

# AerialFormer: Multi-resolution Transformer for Aerial Image Segmentation

Kashu Yamazaki\*, *Member, IEEE*, Taisei Hanyu\*, Minh Tran, Adrian Garcia, Anh Tran, Roy McCann, Haitao Liao, Chase Rainwater, Meredith Adkins, Andrew Molthan, Jackson Cothren, and Ngan Le, *Member, IEEE*

**Abstract**—Aerial Image Segmentation is a top-down perspective semantic segmentation and has several challenging characteristics such as strong imbalance in the foreground-background distribution, complex background, intra-class heterogeneity, inter-class homogeneity, and tiny objects. To handle these problems, we inherit the advantages of Transformers and propose AerialFormer, which unifies Transformers at the contracting path with lightweight Multi-Dilated Convolutional Neural Networks (MD-CNNs) at the expanding path. Our AerialFormer is designed as a hierarchical structure, in which Transformer encoder outputs multi-scale features and MD-CNNs decoder aggregates information from the multi-scales. Thus, it takes both local and global contexts into consideration to render powerful representations and high-resolution segmentation. We have benchmarked AerialFormer on three common datasets including iSAID, LoveDA, and Potsdam. Comprehensive experiments and extensive ablation studies show that our proposed AerialFormer outperforms previous state-of-the-art methods with remarkable performance. Our source code will be publicly available upon acceptance.

**Index Terms**—Aerial Image, Segmentation, Transformers, Dilated Convolution

## I. INTRODUCTION

The use of aerial images provides a view of the Earth from above, which consists of various geospatial objects such as cars, buildings, airplanes, ships, etc., and allows us to regularly monitor the large areas of the planet. The recent advances in sensor technology have opened up the potential use of those remote sensing (RS) images in broader applications thanks to the ability to capture high spatial resolution (HSR) images with abundant spatial details and rich potential semantic content. Aerial image segmentation (AIS) is a particular semantic segmentation task that aims to assign a semantic category to

K. Yamazaki, T. Hanyu, M. Tran, A. Garcia, A. Tran, and N. Le are with Artificial Intelligence and Computer Vision (AICV) Lab, University of Arkansas, 1 University of Arkansas Fayetteville, AR 72701 USA e-mail: {kyamazak, thanyu, minht, ad84, anhtran, thile}@uark.edu

J. Cothren is with Department of Geoscience, Center for Advanced Spatial Technologies, University of Arkansas, 227 N Harmon Ave, Fayetteville, AR 72701, rmccann@uark.edu

R. McCann is with Department of Electrical Engineering, University of Arkansas, 1 University of Arkansas Fayetteville, AR 72701, rmcann@uark.edu

H. Liao and C.Rainwater are with Department of Industrial Engineering, University of Arkansas, 1 University of Arkansas Fayetteville, AR 72701, {liao, cer}@uark.edu

M.Adkins is with Institute for Integrative and Innovative Research, University of Arkansas, 1 University of Arkansas Fayetteville, AR 72701, mmcree@uark.edu

A. Molthan is with NASA Marshall Space Flight Center, Huntsville, AL 35808

\*These authors contributed equally to this work

each image pixel. Thus, AIS plays an important role in the understanding and analysis of remote sensing data, offering both semantic and localization cues for the targets of interest. Understanding and analyzing these objects from top-down perspective offered by remote sensing (RS) imagery is crucial for urban monitoring and planning. This understanding finds utility in numerous practical urban-related applications, such as disaster monitoring [61], agricultural planning [86], street view extraction [20], [63], land change [54], [60], [87], land cover [80], climate change [58], deforested regions [2], etc. However, due to the large footprint of aerial images and limited sensor bandwidth, several challenging characteristics are needed to be investigated. Some critical issues are intra-class heterogeneity (i.e., objects of the same category may be shown in various shapes, texture, colors, scales, and structures), inter-class homogeneity (i.e., objects of the different classes may share the same visual properties) [79], large diversity of resolution and orientation [85], dense and tiny objects [62], background and foreground imbalance [106], high background complexity [106]. As shown in Figure1, the ratio between foreground and background is 2.86%/97.14%; the inter-class homogeneity is presented by Tennis Court and Basketball Court, which share similar appearance; intra-class heterogeneity is presented by Tennis Court which are shown in various appearances.

With the success of deep learning (DL) techniques in extracting rich contextual feature e.g., VGG [64], ResNet [24], InceptionNet [70], [71], MobileNet [29], etc., various semantic segmentation approaches based on those backbones have been proposed such as Unet [59], PSPNet [102], DeepLabV3+ [11], Segmener [65], or UperNet [88]. Most of the existing image segmentation methods are originally proposed for other use cases such as self-driving vehicle [55] and medical imaging [38]. Thus, they do not perform optimally on AIS, resulting in limited accuracy on tiny objects and weak boundary objects. To alleviate those limitations, it is essential to obtain strong semantic representations at both the local level (e.g., boundary) and the global context level (e.g., relationship between objects/classes).

Recently, the great success of Transformer [74] in natural language processing (NLP) has inspired numerous tasks in computer vision including semantic segmentation. Following the Transformer design in NLP, [18] split an image into multiple linearly embedded patches and feed them into a standard Transformer and proposes vision Transformer (ViT) for image classification. Later, [105] adopts ViT as a backbone and proposes SETR to demonstrate the feasibility of using Trans-



Fig. 1: Examples of challenging characteristics in remote sensing image segmentation. (left) the distribution of foreground and background are highly imbalanced. (right-top) objects in some classes are dense and small such that they are hardly identifiable. (right-bottom) within a class, there are large diversity in appearance: intra-class heterogeneity (purple); some different classes share the similar appearance: inter-class homogeneity (pink). The image is from iSAID dataset. Best viewed with color and zoom in.

formers in semantic segmentation. Although the Transformer-based encoder has various benefits, its computational complexity is considerably greater than that of the CNN-based encoder because of its self-attention mechanism with a squared complexity. As a result, it is challenging to process high-resolution images using Transformer-based models. To reduce the computational complexity, some Transformer models such as Swin Transformer [50] propose shifted windows to bring greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. Despite the great potential in various computer vision tasks owing to their strong capability to model long-range dependency using the self-attention mechanism, vision transformers are limited in modeling local visual structures and scale-invariant representations in the context of dense prediction tasks. Unlike vision transformers, Convolution Neural Networks (CNNs) are based on convolution to compute local correlation among neighbor pixels. Consequently, CNNs are good at extracting local features, and scale-invariance and still serve as prevalent backbones in vision tasks. Generally, CNNs and vision transformers focus on different aspects. On one hand, CNNs adopt convolutions allowing CNNs to preferably extract local contextual information and translation invariance. However, this property leads to locality and strong inductive biases. On the other hand, vision transformers adopt self-attention mechanisms for perfectly extracting global and long-range dependencies, but do not capture locality and trans-

lation invariance very well. According to the above-mentioned analyses, we believe CNNs and vision transformers are naturally complementary to each other. Thus, combining these two kinds of CNNs and vision transformers can overcome the weaknesses of two models and strengthen their advantages simultaneously.

In an effort to mitigate the multiple aforementioned challenging characteristics involved in aerial image segmentation, as per our prior analysis, we draw inspiration from the strengths and success of CNNs for exploring the advantages of introducing local visual structures, as well as from the scale-invariant representation in vision transformers. In this paper, we particularly propose AerialFormer, a deep learning network with Swin Transformer encoder and CNNs decoder to efficiently localize objects in aerial images from satellite. Furthermore, we present a new approach that utilizes a convolutional stem network to generate fine feature maps for tiny objects in the encoder. We also introduce a Multi-Dilated Convolution (MDC) block at decoder to effectively extract features while avoiding excessive computational complexity due to its fully convolutional design.

Our contribution is summarized as follows:

- Provide a comprehensive literature review on aerial images segmentation.
- Analyze the current challenging characteristics of aerial images segmentation.
- Propose an effective computation model to leverage the

merits of both vision transformers to capture long-range dependency and CNNs to extract local representation and scale-invariance.

- Propose an CNN stem network to alleviate the potential drawback in using Transformer backbone for dense prediction task.
- Conduct an extensive experiment on the widely recognized three datasets: iSAID, LoveDA, and Potsdam.

## II. RELATED WORKS

Generally, image segmentation is categorized into three tasks: instance segmentation, semantic segmentation, and panoptic segmentation. Each of these tasks is distinguished based on their respective semantic considerations. In this work, we focus on the second task of semantic segmentation, a form of dense prediction task where each pixel from an image is associated with a class label. Different from instance segmentation, it does not distinguish each individual instance of the same object class. The goal of semantic segmentation is to divide an image into several visually meaningful or interesting areas for visual understanding according to semantic information. Semantic segmentation plays an important role in a broad range of applications, e.g., scene understanding, medical image analysis, autonomous driving, video surveillance, robot perception, satellite image segmentation, agriculture analysis, etc. We start this section with reviewing on DL-based semantic image segmentation and the advancements made in Computer Vision with Transformers. Then, we turn our focus to a review of aerial image segmentation using deep neural networks.

### A. DL-based Image Segmentation

Convolutional Neural Networks (CNNs) are widely regarded as the de-facto standard for various tasks within the field of computer vision. Long et al. [51] shows that Fully Convolutional neural (FCNs) can be used to segment images without fully connected layers and it has become one of the principal networks for semantic segmentation. With the advancements brought by the FCNs in semantic segmentation, many improvements have been made by designing the network deeper, wider, or more effective. This includes enlarging the receptive field [9], [11], [14], [27], [40], [95], strengthening context cues [23], [40], [41], [30], [31], [35], [36], [98]–[100] leveraging boundary information [5], [16], [38], [39], [48], [104], and incorporating neural attention [21], [22], [32], [33], [44], [66], [83], [84], [103]. Recently, a new paradigm of neural network architecture that does not employ any convolutions and mainly relies on self-attention mechanism, called Transformers, has become rapidly adopted to CV tasks [6], [42], [49] and achieved promising performance. The core idea behind transformer architecture [74] is the self-attention mechanism to capture long-range relationships. In addition, Transformers can be easily parallelized, facilitating training on larger datasets. Vision Transformer (ViT) [18] is considered one of the first works applied the standard Transformer to vision tasks. Unlike the CNNs structure, the ViT processes the 2D image as a 1D sequence of image

patches. Thanks to the powerful sequence-to-sequence modeling ability of the Transformer, ViT demonstrates superior characterization of extracting global context especially in the lower level features compared to the CNN counterparts. Recent advancements in Transformers over the past few years have demonstrated their effectiveness as backbone networks for visual tasks, surpassing the performance of numerous CNN-based models trained on large datasets. Transformer-based image segmentation approaches [12], [13], [65], [73], [89], [105] inherit the flexibility of Transformers in modeling long-range dependencies, yielding remarkable results. Transformers have been applied with notable success across a variety of computer vision tasks. These include image recognition [18], [72] object detection [6], [68], [108], image segmentation [73], [96], [105], action localization [75], [76], and video captioning [93], [94], thereby showcasing their capability to augment global information.

