MAC: Multi-Scales Attention Cascade for Aerial Image Segmentation

Yubo Wang[1], Zhao Wang[1], Yuusuke Nakano[2]
Katsuya Hasegawa[3], Hiroyuki Ishii[1] and Jun Ohya[1]
[1]*Department of Modern Mechanical and Engineering, Waseda University, Tokyo, Japan*
[2]*Network Service Systems Laboratories, NTT Corporation, Tokyo, Japan*
[3]*Institute of Space and Astronautical Science, JAXA, Kanagawa, Japan*

Keywords: Geospatial Information Processing, Image Processing, Aerial Image Segmentation, Transformer Model, Feature Pyramid Network (FPN).

Abstract: Unlike general semantic segmentation, aerial image segmentation has its own particular challenges, three of the most prominent of which are great object scale variation, the scattering of multiple tiny objects in a complex background and imbalance between foreground and background. Previous affinity learning-based methods introduced intractable background noise but lost key-point information due to the additional interaction between different level features in their Feature Pyramid Network (FPN) like structure, which caused inferior results.We argue that multi-scale information can be further exploited in each FPN level individually without cross-level interaction, then propose a Multi-scale Attention Cascade (MAC) model to leverage spatial local contextual information by using multiple sized non-overlapping window self-attention module, which mitigates the effect of complex and imbalanced background. Moreover, the multi-scale contextual cues are propagated in a cascade manner to tackle the large scale variation problem while extracting further details. Finally, a local channels attention is presented to achieve cross-channel interaction. Extensive experiments verify the effectiveness of MAC and demonstrate that the performance of MAC surpasses those of the state-of-the-art approaches by +2.2 mIoU and +3.1 mFscore on iSAID dataset, by +2.97 mIoU on ISPRS Vaihingen dataset. Code has been made available at https://github.com/EricBooob/Multi-scale-Attention-Cascade-for-Aerial-Image-Segmentation.

1 INTRODUCTION

1.1 Merits and Challenges in Aerial Imagery

High Spatial Resolution (HSR) remote sensing imagery has the hallmark of containing plentiful geospatial information, which provides semantic and localization for the objects of interest, including buildings, vehicles, ships, etc. Understanding these information is essential for various practical purposes, e.g., city monitoring, environment change surveillance, disaster response and route planning. For HSR remote sensing images, aerial image segmentation is an important computer vision task that aims to segment foreground objects and background area while assigning a semantic label to each image pixel from an aerial viewpoints.

However, in contrast to common semantic segmentation task in natural scene, aerial image segmentation contains the three dominant challenging cruxes:

1) *Great object scale variation in the same scene* (Xia et al., 2018, Waqas Zamir et al., 2019). The scale of objects in aerial imagery varies in a quite wide range, which means that extremely tiny and large objects are diffcult to segment.

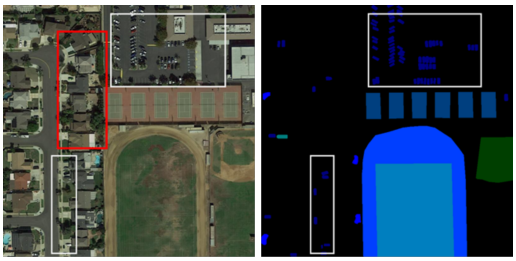
2) *The spreading of a large number of tiny objects in HSR images* (Xia et al., 2018). Numerous tiny objects pervade in the large aerial image, so to recognizing and segmenting them distinctly is an intractable issue, especially for the ambiguous boundaries.

3) *Imbalanced and complex background.* The ratio of foreground is much less than that of the complex background (Waqas Zamir et al., 2019), which brings about noise in modeling while causing serious false positives for outputs.

