. Dang, and T.-T. Nguyen, "Automatic polyp segmentation using modified recurrent residual unet network," *IEEE Access*, vol. 10, pp. 65951–65961, 2022.

[41] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[42] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

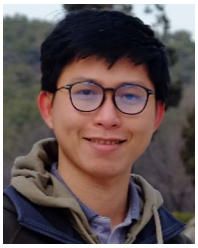
[43] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS-J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020.

[44] "ISPRS. 2D semantic labeling—vaihingen data," Accessed: Jan. 2024. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[45] "ISPRS. 2D semantic labeling contest—potsdam," Accessed: Jan. 2024. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[46] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2021, Art. no. 5603018.

[47] S. Liu, J. Cheng, L. Liang, H. Bai, and W. Dang, "Light-weight semantic segmentation network for UAV remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8287–8296, Aug. 2021.
[48] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607713.
[49] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.
[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 249–256.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.
[52] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?," in *Proc. Brit. Mach. Vis. Conf.*, Bristol, 2013, pp. 32.1–32.11.
[53] X. Meng, Y. Yang, L. Wang, T. Wang, R. Li, and C. Zhang, "Class-guided Swin transformer for semantic segmentation of remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 6517505.
[54] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "Shelfnet for fast semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Seoul, South Korea, 2019, pp. 847–856.
[55] Y. Cao, C. Huo, S. Xiang, and C. Pan, "GFFNet: Global feature fusion network for semantic segmentation of large-scale remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4222–4234, Jan. 2024.
[56] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 1052–1064, Jan. 2023.



Thien Huynh-The (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Kyung Hee University (KHU), Seoul, South Korea, in 2018.

From March 2018 to August 2018, he was a Postdoctoral Researcher with Ubiquitous Computing Laboratory, KHU. From 2018 to 2022, he was a Postdoctoral Researcher with ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi, South Korea. He is currently a Lecturer with the Department of Computer and Communications Engineering, Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam. His current research interests include digital image processing, radio signal processing, computer vision, wireless communications, IoT applications, machine learning, and deep learning.

Dr. Huynh-The was the recipient of Golden Globe Award 2020 for Vietnamese Young Scientist and IEEE ATC Best Paper Award in 2023. He is currently an Editor of IEEE COMMUNICATIONS LETTERS. He was a recipient of the Superior Thesis Prize awarded by KHU.

Son Ngoc Truong received the B.S. and M.S. degrees in electronics engineering from the HCMC University of Technical Education, Ho Chi Minh City, Vietnam, in 2006 and 2011, respectively, and the Ph.D. degree in electronics engineering from Kookmin University, Seoul, South Korea, in 2016.

He was a Research Assistant with Kookmin University, Seoul, South Korea, from 2016 to 2017. He is currently a Lecturer with the HCMC University of Technical Education. His research interests include the circuits and architectures with memristor device, memristor crossbar-based neuromorphic computing system, brain-inspired system, artificial intelligence, and deep learning.

Gia-Vuong Nguyen received the B.S. degree in embedded systems and Internet of Things from the Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam, in 2024.

His current research interests include digital image processing, radio signal processing, computer vision, machine learning, and deep learning.

Mr. Nguyen was a recipient of the REV-ECIT Best Paper Award in 2023.