### B. Aerial Image Segmentation

Computer vision techniques have long been employed for the analysis of satellite images. Historically, satellite images had a lower resolution and the goal of segmentation was primarily to identify boundaries like straight lines and curves in aerial pictures. However, modern satellite imagery possesses significantly higher resolution, and consequently, the demands of segmentation tasks have substantially increased, which include the segmentation of tiny objects, objects with substantial scale variation, and entities exhibiting visual ambiguities. To this end, FCNs and their variants have become the mainstream solution for aerial image segmentation and led to state-of-the-art performance across numerous datasets [11], [28], [47], [53], [67], [92]. To capture contextual interrelations among pixels in remote sensing images, techniques from natural language processing have also been incorporated into aerial image segmentation [97]. By imitating the channel attention mechanism [32], S-RA-FCN [56] designs a spatial relation module to capture global spatial relations, and [57] introduces HMANet with spatial interaction while balancing between the size of receptive field and the computation cost. In HMANet, a region shuffle attention module is proposed to improve the efficiency of the self-attention mechanism by reducing redundant features and forming region-wise representations. In recent years, the advancements in transformer-based networks, which leverage self-attention mechanisms to achieve receptive fields as large as the entire image, have sparked increased interest in their applications. Consequently, there has been a surge in research studies [8], [69], [77], [81], [82], [89], [91] that have integrated Transformers into remote sensing applications. For instance, RSSFormer [91] proposed the Adaptive Transformer Fusion Module to mitigate background noise and enhance object saliency during the fusion of multi-scale features. Some other works [69], [81] adopt Transformers as their backbone.

In this paper, we introduce AerialFormer, an innovative fusion of a Transformer encoder and a multi-dilated CNNs decoder. While Transformer-based approaches excel at modeling long-range dependencies, they face challenges in capturing local details and struggles in handling tiny objects. Thus,

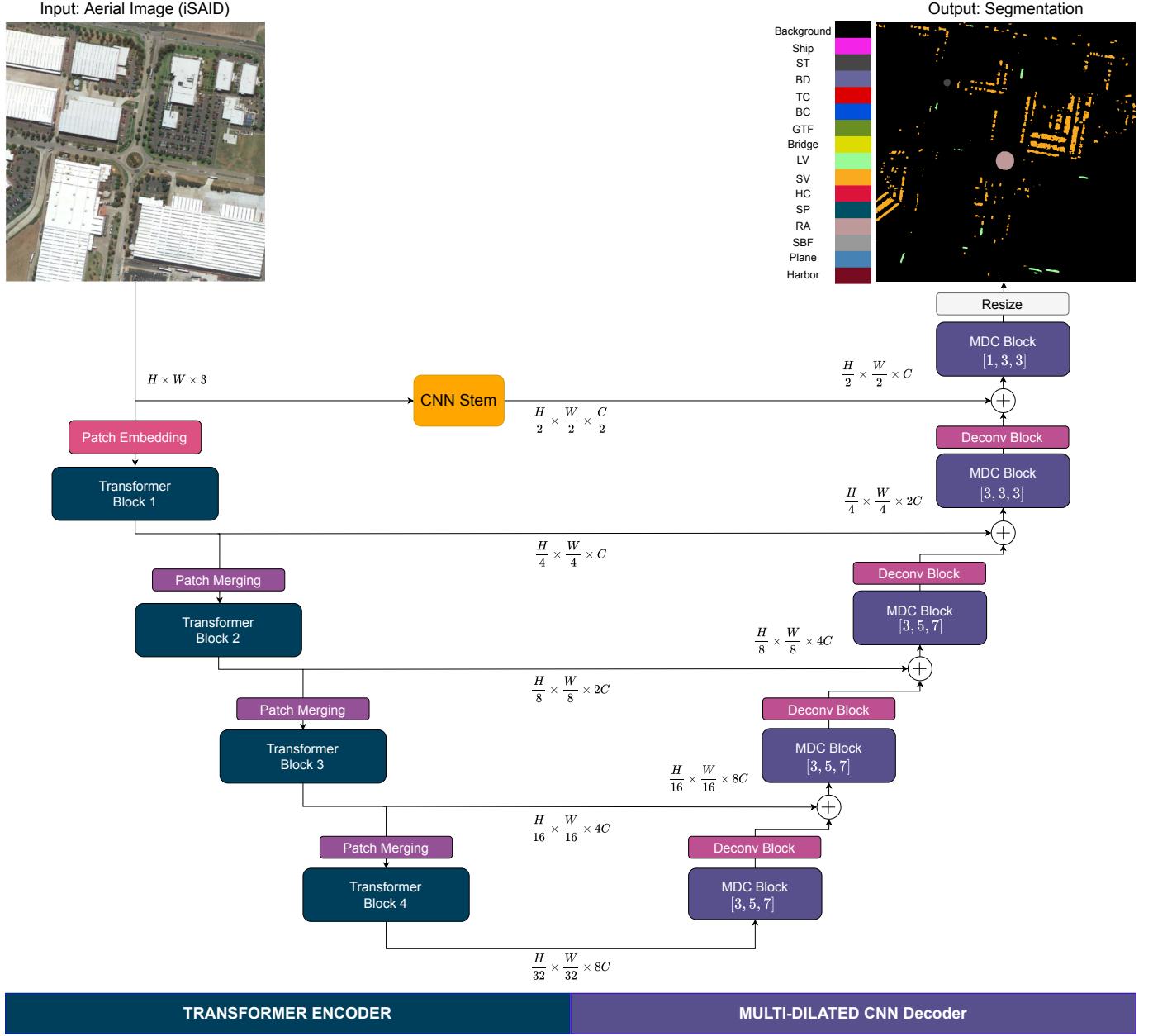


Fig. 2: Overall network architecture of our proposed AerialFormer which consists of three components i.e., Transformer Encoder, CNNs Stem, and Multi-Dilated CNNs Decoder.

our AerialFormer incorporates multi-dilated convolution to capture long-range dependence without increasing the memory footprint at the decoder. Our novel AerialFormer approach combines the strengths of a Transformer encoder and a multi-dilated CNNs decoder, aided by skip connections, to capture both local context and long-range dependencies effectively in aerial image segmentation.

### III. METHODS

#### A. Network Overview

An overview of our AerialFormer architecture is presented in Figure 2. The architecture design is fundamentally rooted in the renowned Unet structure for semantic segmentation [59],

characterized by its encoder-decoder network with use of skip-connections between the matched blocks with identical spatial resolution on both encoder and decoder sides. The composition of our model is threefold: a *Transformer Encoder*, a *CNNs Stem*, and a *Multi-Dilated CNNs Decoder*. The Transformer Encoder is designed as a sequence of  $s$  stages of Transformer Encoder blocks ( $s$  is set as 4 in our architecture) aiming to extract long-range context representation. The CNNs Stem aims to preserve low-level information at high resolution. The latter, Multi-Dilated CNNs (MDC) Decoder consists of  $s+1$  MDC blocks with skip connections to obtain information from multiple scales and wide context. We will detail these components in the following subsections.

Given a high-resolution aerial image, we first overlap par-

tition it into a set of sub-images sized  $H \times W \times 3$ , where 3 corresponds to three color channels. Each sub-image is then fed to the AerialFormer and the output is the segmentation of  $H \times W$ .

### B. Transformer Encoder

The Transformer Encoder starts by processing an input image size of  $H \times W \times 3$ , which is tokenized by the *Patch Embedding layer*, which results in a feature map  $\frac{H}{p} \times \frac{W}{p} \times C$ . The feature map is then passed through a sequence of  $s = 4$  *Transformer Encoder Blocks* and produces multi-level outputs of different sizes at each block:  $\frac{H}{4} \times \frac{W}{4} \times C$ ,  $\frac{H}{8} \times \frac{W}{8} \times 2C$ ,  $\frac{H}{16} \times \frac{W}{16} \times 4C$ , and  $\frac{H}{32} \times \frac{W}{32} \times 8C$ . Each Transformer Encoder Block is followed by a *Patch Merging layer*, which reduces the spatial dimension by half before being passed to the next deeper Transformer Encoder Block.

1) *Patch Embedding*: The Transformer Encoder starts by taking an image  $H \times W \times 3$  as an input and dividing it into patches of size  $p \times p$  in a non-overlapping manner. Each patch is embedded into a vector in dimensional space of  $\mathbb{R}^C$  by a linear projection, which can be simplified as a single convolution operation with the kernel size of  $p \times p$  and the stride of  $p \times p$ . The Patch Embedding produces a feature maps of  $\frac{H}{p} \times \frac{W}{p} \times C$ . The patch size determines the spatial resolution of the input sequence of transformer, and therefore smaller patch size is favored for the dense prediction tasks including semantic segmentation. While ViT [18] is a commonly used vision transformer in computer vision, which processes  $16 \times 16$  patch and is able to capture wider range context, it may not be suitable for capturing detailed information. One of the most challenging aspects of aerial image segmentation is dealing with tiny objects. In the other hand, Swin Transformer [50], one of the transformer variants utilizes a smaller patch of  $4 \times 4$ . Thus, we adopt Swin Transformer [50] to implement Patch Embedding layer to better capture the detailed information of tiny objects in aerial image segmentation.

2) *Transformer Encoder Block*: In general, let  $\mathbf{x} \in R^{h \times w \times d}$  denote the input of a Transformer Encoder Block. Transformer Encoder Block processes the input data with a series of self-attention and feed-forward network with residual connection. To compensate the increase in computation because of the smaller patch size, Swin Transformer [50] utilizes a local self-attention instead of global self-attention. The global self-attention, used in standard Transformers, has a computational cost of  $\mathcal{O}(N^2 \cdot d)$  where  $N$  is the number of tokens (i.e.,  $N = h \times w$ ) and  $d$  is the representation dimension, which can be prohibitively expensive for large images and small patch size. Swin Transformer introduced window-based self-attention (WSA) that divides the image into non-overlapping windows and performs self-attention within each window. With WSA, the computational cost is linear to the number of tokens, i.e.,  $\mathcal{O}(M^2 \cdot N \cdot d)$  where  $M^2$  is the number of patches within a window and  $M^2 \ll N$ . In order to apply the WSA, an input  $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$  is partitioned into a group of local patches  $\mathbf{x}' \in \mathbb{R}^{\frac{h \times w}{M^2} \times M^2 \times d}$  and the first dimension  $\frac{h \times w}{M^2}$  is treated as a batch dimension, i.e., the network parameters are shared along the first dimension. Considering the multi-head attention operation with  $h$  heads, the feature dimension

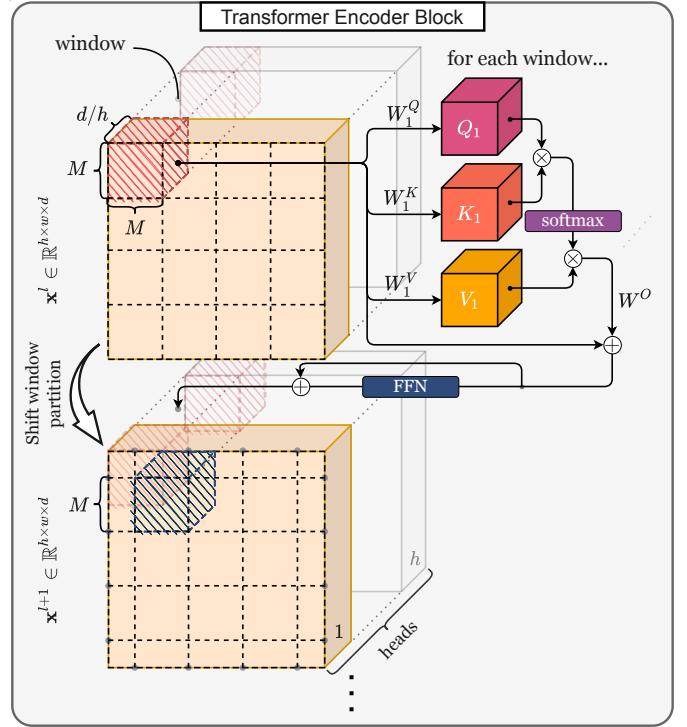


Fig. 3: An illustration of the Transformer Encoder Block.