As shown in Figure. 1, this aerial image example contains objects with multiple scales from the

Wang, Y., Wang, Z., Nakano, Y., Hasegawa, K., Ishii, H. and Ohya, J.
MAC: Multi-Scales Attention Cascade for Aerial Image Segmentation.
DOI: 10.5220/0012343500003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 37-47
ISBN: 978-989-758-684-2; ISSN: 2184-4313



(a) Aerial Image (b) Segmentation map

Figure 1: An example of aerial image (Waqas Zamir et al., 2019), in which, (a) is the original aerial image input and (b) is the corresponding ground-truth of segmentation map. This typical image illustrates the main challenges in aerial image segmentation task. (1) From very small to extremely large, multi-scale objects appear in the same scene (*great scale variation*); (2) the white box shows many small vehicles disperses around the image (*the spreading of a large number of tiny objects*); and (3) the whole image demonstrates the imbalance between foreground and background while the red box shows the complex background (imbalanced and complex background).

very small to the extremely large. In the meantime, the whole image shows the imbalance between foreground and background. In addition,the red box contains the complex background including buildings, trees, etc., and the white box demonstrates that numerous tiny vehicles spread in the whole scene.

1.2 Recent Development on Semantic and Aerial Segmentation

For a general semantic segmentation task, as a result of the impressive success of Fully Convolution Network, i.e., FCN (Long et al., 2015), some of its derivatives (Chen et al., 2017, Chen et al., 2018, Zhao et al., 2017) utilized elaborate dilated convolution layers and pyramid pooling modules to achieve multi-scales contexts aggregation. However, for HSR remote sensing imagery, they obtain inferior outputs due to the imbalance between foreground and background. Meanwhile, the performances of some recent object boundaries enhancing methods (Kirillov et al., 2020, Takikawa et al., 2019) are also limited by the intricate background and intractable tiny objects. Recently, Feature Pyramid Network, i.e., FPN (Lin et al., 2017) has become the most prevalent component to tackle the scale variation problem. Some FPN based methods (Kirillov et al., 2019, Xiao et al., 2018) achieve multiple level feature fusion and representation, but they ignore the imbalanced background and cause serious false positive on their outputs.

In Natural Language Processing (NLP), Transformer (Vaswani et al., 2017) caused a profound change and a large leap forward in contextual information capturing. Inspired by Transformer, many Vision Transformer based methods (Dosovitskiy et al., 2020, Liu et al., 2021, Strudel et al., 2021, Xie et al., 2021, Zheng et al., 2021) have been proposed in Computer Vision (CV). Though these methods can generate accurate prediction on tiny and ambiguous objects, they cannot accurately segment large objects boundary due to the great scale variation in aerial images. For dense affinity learning based methods (Fu et al., 2019, Li et al., 2021, Zheng et al., 2020), their segmentation results are degraded by complex background and noise context. Pointflow (Li et al., 2021) adopts sparse affinity learning by selecting and matching salient points between adjacent level features of FPN. Though it can handle the complex background and noise, this method also results in the lose of tiny objects and weaker prediction for large objects boundaries.

1.3 Essence and Contributions of this Work

In this paper, the aforementioned issues are handled by our proposed Multi-scale Attention Cascade model, which is abbreviated in MAC. On the basis of MAC, self-attention in a different size non-overlapping window is computed to exploit the spatial local contextual cues while mitigating the effect of complex and imbalanced background. Rather than the invariable window size in previous methods (Dosovitskiy et al., 2020, Liu et al., 2021), the key design element of MAC is *mac* module, the strategy of that is illustrated in Figure. 2. Based on *mac* module, a multi-scale window-wise multi-head self-attention is successively stacked in a cascade manner to cope with the great scale-variation in aerial images. Benefiting from these merits, MAC not only generates higher resolution masks on tiny objects, but also better predicts boundaries on very large objects. In addition to the spatial self-attention, we further present a local channel attention at the end of *mac* module to achieve cross-channel interaction and make the homogeneous feature compact along the channel dimension.