$d$  is split into  $h$  identical blocks, i.e.,  $\mathbb{R}^{\frac{h \times w}{M^2} \times M^2 \times \frac{d}{h} \times h}$ . Then, we can formulate the WSA as:

$$\text{WSA}(\mathbf{x}') = [\text{head}_1; \dots; \text{head}_h]W^O \quad (1)$$

where  $[\cdot]$  denotes the channel wise concatenation of tensor,  $W^O \in \mathbb{R}^{d \times d}$  is the output projection weights, and each head  $\text{head}_i$  is calculated as:

$$\text{head}_i = \text{softmax} \left( \frac{Q_i K_i^\top}{\sqrt{d/h}} + B \right) V_i \quad (2)$$

where  $Q_i = \mathbf{x}'_i W_i^Q$ ,  $K_i = \mathbf{x}'_i W_i^K$ ,  $V_i = \mathbf{x}'_i W_i^V \in \mathbb{R}^{M^2 \times \frac{d}{h}}$  are the query, key and value tensors, which are created from the local window with  $M \times M$  patches with  $\frac{d}{h}$  feature dimensions by linearly projecting with learnable weights of  $W^Q$ ,  $W^K$ , and  $W^V \in \mathbb{R}^{\frac{d}{h} \times \frac{d}{h}}$ .  $B \in \mathbb{R}^{M^2 \times M^2}$  is the relative position bias [50] that introduces relative positional information to the model.

Because the WSA applies the self-attention on the local window, WSA alone cannot obtain a global context of the image. To alleviate this issue, Swin Transformer stacks Transformer blocks using WSA and alternates the window location by half of the window size to gradually build global context by integrating information from different windows. Specifically, the Swin Transformer block consists of a shifted WSA, followed by a 2-layer FFN with GELU activation function in between, which is formulated as:

$$\begin{aligned} \hat{\mathbf{x}}^l &= \mathbf{x}^l + \text{WSA}(\text{norm}(\mathbf{x}^l)) \\ \mathbf{x}^{l+1} &= \hat{\mathbf{x}}^l + \text{FFN}(\text{norm}(\hat{\mathbf{x}}^l)) \end{aligned} \quad (3)$$

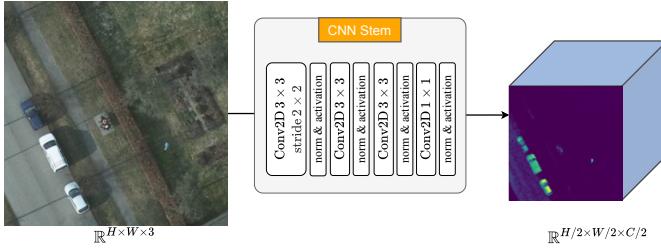


Fig. 4: An illustration of the CNN Stem. The Stem takes the input image and produces feature maps with half of the original spacial resolution.

where the *norm* indicates the LayerNorm [3] operation, FFN indicates the feed-forward network, and partitioning of the input  $\mathbf{x}$  is shifted by  $(\lfloor \frac{M}{2}, \frac{M}{2} \rfloor)$  from the regularly partitioned windows when layer  $l$  is even. This process is illustrated in Figure 3. For each Transformer Encoder Block, we denote the set of the total number of layers as  $\mathcal{L}_s$ .

3) *Patch Merging*: In order to generate a hierarchical representation, the spatial resolution of each Transformer Encoder Block is reduced by half through the Patch Merging layer. The Patch Merging layer takes a feature map size of  $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$  as an input. The layer first splits and gathers the feature in a checkerboard pattern, creating four sub-feature maps  $\mathbf{x}_1$  to  $\mathbf{x}_4$  with half of the spatial dimension of the original feature map, where  $\mathbf{x}_1$  contains pixels from 'black' squares in even rows,  $\mathbf{x}_2$  from 'white' squares in even rows,  $\mathbf{x}_3$  from 'black' squares in odd rows, and  $\mathbf{x}_4$  from 'white' squares in odd rows. Then these four feature maps are concatenated along the channel dimension, resulting in a tensor of size  $h/2 \times w/2 \times 4d$ . Finally, the linear projection is applied to reduce the channel dimension from  $4d$  to  $2d$ .

### C. CNN Stem

Although our Transformer encoder is favored in semantic segmentation by using smaller patch size, it may discard the fine-grained details that are especially important in aerial images, which contain tiny and dense objects. To this end, we propose a simple yet effective way to inject the low-level features of the input image to our decoder through a convolutional stem module. This module is expected to model the local spatial contexts of images parallel with the patch embedding layer. As shown in the Figure 4, our CNN Stem consists of four convolutional layers, each followed by BatchNorm [34] and GELU [26] activation layers. The first  $3 \times 3$  convolutional layer with stride of  $2 \times 2$  reduces the input spacial size into half and through the following three layers of convolution, we obtain local features for tiny and dense objects.

### D. Multi-Dilated CNNs Decoder

While local fine-grained feature is important for segmenting tiny objects, we want to consider the global context at the same time. In the decoder, we propose to use multiple dilated convolutional operations in parallel with different dilation

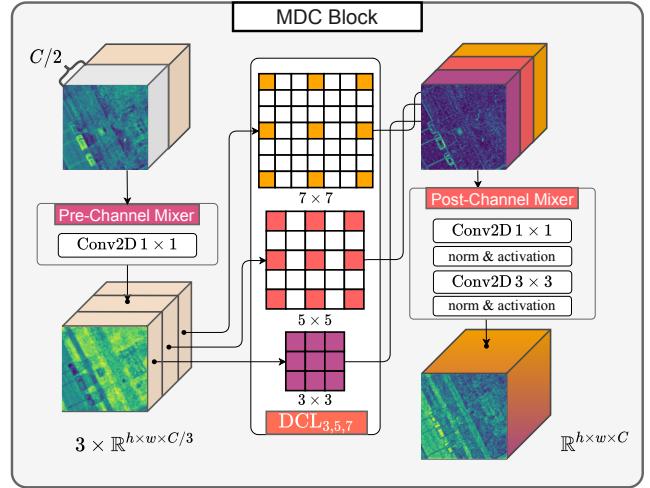


Fig. 5: An illustration of the MDC Block, which consists of Pre-Channel Mixer, DCL, and Post-Channel Mixer.

rates to obtain wider context for decoding without any additional parameters. The Multi-Dilated CNNs Decoder contains a sequence of Multi-Dilated CNNs (MDC) Block followed by Deconvolutional (Deconv) block, which are detailed as follows.

1) *MDC Block*: A MDC Block is defined by three params  $[r_1, r_2, r_3]$  corresponding to three receptive fields, and consists of three parts of Pre-Channel Mixer, Dilated Convolutional Layer (DCL) and Post-Channel Mixer.

The MDC Block starts by applying Pre-Channel Mixer to the input, which is the concatenation of previous MDC block's output and the skip connection from the mirrored encoder, in order to exchange the information in channel dimension. The channel mixing operation can be implemented with any operator that enforces the information exchange in channel dimension. Here, Pre-Channel Mixer is implemented as a point-wise convolution layer without any normalization or activation layer.

The DCL utilizes three convolutional kernels with different dilation rates of  $d_1, d_2$ , and  $d_3$ , which allows to obtain multi-scale receptive fields.

We can calculate the length of one side of a receptive field  $r$  of dilated convolution given a kernel size  $k$  and a dilation rate  $d$  as follows:

$$r_i = d_i(k - 1) + 1 \quad (4)$$

where the kernel size  $k$  is established as 3 for receptive fields that exceed  $3 \times 3$  in size, and as 1 for those receptive fields that are smaller.

We will denote the dilated convolutional operation with receptive field of  $r \times r$  as  $\text{Conv}_r(\cdot)$ . Then we can formulate our DCL as follows:

$$\text{DCL}_{r_1, r_2, r_3}(\mathbf{x}) = [\text{Conv}_{r_1}(\mathbf{x}_1); \text{Conv}_{r_2}(\mathbf{x}_2); \text{Conv}_{r_3}(\mathbf{x}_3)] \quad (5)$$

where  $\mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2; \mathbf{x}_3]$ , i.e., the tensor after the Pre-Channel Mixer  $\mathbf{x}$  is sliced into three sub-tensors with equivalent

TABLE I: Performance comparison on **iSAID** *valset* between our AerialFormer and other SOTA approaches. We report performance on mIoU and IoU for each category. The **bold** and *italic-underline* values in each column show the best and the second best performances.

Method	Year	mIoU ↑	IoU per category* ↑														
			Vehicles					Artifacts				Fields					
			LV	SV	Plane	HC	Ship	ST	Bridge	RA	Harbor	BD	TC	GTF	SBF	SP	BC
UNet [59]	2015	37.4	49.9	35.6	74.7	0.0	49.0	0.0	7.5	46.5	45.6	6.5	78.6	5.5	9.7	38.0	22.9
PSPNet [102]	2017	60.3	58.0	43.0	79.5	10.9	65.2	52.1	32.5	68.6	54.3	75.7	85.6	60.2	71.9	46.8	61.1
DeepLabV3 [10]	2017	59.0	54.8	33.7	75.8	31.3	59.7	50.5	32.9	66.0	45.7	77.0	84.2	59.6	72.1	44.7	57.9
DeepLabV3+ [11]	2018	61.4	61.9	46.7	82.1	0.0	66.2	71.5	37.5	63.1	56.9	73.1	87.2	56.2	73.8	46.6	59.8
HRNet [67]	2019	62.3	61.6	48.5	82.3	6.9	67.5	70.3	38.4	65.7	54.7	75.4	87.1	55.5	75.5	46.4	62.1
FarSeg [106]	2020	63.7	60.6	46.3	82.0	35.8	65.4	61.8	36.7	71.4	53.9	77.7	86.4	56.7	72.5	51.2	62.1
HMANet [57]	2021	62.6	59.7	50.3	83.8	32.6	65.4	70.9	29.0	62.9	51.9	74.7	88.7	54.6	70.2	51.4	60.5
PFNet [47]	2021	66.9	64.6	50.2	85.0	37.9	70.3	74.7	45.2	71.7	59.3	77.8	87.7	59.5	75.4	50.1	62.2
Segformer [89]	2021	65.6	64.7	51.3	85.1	40.3	70.8	73.9	40.8	60.9	56.9	74.6	87.9	58.9	75.0	51.2	59.1
FactSeg [53]	2022	64.8	62.7	49.5	84.1	<u>42.7</u>	68.3	56.8	36.3	69.4	55.7	78.4	88.9	54.6	73.6	51.5	64.9
BSNet [28]	2022	63.4	63.4	46.6	81.8	31.8	65.3	69.1	41.3	70.0	57.3	76.1	86.8	50.3	70.2	48.8	55.9
AANet [92]	2022	66.6	63.2	48.7	84.6	41.8	71.2	65.7	40.2	72.4	57.2	<u>80.5</u>	88.8	60.5	73.5	<u>52.3</u>	<u>65.4</u>
RSP-Swin-T [77]	2022	64.1	62.0	50.6	85.2	37.6	67.0	74.6	44.3	64.9	53.8	73.7	70.7	60.1	76.2	46.8	59.0
Ringmo [69]	2022	67.2	63.9	51.2	85.7	40.1	<u>73.5</u>	73.0	43.2	67.3	58.9	77.0	89.1	<u>63.0</u>	<b>78.5</b>	48.9	62.5
RSSFormer [91]	2023	65.9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
<b>AerialFormer-T</b>	—	67.5	<u>67.0</u>	52.6	86.1	42.0	68.6	<u>74.9</u>	<u>45.3</u>	<u>73.0</u>	58.2	77.5	88.8	57.5	75.1	50.5	63.4
<b>AerialFormer-S</b>	—	<u>68.4</u>	66.5	<u>53.6</u>	<b>86.5</b>	40.0	72.1	74.1	44.8	<b>74.0</b>	<b>60.9</b>	78.8	<u>89.2</u>	59.5	77.0	52.1	<b>66.5</b>
<b>AerialFormer-B</b>	—	<b>69.3</b>	<b>67.8</b>	<b>53.7</b>	<b>86.5</b>	<b>46.7</b>	<b>75.1</b>	<b>76.3</b>	<b>46.8</b>	66.1	<u>60.8</u>	<b>81.5</b>	<b>89.8</b>	<b>65.0</b>	<u>78.3</u>	<b>52.4</b>	62.4