Different from the previous affinity learning methodology (Li et al., 2021, Zheng et al., 2020) to select and match the contextual information between different levels of FPN, the central idea of MAC is to operate the cascaded module in each pyramid level individually. This methodology is motivated by the analysis on the success of FPN: 1) by constructing feature pyramid and leveraging multi-scale feature fusion, FPN obtains better feature representation, and 2) each level of FPN output accounts for the predic-

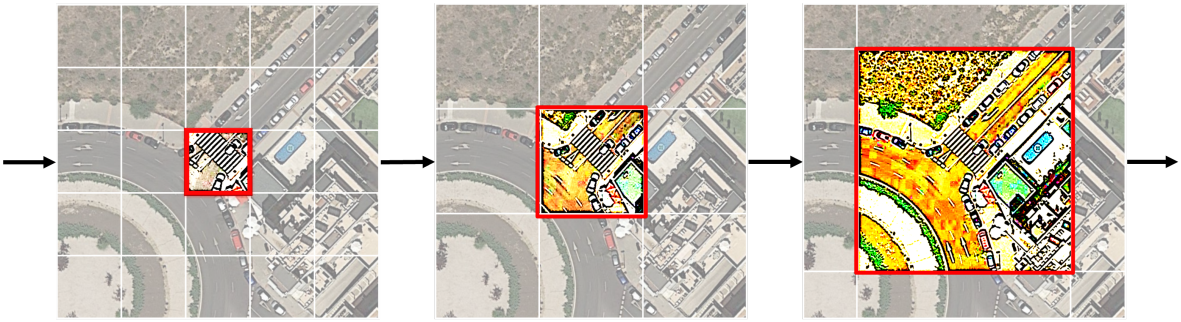


Figure 2: Strategy illustration of the **mac** module, which consists of three consecutive cascade stages. Unlike previous fixed-size window partition schemes (Dosovitskiy et al., 2020, Liu et al., 2021), mac computes self-attention within varied-size windows at a different cascade stage, specifically, a small window for small scale objects as well as a large window for large scale objects . In accordance with the window size, the connection of the three stage follows a tiny-to-large strategy in order to handle the severe scale-variation problem of aerial imagery.

tion of objects within its scale range, i.e., divide-and-conquer (Chen et al., 2021, Jin et al., 2022). From the bottom-up pathway of backbone and top-down pathway of FPN, each level feature can cover a wider scale range of objects. Therefore, rather than matching contextual cues and fusing semantic information between different FPN levels which causes inferior outputs due to the loss of smaller scale information, we argue that feature at each level provides sufficient different scale information and needs to be further devised and explored. From this starting point, we use mac module after each level of FPN output individually. Effectiveness of the proposed methodology is verified by detailed ablation studies and comparison analysis in the experiment part.

On the basis of the cascading spatial-dimension multiple-size window-wise multi-head self-attention and local channel attention, we further exploit the multi-scale feature representation of FPN and propose the MAC model for aerial image segmentation task. Specifically, MAC outperforms the state-of-the-art (SoTA) method PointFlow (Li et al., 2021) by +2.2 % mIoU and +3.1 % mFscore on iSAID dataset, by +2.97 % mIoU on ISPRS Vaihingen dataset. Moreoever, we benchmark the recent state-of-the-art Transformer based semantic segmentation methods on iSAID datasets. The main contributions of our work can be summarized as follows:

1) A Multi-scale Attention Cascade model, a.k.a, MAC, is proposed to solve the aerial image segmentation task in HSR remote sensing imagery.

2) To handle the serious scale variation problem while suppressing the complex and imbalanced background, we propose a generic *mac* module by a cumulating multiple size window-wise self-attention and one local channel attention.

3) Extensive experiments are conducted to verify that the feature of each level of FPN provides sufficient different scale information, which allows further exploiting.

4) We benchmark current Transformer-based methods on iSAID, ISPRS datasets and comparison results show MAC achieves *state-of-the-art* performance.