channel length. As we split feature to process with DCL with three different spacial resolution, we applied a Post-Channel Mixer to exchange the information from the three convolutional layers. We implemented the Post-Channel Mixer with sequence of point-wise and  $3 \times 3$  convolution layer, each of which are followed by BatchNorm and ReLU activation layers. This lets us formulate the Multi-Dilated Convolution (MDC) block as follows. The entire operation for MDC Block is illustrated in Figure 5.

$$\text{MDC}(\mathbf{x}) = \text{PostMixer}(\text{DCL}_{r_1, r_2, r_3}(\text{PreMixer}(\mathbf{x}))) \quad (6)$$

where PreMixer refers to the Pre-Channel Mixer and PostMixer refers to the Post-Channel Mixer.

2) *Deconv Block*: The Deconv Block employs the transposed convolution layer, which serves to increase the spatial dimensions of the feature map by a factor of two, while concurrently decreasing the channel dimension by half. We also add the BatchNorm and ReLU activation layers after the transposed convolution operation.

#### E. Loss Function

We supervise the network with Cross Entropy Loss, which can be formulated as follows:

\*Categories in iSAID dataset: Large Vehicle (LV), Small Vehicle (SV), Plane, Helicopter (HC), Ship, Storage Tank (ST), Bridge, Roundabout (RA), Harbor, Baseball Diamond (BD), Tennis Court (TC), Ground Track Field (GTF), Soccerball Field (SBF), Swimming Pool (SP), and Basketball Court (BC).

$$\mathcal{L}_{CE} = - \sum_{i=1}^n t_i \log(p_i) \quad (7)$$

where  $t_i$  represents the ground truth and  $p_i$  is the softmax probability for the  $i^{th}$  class.

## IV. EXPERIMENTS

### A. Datasets

Our AerialFormer is benchmarked on three standard aerial imaging datasets, i.e., iSAID, Potsdam, and LoveDA as below. **iSAID**: iSAID dataset [85] is a large-scale and densely annotated aerial segmentation dataset that contains 655,451 instances of 2,806 high-resolution images for 15 classes (i.e., ship (Ship), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground field track (GTF), bridge (Bridge), large vehicle (LV), small vehicle (SV), helicopter (HC), swimming pool (SP), roundabout (RA), soccerball field (SBF), plane (Plane), and harbor (Harbor)). This dataset is challenging due to the presence of a large number of objects per image, limited appearance details, a variety of tiny objects, large-scale variations, and high-class imbalance. These images were collected from multiple sensors and platforms with multiple resolutions and image sizes ranging from  $800 \times 800$  pixels to  $4000 \times 13,000$  pixels. Follow the experiment setup [106], [47], the dataset is split into 1,411/458/937 images for train/val/test. We train the network on the trainset and benchmark on the valset. Each image is

TABLE II: Performance comparison on **Potsdam** *valset with clutter*. We report performance on mIoU, OA, mF1, and F1 score for each category. Note that both train and evaluation are done on the eroded dataset. The **bold** and *italic-underline* values in each column show the best and the second best performances.

<b>Method</b>	<b>Year</b>	<b>mIoU</b> $\uparrow$	<b>OA</b> $\uparrow$	<b>mF1</b> $\uparrow$	<b>F1 per category</b> $^*\uparrow$					
		<b>Imp. Surf.</b>	<b>Building</b>	<b>Low Veg.</b>	<b>Tree</b>	<b>Car</b>	<b>Clutter</b>			
FCN [51]	2015	64.2	—	75.9	87.6	91.6	77.8	84.6	73.5	40.3
PSPNet [102]	2017	77.1	90.1	85.6	92.6	96.2	86.2	88.0	95.3	55.4
DeeplabV3 [10]	2017	77.2	90.0	85.6	92.4	95.9	86.4	87.6	94.9	56.7
UPerNet [88]	2018	76.8	89.7	85.6	92.5	95.5	85.5	87.5	94.9	58.0
DeepLabV3+ [11]	2018	77.1	90.1	85.6	92.6	96.4	86.3	87.8	95.4	55.1
Denseaspp [95]	2018	64.7	—	76.4	87.3	91.1	76.2	83.4	77.1	43.3
DANet [19]	2019	65.3	—	77.1	88.5	92.7	78.8	85.7	73.7	43.2
EMANet [46]	2019	65.6	—	77.7	88.2	92.7	78.0	85.7	72.7	48.9
CCNet [33]	2019	64.3	—	75.9	88.3	92.5	78.8	85.7	73.9	36.3
SCAttNet V2 [43]	2020	68.3	88.0	78.4	81.8	88.8	72.5	66.3	80.3	20.2
PFNet [47]	2021	75.4	—	84.8	91.5	95.9	85.4	86.3	91.1	58.6
Segformer [89]	2021	78.0	90.5	86.4	92.9	96.4	86.9	88.1	95.2	58.9
<b>AerialFormer-T</b>	—	<u>79.5</u>	91.1	<u>87.5</u>	<b>93.5</b>	<u>96.9</u>	87.2	<u>89.0</u>	<u>95.9</u>	<b>62.5</b>
<b>AerialFormer-S</b>	—	79.3	<u>91.3</u>	87.2	<b>93.5</b>	97.0	<u>87.7</u>	88.9	<b>96.0</b>	60.2
<b>AerialFormer-B</b>	—	<b>79.7</b>	<b>91.4</b>	<b>87.6</b>	<b>93.5</b>	<b>97.2</b>	<b>88.1</b>	<b>89.3</b>	95.7	<u>61.9</u>

overlap-partitioned into a set of sub-images sized of  $896 \times 896$  with a step size of 512 by 512.

**Potsdam:** Potsdam dataset [1] contains 38 high resolution images of  $6,000 \times 6,000$  pixels over Potsdam City, Germany, and the ground sampling distance is 5 cm. The dataset is split into 24 images for training and 14 images for validation/testing. There are two modalities included in Potsdam dataset, i.e., true orthophoto (TOP) and digital surface model (DSM). While DSM consists of the near infrared (NIR) band, TOP is corresponding to RGB image. In this work, we use TOP images from Potsdam and ignore DSM images. The dataset offers two types of annotations with non-eroded (NE) and eroded (E) options, which respectively with and without the boundary. To avoid ambiguity in labeling boundaries, all experimental results are performed and benchmarked on the eroded boundary dataset. Follow experiment setup [25], [77] we divide the dataset into 24 images for training and 14 images for testing. The testset of 14 images including 2\_13, 2\_14, 3\_13, 3\_14, 4\_13, 4\_14, 4\_15, 5\_13, 5\_14, 5\_15, 6\_13, 6\_14, 6\_15, and 7\_13. The dataset consists of six categories of surfaces, building, low vegetation, tree, car, and clutter/background. We report the performance in two cases of with and without clutter. Each image is overlap-partitioned into a set of sub-images sized of  $512 \times 512$  with a step size of 256 by 256.

**LoveDA:** LoveDA dataset [80] consists of 5,987 high resolution images of  $1024 \times 1024$  pixels and 30 cm in spatial resolution. The data include 18 complex urban and rural scenes and 166,768 annotated objects from three different cities (Nanjing, Changzhou, and Wuhan) in China. In alignment with the experimental setup delineated in [80], we partition the

dataset into 2,522/1,669/1,796 images for training, validation, and testing, respectively. In evaluation scenarios involving the test set, we amalgamate the training and validation sets of LoveDA to create a combined trainval set, while keeping the test set unaltered.

### B. Evaluation Metrics

To evaluate the performance, we adopt three commonly used metrics: mean intersection over union (mIoU), overall accuracy (OA), and mean F1 score (mF1).

These metrics are computed based on four fundamental values, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The calculation of these four values involves the utilization of the prediction  $P \in \mathbb{R}^{L \times H \times W}$  and class-wise binary groundtruth mask  $GT \in \mathbb{R}^{L \times H \times W}$ , where  $H$  and  $W$  are height and width of the input image and  $L$  is the number of classes/categories existing in the input. In the context of multi-class segmentation, these values are computed for each class  $l \in [1, 2, \dots, L]$  across all pixels.

$$\begin{aligned}
 TP_l &= \sum_{h=1}^H \sum_{w=1}^W GT_{l,h,w} \wedge P_{l,h,w} \\
 TN_l &= \sum_{h=1}^H \sum_{w=1}^W \neg(GT_{l,h,w} \vee P_{l,h,w}) \\
 FP_l &= \sum_{h=1}^H \sum_{w=1}^W \neg GT_{l,h,w} \wedge P_{l,h,w} \\
 FN_l &= \sum_{h=1}^H \sum_{w=1}^W GT_{l,h,w} \wedge \neg P_{l,h,w}
 \end{aligned} \tag{8}$$

\*Categories in Potsdam dataset with Clutter: Impervious Surface (Imp.Surf), Building, Low Vegetation (Low Veg.), Tree, Car, and Clutter/Background.

TABLE III: Performance comparison on **Potsdam** *valset without clutter*. We report performance on mIoU, OA, mF1, and F1 score for each category. Note that both train and evaluation are done on the eroded dataset and we ignored the clutter category. The **bold** and *italic-underline* values in each column show the best and the second best performances.