2 RELATED WORK

CNN Based General Semantic Segmentation. FCN (Long et al., 2015) serves as a milestone in modern semantic segmentation tasks, in which an end-to-end, pixel-to-pixel prediction is produced for the input image. To increase the details of segmentation results, UNet (Ronneberger et al., 2015) and the subsequent SegNet (Badrinarayanan et al., 2017) construct encoder-decoder architectures to achieve high-resolution and semantically meaningful features. The approaches that followed improve the segmentation outputs by leveraging multi-scale contexts aggregation. DeepLab series (Chen et al., 2017, Chen et al., 2018) perform dilated convolution layers to obtains features from various receptive field as well as devises Atrous Spatial Pyramid Pooling (ASPP) for feature fusion, while PSPNet (Zhao et al., 2017) operates pooling at a different grid scale to generate a feature pyramid. Moreover, some recent studies focusing on exploiting the boundary information to improve segmentation outputs (Kirillov et al., 2020, Takikawa et al., 2019). For example, PointRend (Kirillov et al., 2020) iteratively selects uncertain boundary points then, computes point-wise feature representation, and then predicts labels in a coarse-to-fine manner.

FPN (Lin et al., 2017) provides a paradigm for multi-level feature fusion. Recently, various FPN-like model have been proposed to achieve better feature representation. UPerNet (Xiao et al., 2018) specifi-

cally defines semantic levels from the lowest texture level and the middle object level to the highest scene level. Panoptic FPN (Kirillov et al., 2019) designs a semantic segmentation branch at the back of FPN to fuse all level outputs of a feature pyramid into a single output. For general semantic segmentation methods, though multi-scale fusion or boundary modeling is used to obtain finer segmentation outputs, the results are limited due to the great imbalance between foreground and background. In the meantime, intricate background fools their context modeling and cause the resultant inferior performance.

Self-Attention Based Semantic Segmentation. With the magnificent feats achieved by the self-attention mechanism and Transformer (Vaswani et al., 2017) in NLP, further exploration of the Transformer has gradually emerged in vision tasks (Dosovitskiy et al., 2020, Liu et al., 2021, Strudel et al., 2021, Xie et al., 2021, Zheng et al., 2021). For image classification, SoTA performance is shown by ViT (Dosovitskiy et al., 2020), which proposes a pure Transformer model. Inspired by the success of ViT, SETR (Zheng et al., 2021) utilizes ViT as its backbone and a CNN decoder to solve semantic segmentation as a sequence-to-sequence task. Apart from SETR, Segmentor (Strudel et al., 2021) applies a point-wise linear layer after the ViT backbone to produce patch-level class logits. Meanwhile, Segformer (Xie et al., 2021) designs a hierarchical Transformer encoder to achieve a larger receptive field with a light-weight multilayer perceptron to predict a segmentation mask. Recently, with the shift window methodology, Swin Transformer (Liu et al., 2021) provides a SoTA backbone for vision task, in which UperNet (Xiao et al., 2018) is utilized as the segmentation head. Though these Transformer-based methods benefit from obtaining contextual information, their effectiveness for aerial images is limited due to the imbalance between foreground and background and the large scale-variation problem.

Based on self-attention mechanism, some affinity learning method have been proposed. DANet (Fu et al., 2019) presents a dual attention network for scene segmentation, in which a position attention module as well as a channel attention module are designed to distinguish confusing categories. To solve the challenges in aerial image, Farseg (Zheng et al., 2020) and Pointflow (Li et al., 2021) construct FPN-like structure to fuse semantic information between different FPN level. In particular, Farseg proposed a Foreground-Scene Relation Module to align high-level scene feature with low-level relevant context feature. However, the whole spatial contextual cues matching of Farseg lead to loss of small objects and large computation complexity. The baseline of our work is PointFlow, in which object cues are selected via salient and boundary points to handle the disturbance from a complex background. However, it worth noting that details of tiny object and large object boundaries are deteriorated by such a point-selection methodology.

Apart from the aforementioned methods, the design of our proposed MAC is based on two analysis. We believe that the window-wise multi-head self-attention can tackle the imbalanced and complex background in aerial images while obtaining more accurate details for both tiny and large details. In addition, each level of FPN is proved (Chen et al., 2021, Jin et al., 2022) to cover a wider scale range of objects. As a result, compared with matching scale information between different levels of PFN (Li et al., 2021, Zheng et al., 2020) we argue that further exploration at an individual FPN level can benefit from the "divide-and-conquer" merit of Feature Pyramid and achieve better performance on handling scale-variation problem.

3 METHOD

To handle the large scale-variation problem and spreading of tiny objects while overcoming the imbalanced and complex background, we propose a Multi-scale Attention Cascade model (a.k.a, MAC), in which Swin Transformer Tiny, i.e., Swin-T (Liu et al., 2021) and FPN (Lin et al., 2017) serve as backbone and neck, respectively. To further explore the merged pyramid features, at the output of each level of FPN, a multi-scale attention cascade (*mac*) module is implemented, which contains three successive different scales multi-head self-attention (W-MSA cascade) spatially and a local cross-channel attention (channel interaction) for dimension interaction while making homogeneous feature compact. Afterwards, the multi-level features are reshaped into the same size and concatenated together in channel dimension. Finally, a concise segmentation decoder is proposed for interacting feature globally and generate segmentation result. The explicit architecture of the proposed model is demonstrated in Figure. 3.

3.1 FPN-Based Segmentation Framework

In this section, we start with a brief review of FPN (Lin et al., 2017). Given the input image $I \in \mathbb{R}^{H\times W\times 3}$ multi-scale and resolution features $C_i = \{C_2, \cdots, C_5\}$ are generated by the backbone (Simonyan and Zisser-

man, 2014, He et al., 2016, Liu et al., 2021) through a bottom-up pathway. As the neck part, FPN utilizes a lateral connection f_i and up-sampling $Up_{2\times}$ to make the shapes and channels of different feature map consistent. Afterwards, FPN builds a feature pyramid by fusing (adding) the adjacent feature maps together in a pixel-wise manner and propagates them via a top-down pathway. After multi-scale feature fusion, one 3×3 convolution layer is implemented on each merged feature map to solve the aliasing effect. In addition, an extra global context feature is obtained by operating a Pyramid Pooling Module, i.e., *PPM* (Zhao et al., 2017) on C_5. Finally, pyramid features $P_i = \{P_2, \cdots, P_6\}$ with a fixed number of channels (usually 256-D) are generated. The whole process is as follows,

$$\begin{aligned} P_i &= PPM(C_5), \quad i = 6 \\ P_i &= f_i(C_i), \quad i = 5 \\ P_i &= f_i(C_i) + Up_{2\times}(f_{i+1}(C_{i+1})), \quad 2 \le i \le 4 \end{aligned} \tag{1}$$

With such a feature interaction and fusion process, multi-level features can cover various receptive fields so that each level output of FPN (i.e., P_i) contains sufficient context information for a different scale range. Therefore, FPN achieves better feature representations while dividing multi-scale target objects into a multi-scale range to handle them in a divide-and-conquer manner.

3.2 Window-Wise Multi-Head Self-Attention

To calculate local multi-head self-attention (MSA), the input feature map $\mathcal{F} = \mathbb{R}^{h\times w\times C}$ is evenly split into numerous non-overlapping window $M_i = \{M_1, \cdots, M_n\} = \mathbb{R}^{N\times k\times k\times C}$, in which $k\times k$ is the size of each window and $N = h\times w/k^2$ is the number of the windows. Each window is flattened into a 1-D sequence $M_i \in \mathbb{R}^{k^2\times C}$. The first process of window-wise MSA is using linear projection to map M_i and then reshape it into $Q, K, V \in \mathbb{R}^{r\times k^2\times C/r}$, in which Q is query, K is key, V is value and r is number of heads. Furthermore, a relative position bias $B \in \mathbb{R}^{k^2\times k^2}$ is added to capture positional information, then MSA is computed as follows:

$$Attn(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \tag{2}$$

in which $d = C/r$ and $1/\sqrt{d}$ is scale factor. The shape of MSA feature map $Attn \in \mathbb{R}^{r\times k^2\times C/r}$ will be reverse into $Attn' \in \mathbb{R}^{H\times W\times C}$. The outputs $\mathcal{F}' \in \mathbb{R}^{H\times W\times C}$ of window wise MSA are obtained in a residual manner,

$$\mathcal{F}' = \mathcal{F} + Attn' \tag{3}$$

Afterwards the output of window wise MSA is operated by the followed Feed Forward Network, i.e., FFN, the details is shown as follows,

$$\bar{\mathcal{F}} = Mlp(Norm(\mathcal{F}')) + \mathcal{F}' \tag{4}$$

where *Norm* is LayerNorm (LN) (Ba et al., 2016) layer and Mlp consists of two consecutive connections of linear layers and dropout layer. The final output $\bar{\mathcal{F}} \in \mathbb{R}^{H\times W\times C}$ is obtained.

3.3 Multi-Scale Attention Cascade (*mac*) Module

The central idea of MAC is to exploit more detailed scale information in the wide scale range covered by feature pyramid. It is worth noting that, rather than previous affinity learning-based methods that implement additional cross-level interaction, *mac* module is operated at each level of FPN individually. The mac module contains three $Cas_i, i \in \{1,2,3\}$ stages (W-MSA cascade) and a local channel attention (LCA). The whole process is illustrated in Figure. 4.

At each stage of W-MSA cascade, the feature map is spatially divided into different size window, and the size of the window for each stage is $k^i \times k^i = \{2\times2, 4\times4, 7\times7\}$. With these different sized windows, we first compute self-attention in the small area (2×2) and then extend the area gradually into the medium (4×4) and the large (7×7). Given one level pyramid feature $P_i \in \mathbb{R}^{h_i\times w_i\times256}$, it is first fed into one 1×1 convolution layer to reduce dimension into $\bar{P}_i \in \mathbb{R}^{h_i\times w_i\times192}$. The implementation details of W-MSA cascade are as follows,

$$Cas^i_{out} = Cas_3(Cas_2(Cas_1(\bar{P}_i))) \tag{5}$$

where Cas_i denotes the $k^i \times k^i$ window-wise MSA with a FFN and $Cas^i_{out} \in \mathbb{R}^{h_i\times w_i\times192}$. In addition, to achieve channel interaction at each FPN level, local channel attention (LCA) is implemented on Cas^i_{out} to make the homogeneous feature compact along the channel dimension (Jin et al., 2022, Wang et al., 2020), which is shown in Figure. 5.

In LCA, Cas^i_{out} passes through an Adaptive average pooling layer, a *1D Convolution* layer with kernel size=3 and a Sigmoid activate function in sequence. Afterward, with a point-wise multiplication, $Cas_{out^{i'}}$ is obtained. Finally, the output from the LCA will be resized into 1/4 size of input image $I \in \mathbb{R}^{H\times W\times3}$ by bilinear interpolation to generate the multi-scale contextual feature $Cas_{out^{i'}} \in \mathbb{R}^{H/4\times W/4\times192}$ of each FPN level, which is shown as follows, where *Resize* denotes feature resize via bilinear interpolation.