<b>Method</b>	<b>Year</b>	<b>mIoU</b> $\uparrow$	<b>OA</b> $\uparrow$	<b>mF1</b> $\uparrow$	<b>F1 per category*</b> $\uparrow$				
		<b>Imp.</b>	<b>Surf.</b>	<b>Building</b>	<b>Low Veg.</b>	<b>Tree</b>	<b>Car</b>		
DeepLabV3+ [11]	2018	81.7	89.6	89.8	92.3	95.5	85.7	86.0	89.4
DANet [19]	2019	—	89.7	89.1	91.6	96.4	86.1	88.0	83.5
LANet [17]	2020	—	90.8	92.0	93.1	97.2	87.3	88.0	94.2
S-RA-FCN [56]	2020	72.5	88.5	89.6	90.7	94.2	83.8	85.8	93.6
FFPNet [90]	2020	86.2	91.1	92.4	93.6	96.7	87.3	88.1	96.5
ResT [101]	2021	85.2	90.6	91.9	92.7	96.1	87.5	88.6	94.8
ABCNet [45]	2021	86.5	91.3	92.7	93.5	96.9	87.9	89.1	95.8
Segmenter [65]	2021	80.7	88.7	89.2	91.5	95.3	85.4	85.0	88.5
TransUNet [8]	2021	86.1	—	88.1	92.4	94.9	82.9	88.9	91.3
HMANet [57]	2021	87.3	92.2	93.2	93.9	97.6	88.7	89.1	96.8
DC-Swin [81]	2022	87.6	92.0	93.3	94.2	97.6	88.6	89.6	96.3
BSNet [28]	2022	77.5	90.7	91.5	92.4	95.6	86.8	88.1	94.6
UNetFormer [82]	2022	86.8	91.3	92.8	93.6	97.2	87.7	88.9	96.5
FT-UNetformer [82]	2022	87.5	92.0	93.3	93.9	97.2	88.8	<b>89.8</b>	96.6
UperNet RSP-Swin-T [77]	2022	—	90.8	90.0	92.7	96.4	86.0	85.4	89.8
UperNet-RingMo [69]	2022	—	91.7	91.3	93.6	97.1	87.1	86.4	92.2
<b>AerialFormer-T</b>	—	88.5	93.5	93.7	95.2	98.0	89.1	89.1	97.3
<b>AerialFormer-S</b>	—	<u>88.6</u>	<u>93.6</u>	<u>93.8</u>	<u>95.3</u>	<b>98.1</b>	<u>89.2</u>	89.1	<u>97.4</u>
<b>AerialFormer-B</b>	—	<b>89.1</b>	<b>93.9</b>	<b>94.1</b>	<b>95.5</b>	<b>98.1</b>	<b>89.8</b>	<b>89.8</b>	<b>97.5</b>

Based on the four values above, we calculate the IoU, Accuracy (Acc), and F1 of an individual category  $l$  as follows:

$$\text{IoU}_l = \frac{TP_l}{TP_l + FN_l + FP_l} \quad (9)$$

$$\text{Acc}_l = \frac{TP_l + TN_l}{TP_l + TN_l + FN_l + FP_l} \quad (10)$$

$$\text{F1}_l = \frac{2TP_l}{2TP_l + FN_l + FP_l} \quad (11)$$

We usually refer  $\text{IoU}_l$ ,  $\text{Acc}_l$ , and  $\text{F1}_l$  as IoU, Acc, F1 of the category  $l$ . We further compute the mIoU, OA, and mF1 as the arithmetic means of the IoU, accuracy, and F1 score, respectively, for each class category.

$$\text{mIoU} = \frac{1}{L} \sum_{l=1}^L \text{IoU}_l \quad (12)$$

$$\text{OA} = \frac{1}{L} \sum_{l=1}^L \text{Acc}_l \quad (13)$$

$$\text{mF1} = \frac{1}{L} \sum_{l=1}^L \text{F1}_l \quad (14)$$

### C. Implementation Details

We trained our AerialFormer-T on a single RTX 8000 GPU, and our AerialFormer-S and AerialFormer-B on two RTX 8000 GPUs. We employed the Adam [37] optimizer with learning rate of  $6 \times 10^{-5}$ , weight decay of 0.01, betas of (0.9, 0.999) and batch size of 8. The experimental models are trained for 160k iterations for LoveDA and Potsdam dataset and 800k iterations for iSAID dataset. During the all training processes, we applied data augmentation such as random horizontal flipping and photometric distortions. Our AerialFormer has been trained on three different backbones, i.e., Swin Transformer-Tiny (Swin-T), Swin Transformer-Small (Swin-B), and Swin Transformer-Base (Swin-B). The first two backbones were pre-trained on Imagenet-1K dataset [15] and the last backbone was pre-trained on Imagenet-22k dataset [15]. As a result, we will conduct the experimental performance on three models AerialFormer-T, AerialFormer-S, and AerialFormer-B. As introduced in section III-B, we delineate the model hyperparameters: the number of channels  $C$ , window size  $M^2$ , and a set of layers  $\mathcal{L} = \{\mathcal{L}_s\}_{s=1}^{s=4}$  in Transformer Encoder Blocks, that are specific to each model, as follows:

- AerialFormer-T:  $C = 96$ ,  $M^2 = 7^2$ ,  $\mathcal{L} = \{2, 2, 6, 2\}$
- AerialFormer-S:  $C = 96$ ,  $M^2 = 7^2$ ,  $\mathcal{L} = \{2, 2, 18, 2\}$
- AerialFormer-B:  $C = 128$ ,  $M^2 = 12^2$ ,  $\mathcal{L} = \{2, 2, 18, 2\}$

In addition to the aforementioned parameters, we also take note of the receptive field sizes of the MDC Decoder, which remain constant across the models, detailed as follows:

\*Categories in Potsdam dataset without Clutter: Impervious Surface (Imp.Surf), Building, Low Vegetation (Low Veg.), Tree, and Car.

TABLE IV: Performance comparison on **LoveDA** testset dataset between our AerialFormer and other existing SOTA semantic segmentation approaches. The evaluation is based on a submission to the official server. We report performance on mIoU and IoU for each category. The **bold** and *italic-underline* values in each column show the best and the second best performances.

<b>Method</b>	<b>Year</b>	<b>mIoU</b>	<b>IoU per category ↑</b>						
			<b>Background</b>	<b>Building</b>	<b>Road</b>	<b>Water</b>	<b>Barren</b>	<b>Forest</b>	<b>Agriculture</b>
FCN [51]	2015	46.7	42.6	49.5	48.1	73.1	11.8	43.5	58.3
UNet [59]	2015	47.8	43.1	52.7	52.8	73.1	10.3	43.1	59.9
LinkNet [7]	2017	48.5	43.6	52.1	52.5	76.9	12.2	45.1	57.3
SegNet [4]	2017	47.3	41.8	51.8	51.8	75.4	10.9	42.9	56.7
UNet++ [107]	2018	48.2	42.9	52.6	52.8	74.5	11.4	44.4	58.8
DeeplabV3+ [11]	2018	47.6	43.0	50.9	52.0	74.4	10.4	44.2	58.5
FarSeg [106]	2020	48.2	43.4	51.8	53.3	76.1	10.8	43.2	58.6
TransUNet [8]	2021	48.9	43.0	56.1	53.7	78.0	9.3	44.9	56.9
Segmenter [65]	2021	47.1	38.0	50.7	48.7	77.4	13.3	43.5	58.2
Segformer [89]	2021	49.1	42.2	56.4	50.7	78.5	17.2	45.2	53.8
DC-Swin [81]	2022	50.6	41.3	54.5	56.2	78.1	14.5	<u>47.2</u>	62.4
ViTAE-B+RVSA [78]	2022	<u>52.4</u>	—	—	—	—	—	—	—
FactSeg [53]	2022	48.9	42.6	53.6	52.8	76.9	16.2	42.9	57.5
UNetFormer [82]	2022	<u>52.4</u>	44.7	58.8	54.9	79.6	<b>20.1</b>	46.0	62.5
RSSFormer [91]	2023	<u>52.4</u>	<b>52.4</b>	<b>60.7</b>	55.2	76.3	18.7	45.4	58.3
<b>AerialFormer-T</b>	—	52.0	45.2	57.8	56.5	79.6	<u>19.2</u>	46.1	59.5
<b>AerialFormer-S</b>	—	<u>52.4</u>	46.6	57.4	<u>57.3</u>	<u>80.5</u>	15.6	46.8	<u>62.8</u>
<b>AerialFormer-B</b>	—	<b>54.1</b>	<u>47.8</u>	<b>60.7</b>	<b>59.3</b>	<b>81.5</b>	17.9	<b>47.9</b>	<b>64.0</b>

$[r_1, r_2, r_3] = \{[1, 3, 3], [3, 3, 3], [3, 5, 7], [3, 5, 7], [3, 5, 7]\}$  as demonstrated in Figure 2.

It is worth highlighting that, relative to the commonly utilized CNN backbones, our model does not significantly increase computational cost, as computational complexities of Swin-T and Swin-S align closely with those of ResNet-50 and ResNet-101, respectively.

#### D. Quantitative Results and Analysis

The quantitative performance comparisons between our AerialFormer with other existing methods are presented in Tables I, II, III, IV for three different datasets under various settings of iSAID (valset), Potsdam (with clutter), Potsdam (without clutter), and LoveDA (testset) , respectively. For each dataset, we report the performance of the proposed AerialFormer on three backbones of Swin-T, Swin-S and Swin-B and name them as AerialFormer-T, AerialFormer-S, AerialFormer-B, respectively. We compare our AerialFormer with both CNN-based and Transformer-based image segmentation methods. The comparison on each dataset is detailed as follows:

1) **iSAID semantic segmentation results:** Performance comparisons of our proposed AerialFormer with existing state-of-the-art methods on the iSAID dataset are presented in Table I. iSAID dataset consists of 15 categories and divided into three groups of vehicles, artifacts and fields. In general, we observe that our AerialFormer-B achieves the best performance, while both AerialFormer-S and AerialFormer-T obtain comparable results as the second best methods. All three models outperform other existing methods significantly.

Specifically, our AerialFormer-T obtains a mIoU of 67.5%, AerialFormer-S achieves a mIoU of 68.4%, and AerialFormer-B attains a mIoU of 69.3%. Those results present improvements of 0.3%, 1.2% and 2.1% over the previous highest score of 67.2% from RingMo [69]. Moreover, on some small and dense classes (e.g. small vehicles (SV), planes, helicopters (HC), etc), our AerialFormer gains a big margin compared to the existing methods. Take small vehicles (SV) class as an example, our AerialFormer-T achieve 1.4% IoU gain, AerialFormer-S gains 2.4% IoU margin, AerialFormer-B gains 2.5% IoU margin better than the best existing method i.e., RingMo [69]. It is worthy noting that RingMo utilizes Swin-B as its backbone, which share similar computational cost with our AerialFormer-B. This analysis further shows that both our AerialFormer-T and AerialFormer-S, despite being smaller models, outperform the best existing method, RingMo.

2) **Potsdam semantic segmentation results:** We analyze segmentation performance on Potsdam dataset in two cases of with and without Clutter/Background and the results are summarized in Table II and Table III, respectively. Clutter class is the most challenging class as it can contain anything except for the five name classes of Impervious Surface, Building, Low Vegetation, Tree, Car. Similar other existing work [43], [69], [82], we benchmark our AerialFormer using various metrics of mIoU, OA, mF1 and F1 per category.

**Potsdam with Clutter:** Table II reports the performance comparisons between our AerialFormer with the existing methods on 6 classes (i.e. including Clutter class). It is note that among all existing methods, Segformer [89] is a strong Transformer-based segmentation model and obtains the best performance.

Our model gains a remarkable improvement of 1.7% in mIoU, 0.9% in OA, and 1.2% in mF1 compared with the best existing methods Segformer.

Different from experiment on iSAD (section IV-D1), the tradeoff between performance and model size doesn't seem favorable for this dataset. We speculate that the cause for this could be the difference in the spatial resolution of the datasets. As per [52], while the iSAID dataset includes images with spatial resolutions of up to  $0.3\text{ m}$ , the spatial resolution of the Potsdam dataset is finer at  $0.05\text{ m}$ . Consequently, objects in the Potsdam dataset are represented with more pixels, appearing much larger. This might lessen the requirement for architectural enhancements specifically aimed at improving the segmentation of tiny objects.

As the most challenging category, F1 score on Clutter is lowest compared to other five categories. Because of the challenging Clutter category, many methods have ignored this category and focused on training the network on only 5 other categories as shown in Table III below.

**Potsdam without Clutter:** In this experimental setting, the review shows that FT-UNetformer [82], HMANet [57] and DC-Swin [81] obtained the best score on mIoU, OA, mF1 metrics and none of them can achieve the best score on all three metrics. In the other hand, our AerialFormer-B obtains the best score on all three metrics and gains an improvement of 1.6% mIoU, 1.7% OA, and 0.9% mF1 compared to FT-UNetformer, HMANet, DC-Swin, respectively. Compared to Table II which contains Clutter, we can see that Clutter, when ignored, tends to alleviate the ambiguity amongst the remaining classes.

Similar to the observation on iSAD dataset (section IV-D1), we observe that AerialFormer-B achieves the best performance, while both AerialFormer-S and AerialFormer-T obtain comparable results as the second best methods on Potsdam dataset in both settings of with and without Clutter category.

3) **LoveDA semantic segmentation results:** We report performance comparisons with existing methods on *testset* splits of LoveDA dataset in Table IV. In this experiment, we evaluated our method on the public test server \* by sending our predictions. Our smaller model, AerialFormer-S, achieves comparable performance to the existing state-of-the-art methods, such as UNetFormer [82] and RSSFormer [91], with an mIoU (mean Intersection over Union) of 52.4%. Whereas, our best model, AerialFormer-B, shows a significant improvement of 1.7% in mIoU compared to the existing state-of-the-art methods. Notably, AerialFormer-B outperforms the existing methods by 4.1% IoU for the Road category, 5.2% IoU for the Water category, 2.5% IoU for the Forest category, and 5.7% IoU for the Agriculture category. Particularly, 'Road' category, which is typically characterized by narrow and elongated features. Segmenting such objects necessitates both local and global perspectives, a capability our model exhibits effectively.

4) **Network Complexity:** Besides qualitative analysis, we also include an analysis of the network complexity, as presented in Table V. In this section, we provide details on

the model parameters (MB), computation (GFLOPs), and inference time (seconds per image) for our AerialFormer, and compare it with two baseline models: Unet [59], which is a CNN-based network, and TransUnet [8], which is a Transformer-based network. To calculate the inference time, we averaged the results of 10,000 runs of the model using  $512 \times 512$  input with a batch size of 1. While our AerialFormer-T has a similar model size and inference time to Unet [59], it requires fewer computational resources and achieves significantly higher performance. For example, it achieves a 31.9% improvement in mIoU on the iSAID dataset. When compared to TransUnet [8], our AerialFormer-T has a comparable inference time of 0.027 seconds per image, as opposed to 0.038 seconds per image. Additionally, it requires a smaller model size, incurs lower computational costs, and achieves higher performance. For instance, it gains a 3.0% mIoU improvement on the Potsdam validation set without 'Clutter' class, and a 5.2% mIoU improvement on the LoveDA test set. Even with slightly longer inference times, our models still meet real-time speed requirements. The smallest model in our series, AerialFormer-T, can perform inference at a rate of 37 images per second, while AerialFormer-S achieves 25.6 images per second. Even the largest model, AerialFormer-B, with a model size of 113.82MB, can achieve real-time inference speed at 15.4 images per second.

Method	Params (MB)	GFLOPs (GB)	Inference Time (s)
Unet [59]	29.1	203.4	0.038
TransUnet [8]	90.7	233.7	0.023
AerialFormer-T	42.7	49.0	0.027
AerialFormer-S	64.0	72.2	0.039
AerialFormer-B	113.8	126.8	0.065

TABLE V: Performance comparison of our models with different sizes of the backbone.

### E. Qualitative Results and Analysis

We will now present the qualitative results obtained from our model, comparing them

with well-established and robust baseline models, specifically PSPNet [102] and DeepLabV3+ [11]. In this section, we will illustrate the advantages of our AerialFormer in dealing with challenging characteristics of remote sensing images.

**Tiny objects:** As evidenced in Fig. 6, our model, AerialFormer, is capable of accurately identifying and segmenting tiny objects like cars on the road, which might only be represented by approximately  $10 \times 5$  pixels. This showcases the model's remarkable capability to handle small object segmentation in high-resolution aerial images. Additionally, our model demonstrates the ability to accurately segment cars that are not present in the ground truth labels (red boxes). However, this poses a problem in evaluating our model, as its prediction could be penalized as false positive even if the prediction is correct based on the given image.

\*<https://codalab.lisn.upsaclay.fr/competitions/421>

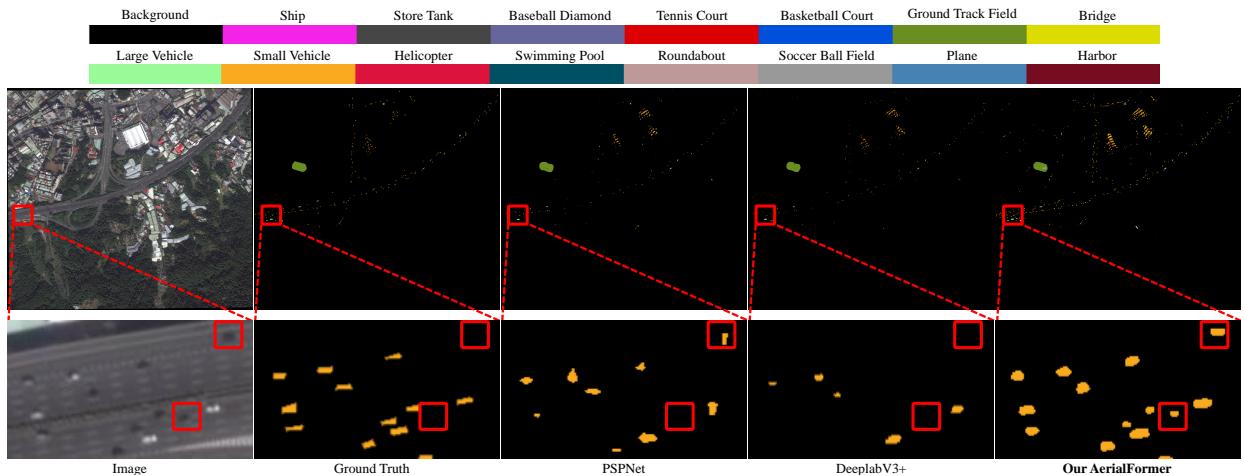


Fig. 6: Qualitative comparison between our AerialFormer with PSPNet [102], DeepLabV3+ [11] on **tiny objects**. From left to right are the original image, Groundtruth, PSPNet, DeepLabV3+, and our AerialFormer. The first row is overall performance and the second row is zoom-in region. We note that some of the objects that are evident in the input are ignored in the ground truth label.

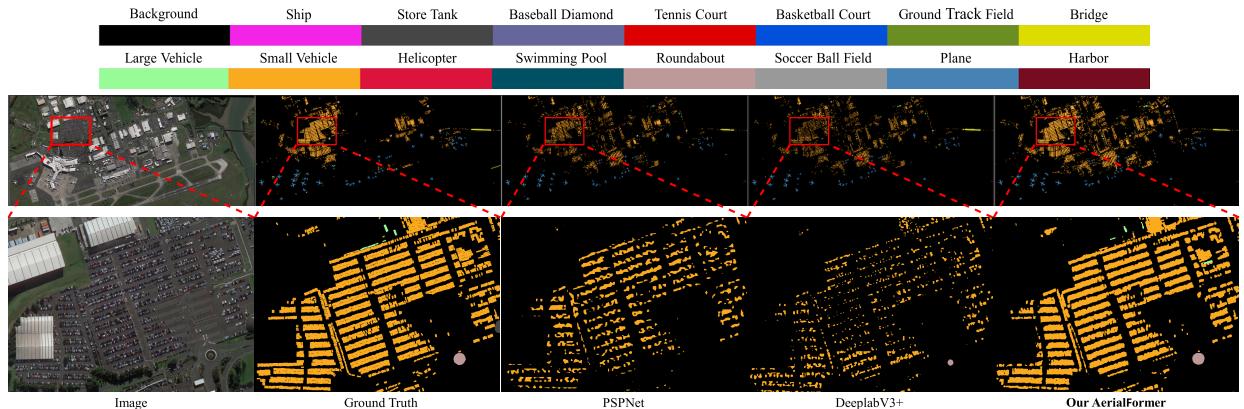


Fig. 7: Qualitative comparison between our AerialFormer with PSPNet [102], DeepLabV3+ [11] on **dense objects**. From left to right are the original image, Groundtruth, PSPNet, DeepLabV3+, and our AerialFormer. The first row is overall performance and the second row is zoom-in region.

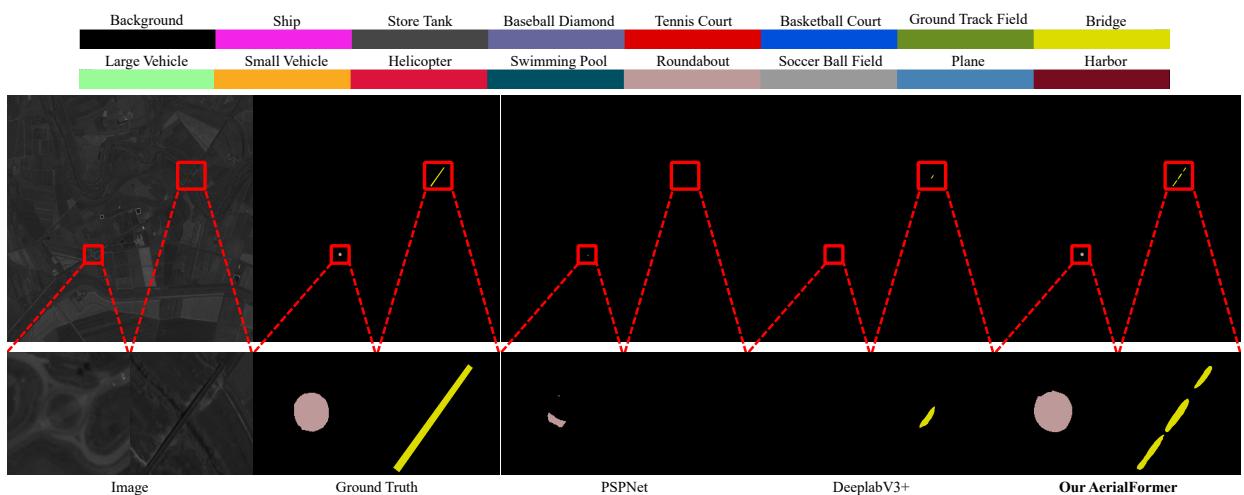


Fig. 8: Qualitative comparison between our AerialFormer with PSPNet [102], DeepLabV3+ [11] on **foreground-background imbalance**. From left to right are the original image, Groundtruth, PSPNet, DeepLabV3+, and our AerialFormer. The first row is overall performance and the second row is zoom-in region.

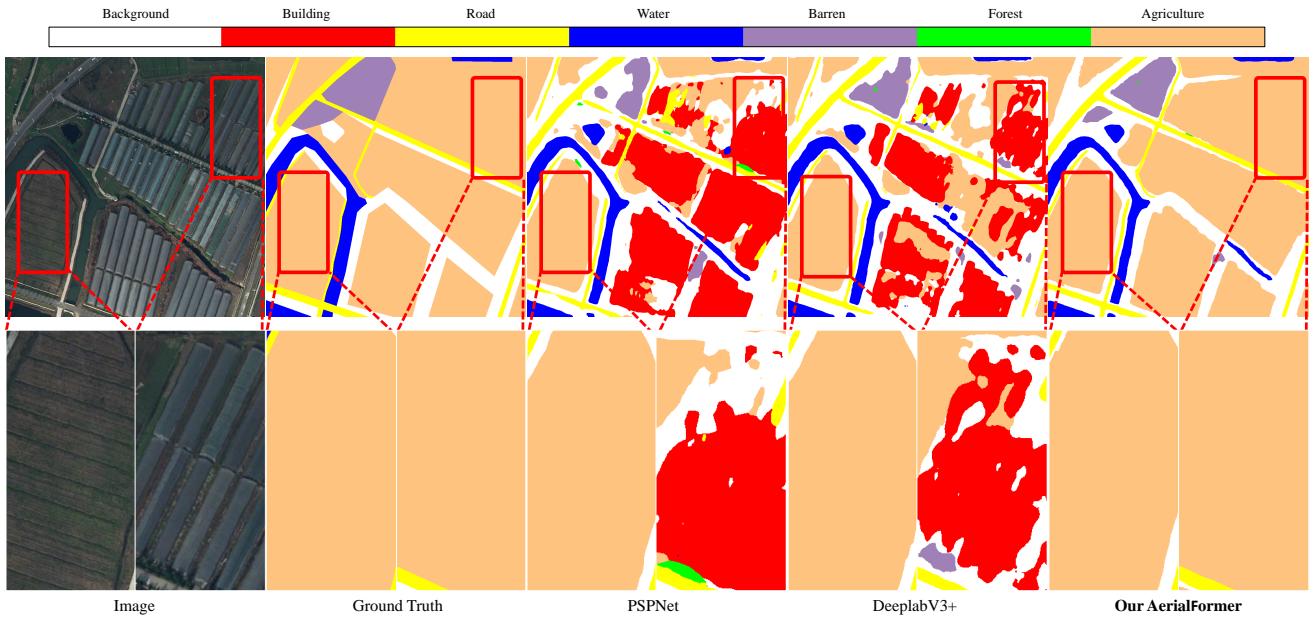


Fig. 9: Qualitative comparison between our AerialFormer with PSPNet [102], DeepLabV3+ [11] on **Intra-class heterogeneity**: the regions highlighted in the box are both classified under the 'Agriculture' category. However, one region features green lands, while the other depicts greenhouses. From left to right are the original image, Groundtruth, PSPNet, DeepLabV3+, and our AerialFormer.

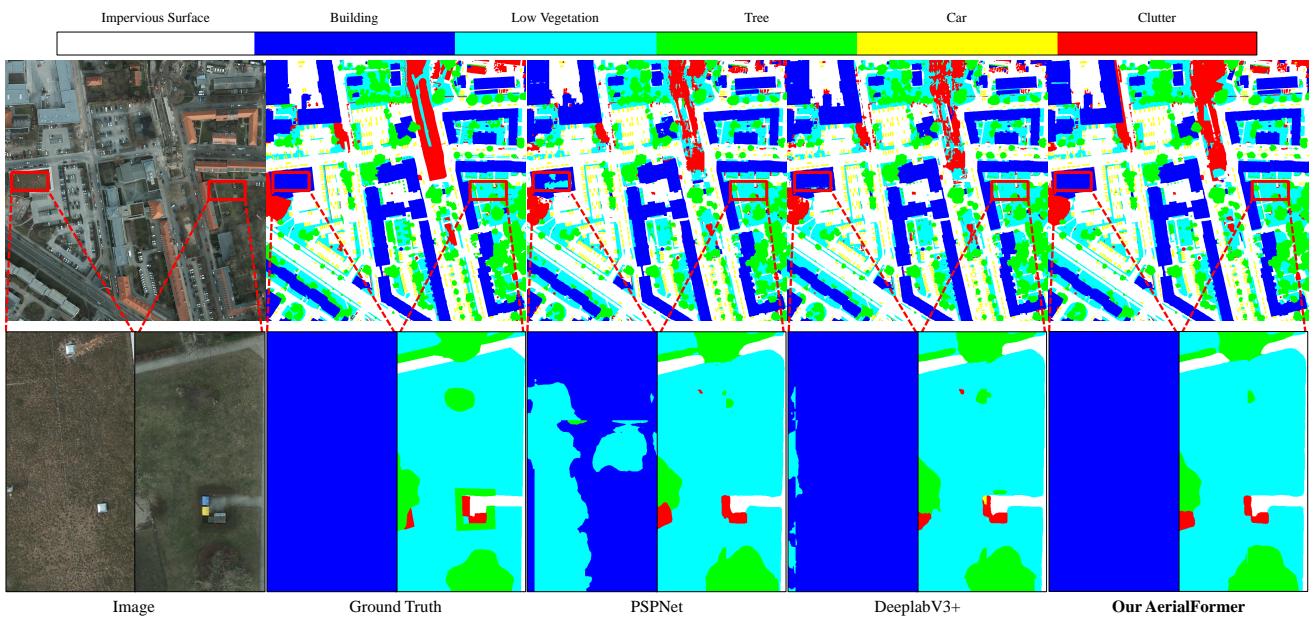


Fig. 10: Qualitative comparison between our AerialFormer with PSPNet [102], DeepLabV3+ [11] on **inter-class homogeneity**: the regions highlighted in the box share similar visual characteristics but one region is classified as a 'Building' while the other is classified as belonging to the 'Low Vegetation' category. From left to right are the original image, Groundtruth, PSPNet, DeepLabV3+, and our AerialFormer. The first row is overall performance and the second row is zoom-in region.

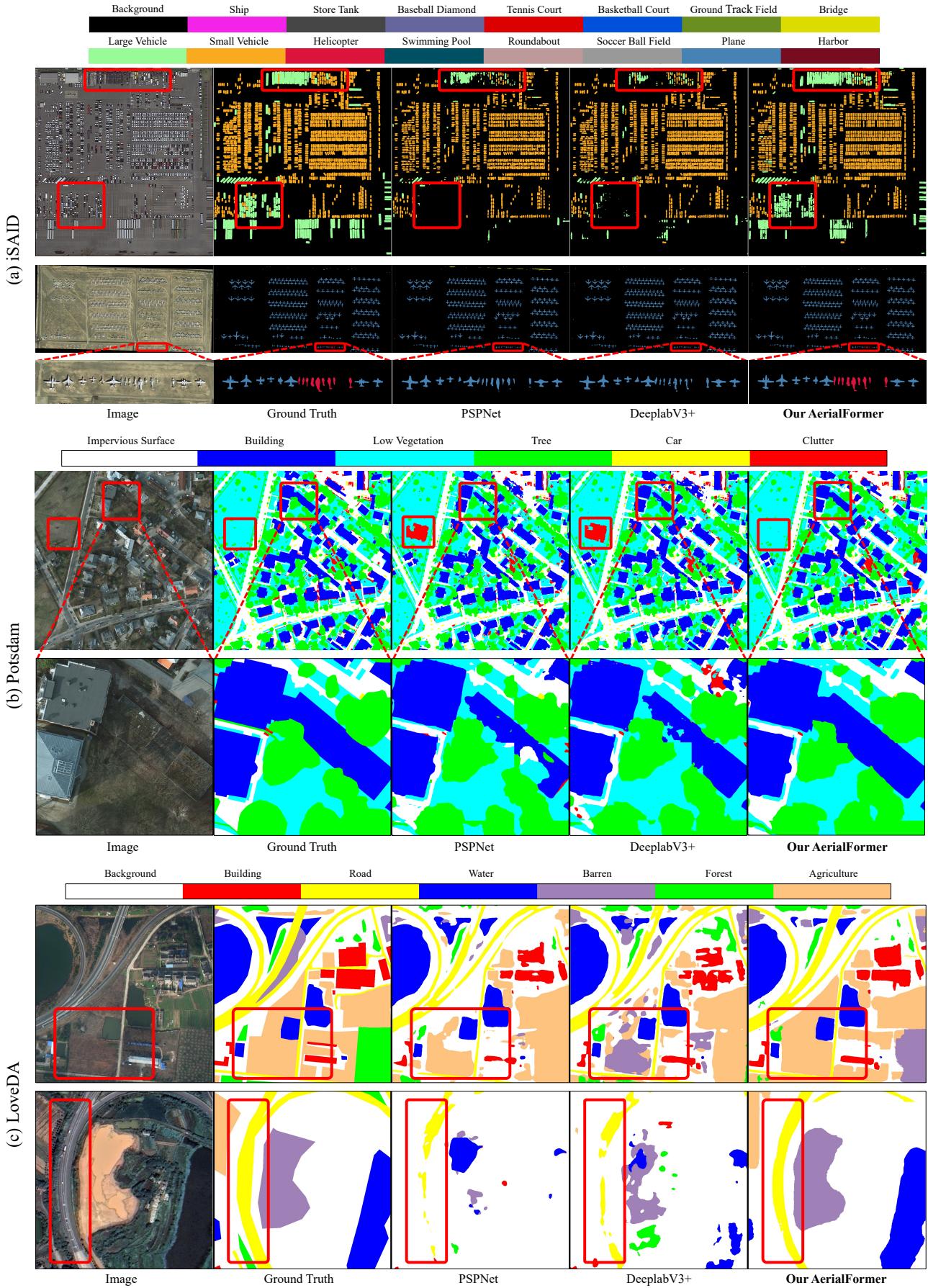


Fig. 11: Qualitative comparison on various datasets: (a) iSAID, (b) Potsdam, and (c) LoveDA. From left to right: original image, ground truth, PSPNet, DeeplabV3+, and our AerialFormer. We highlight the major difference in red boxes.

Dense objects: Fig. 7 demonstrates the proficient ability of our model in accurately segmenting dense objects, particularly clusters of small vehicles, which often pose challenges for baseline models. The baseline models frequently overlook or struggle with identifying such objects. We ascribe the success of our model in segmenting dense objects to the MDC decoder that can capture the global context and the CNN stem that can bring the local details of the tiny objects.

Foreground-background imbalance: As mentioned in Section I, Introduction, the iSAID dataset exhibits a notable foreground and background imbalance. This imbalance is particularly evident in Fig. 8, where certain images contain only a few labeled objects. Despite this extreme imbalance, AerialFormer demonstrates its capability to accurately segment the objects of interest, as depicted in the figure.

Intra-class heterogeneity: Fig. 9 visually demonstrates the existence of intra-class heterogeneity in aerial images, where objects of the same category can appear in diverse shapes, textures, colors, scales, and structures. The red boxes indicate two regions that are classified as belonging to the 'Agriculture' category. However, their visual characteristics significantly differ due to the presence of greenhouses. Notably, while baseline models encounter challenges in correctly classifying the region with greenhouses, misclassifying it as 'Building', our proposed model successfully identifies and labels the region as 'Agriculture'. This showcases the superior performance and effectiveness of our model in handling the complexities of intra-class variations in aerial image analysis tasks.

Inter-class heterogeneity: Fig. 10 illustrates the inter-class homogeneity in aerial images, where objects of different classes may exhibit similar visual properties. The regions enclosed within the red boxes represent areas that exhibit similar visual characteristics, i.e., rooftop greened with lawn and the park. However, there is a distinction in the classification of these regions, with the former being labeled as 'Building' and the latter falling into the 'Low Vegetation' category. While the baseline models are confused by the appearance and produce mixed prediction, we see our model can produce more robust result.

Overall performance: Fig. 11 showcases these qualitative outcomes across three datasets: (a) iSAID, (b) Potsdam, and (c) LoveDA. Each dataset possesses unique characteristics and presents a wide spectrum of challenges encountered in aerial image segmentation. We highlight the major difference among methods in red boxes. Fig. 11 (a) visually demonstrates the efficiency of our model in accurately recognizing *dense and tiny objects*. Unlike the baseline models, which often overlook or misclassify these objects into different categories, our model exhibits its robustness in handling dense and tiny objects, e.g., Small Vehicle (SV) and Helicopter (HC).

As depicted in Fig. 11 (b), our model demonstrates a reduced level of inter-class confusion in comparison to the baseline models. An instance of this is evident in the prediction of building structures, where the baseline models exhibit confusion. In contrast, our model delivers predictions closely aligned with the ground truth. Similarly, in Fig. 11 (c), our model's predictions are less noisy, further asserting its robustness in scenarios where scenes belong to different categories

but exhibit similar visual appearances. As in the quantitative analysis, the performance of our model on the 'Road' class is visually appealing. Our model's ability to accurately delineate road structures, despite their narrow and elongated features, is visibly superior.

## V. CONCLUSION

In this study, we have introduced AerialFormer, a novel approach specifically designed to address the unique and challenging characteristics encountered in remote sensing image segmentation. These challenges include the presence of tiny objects, dense objects, foreground-background imbalance, intra-class heterogeneity, and inter-class homogeneity. To overcome these challenges, we designed AerialFormer by combining the strengths of both Transformers and CNNs architectures, creating a hybrid model that incorporates a Transformer encoder with a multi-dilated CNN decoder. Furthermore, we incorporated a CNN Stem module to facilitate the transmission of low-level, high-resolution features to the decoder. This comprehensive design allows AerialFormer to effectively capture global context and local features simultaneously, significantly enhancing its ability to handle the complexities inherent in aerial images.

We have evaluated our proposed AerialFormer using three different backbone sizes: Swin Transformer-Tiny, Swin Transformer-Small, and Swin Transformer-Base. Our model was benchmarked on three standard datasets: iSAID, Potsdam, and LoveDA. Through extensive experimentation, we demonstrated that AerialFormer-T and AerialFormer-S, with smaller model sizes and lower computational costs, achieve performance that is either superior or comparable to existing state-of-the-art methods, ranking them as the second-best performers. Moreover, our proposed AerialFormer-B surpasses all existing state-of-the-art methods, showcasing its exceptional performance in the field of remote sensing image segmentation.

## ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation (NSF) under Award No OIA-1946391 RII Track-1, NSF 1920920 RII Track 2 FEC, NSF 2223793 EFRI BRAID, NSF 2119691 AI SUSTEIN.

## REFERENCES

- [1] 2d semantic labeling contest - potsdam. *International Society for Photogrammetry and Remote Sensing*.
- [2] RB Andrade, GAOP Costa, GLA Mota, MX Ortega, RQ Feitosa, PJ Soto, and Christian Heipke. Evaluation of semantic segmentation methods for deforestation detection in the amazon. *ISPRS Archives*; 43, B3, 43(B3):1497–1505, 2020.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
- [5] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3602–3610, 2016.
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.
- [7] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017.
- [8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.
- [13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.
- [14] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [16] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6819–6829, 2019.
- [17] Lei Ding, Hao Tang, and Lorenzo Bruzzone. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):426–435, 2020.
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [19] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.
- [20] David Griffiths and Jan Boehm. Improving public data for building segmentation from convolutional neural networks (cnns) for fused airborne lidar and image data using active contours. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154:70–83, 2019.
- [21] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5038–5047, 2017.
- [22] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3562–3572, 2019.
- [23] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [25] Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.
- [26] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [27] Dinh-Hieu Hoang, Gia-Han Diep, Minh-Triet Tran, and Ngan T H Le. Dam-al: Dilated attention mechanism with attention loss for 3d infant brain image segmentation. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 660–668, 2022.
- [28] Jianlong Hou, Zhi Guo, Youming Wu, Wenhui Diao, and Tao Xu. Bsnet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–22, 2022.
- [29] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [30] Chi-Wei Hsiao, Cheng Sun, Hwann-Tzong Chen, and Min Sun. Specialize and fuse: Pyramidal output representation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7137–7146, 2021.
- [31] Hanze Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 1–17. Springer, 2020.
- [32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [33] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.
- [34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.
- [35] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7231–7241, 2021.
- [36] Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7189–7198, 2021.
- [37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [38] Ngan Le, Toan Bui, Viet-Khoa Vo-Ho, Kashu Yamazaki, and Khoa Luu. Narrow band active contour attention model for medical segmentation. *Diagnostics*, 11(8):1393, 2021.
- [39] Ngan Le, Trung Le, Kashu Yamazaki, Toan Bui, Khoa Luu, and Marios Savvides. Offset curves loss for imbalanced problem in medical segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9189–9195. IEEE, 2021.
- [40] Ngan Le, Kashu Yamazaki, Kha Gia Quach, Dat Truong, and Marios Savvides. A multi-task contextual atrous residual network for brain tumor detection & segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5943–5950. IEEE, 2021.

- [41] T Hoang Ngan Le, Chi Nhan Duong, Ligong Han, Khoa Luu, Kha Gia Quach, and Marios Savvides. Deep contextual recurrent residual networks for scene labeling. *Pattern Recognition*, 80:32–41, 2018.
- [42] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.
- [43] Haifeng Li, Kaijian Qiu, Li Chen, Xiaoming Mei, Liang Hong, and Chao Tao. Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 18(5):905–909, 2020.
- [44] Hancho Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.
- [45] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter M Atkinson. Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181:84–98, 2021.
- [46] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019.
- [47] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2021.
- [48] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020.
- [49] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.
- [50] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [51] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [52] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021.
- [53] Ailong Ma, Junjue Wang, Yanfei Zhong, and Zhuo Zheng. Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.
- [54] Diego Marcos, Michele Volpi, Benjamin Kellenberger, and Devis Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS journal of photogrammetry and remote sensing*, 145:96–107, 2018.
- [55] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.
- [56] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7557–7569, 2020.
- [57] Ruigang Niu, Xian Sun, Yu Tian, Wenhui Diao, Kaiqiang Chen, and Kun Fu. Hybrid multiple attention network for semantic segmentation in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.
- [58] Saffron J O’neill, Maxwell Boykoff, Simon Niemeyer, and Sophie A Day. On the use of imagery for climate change engagement. *Global environmental change*, 23(2):413–421, 2013.
- [59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [60] Abdus Samie, Azhar Abbas, Muhammad Masood Azeem, Sidra Hamid, Muhammad Amjad Iqbal, Shaikh Shamim Hasan, and Xiangzheng Deng. Examining the impacts of future land use/land cover changes on climate in punjab province, pakistan: implications for environmental sustainability and economic growth. *Environmental Science and Pollution Research*, 27:25415–25433, 2020.
- [61] Guy JP Schumann, G Robert Brakenridge, Albert J Kettner, Rashid Kashif, and Emily Niebuhr. Assisting flood disaster response with earth observation data and products: A critical assessment. *Remote Sensing*, 10(8):1230, 2018.
- [62] Ayesha Shafique, Guo Cao, Zia Khan, Muhammad Asad, and Muhammad Aslam. Deep learning-based change detection in remote sensing images: A review. *Remote Sensing*, 14(4):871, 2022.
- [63] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, Ruili Wang, and Jie Yang. Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):4673–4688, 2020.
- [64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [65] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmente: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.
- [66] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 347–365. Springer, 2020.
- [67] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.
- [68] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.
- [69] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.
- [70] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [71] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [72] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jegou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.
- [73] Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. *British Machine Vision Conference (BMVC)*, 2022.
- [74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [75] Khoa Vo, Hyekang Joo, Kashu Yamazaki, Sang Truong, Kris Kitani, Minh-Triet Tran, and Ngan Le. AEI: Actors-Environment Interaction with Adaptive Attention for Temporal Action Proposals Generation. *BMVC*, 2021.
- [76] Khoa Vo, Sang Truong, Kashu Yamazaki, Bhiksha Raj, Minh-Triet Tran, and Ngan Le. Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation. *International Journal of Computer Vision*, pages 1–22, 2022.
- [77] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.
- [78] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer towards

- remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.
- [79] Fang Wang, Shihao Piao, and Jindong Xie. Cse-hrnet: A context and semantic enhanced high-resolution network for semantic segmentation of aerial imagery. *IEEE Access*, 8:182475–182489, 2020.
- [80] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021.
- [81] Libo Wang, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.
- [82] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.
- [83] Wenguang Wang, Tianfei Zhou, Siyuan Qi, Jianbing Shen, and Song-Chun Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3508–3522, 2021.
- [84] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
- [85] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.
- [86] Marie Weiss, Frédéric Jacob, and Grgory Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, 236:111402, 2020.
- [87] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023.
- [88] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on computer vision (ECCV)*, pages 418–434, 2018.
- [89] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.
- [90] Qingsong Xu, Xin Yuan, Chaojun Ouyang, and Yue Zeng. Spatial-spectral ffpnet: Attention-based pyramid network for segmentation and classification of remote sensing images. *arXiv preprint arXiv:2008.08775*, 2020.
- [91] Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Transactions on Image Processing*, 32:1052–1064, 2023.
- [92] Gunagkuo Xue, Yikun Liu, Yuwen Huang, Mingsong Li, and Gongping Yang. Aanet: an attention-based alignment semantic segmentation network for high spatial resolution remote sensing images. *International Journal of Remote Sensing*, 43(13):4836–4852, 2022.
- [93] Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, and Ngan Le. Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3656–3661. IEEE, 2022.
- [94] Kashu Yamazaki, Khoa Vo, Sang Truong, Bhiksha Raj, and Ngan Le. Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. *The Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023.
- [95] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.
- [96] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019.
- [97] Hongfeng You, Shengwei Tian, Long Yu, and Yalong Lv. Pixel-level remote sensing image recognition based on bidirectional word vectors. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):1281–1293, 2019.
- [98] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12416–12425, 2020.
- [99] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.
- [100] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.
- [101] Qinglong Zhang and Yu-Bin Yang. Rest: An efficient transformer for visual recognition. *Advances in Neural Information Processing Systems*, 34:15475–15485, 2021.
- [102] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [103] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018.
- [104] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13666–13675, 2020.
- [105] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.
- [106] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105, 2020.
- [107] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.
- [108] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.