$$Cas_{out^{i'}} = Resize(LCA(Cas_{out^i})) \tag{6}$$

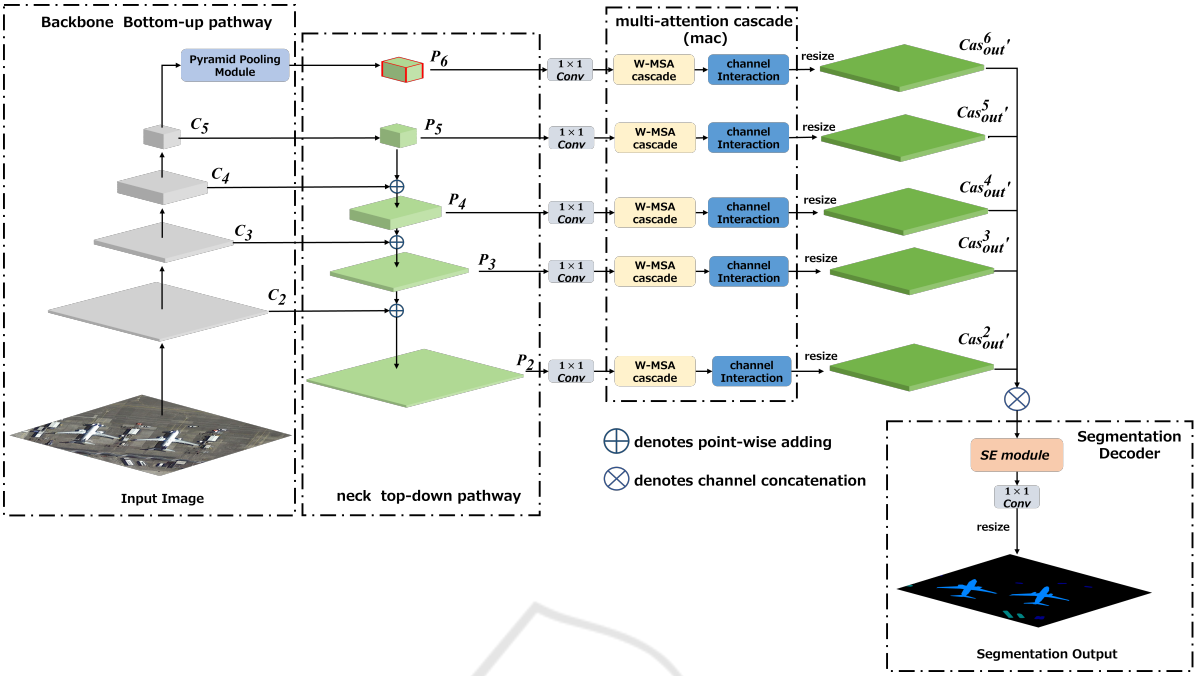


Figure 3: Overview of the proposed MAC model. Notably, the output feature of different FPN level do not have additional cross-level interaction before the segmentation decoder.

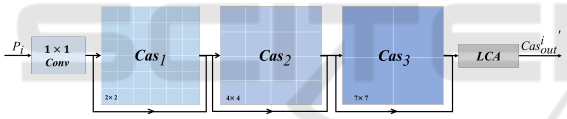


Figure 4: Illustration of detailed process of multi-scale attention cascade (*mac*) module.

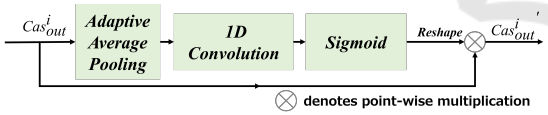


Figure 5: Detailed process of local channel attention (LCA).

3.4 Segmentation Decoder

The segmentation decoder is designed to fuse multi-level feature $Cas_{out^{i'}} \in \mathbb{R}^{H/4\times W/4\times 192}, i = \{2,\cdots,6\}$ together to obtain the final segmentation output $Out \in \mathbb{R}^{H\times W\times class}$, where *class* is the number of categories for segmentation targets. Notably, after implementing mac module on FPN, We argue that there is an overlap in scale between different levels. Therefore, to match the same scale in different level, we first operate channel concatenation to fuse all level features into one feature map. Afterwards, we leverage the Squeeze and Excitation (SE) module (Hu et al., 2018) on it to achieve global channel attention. The final segmentation output is obtained by one 1×1 convolution layer. The process is shown as follows:

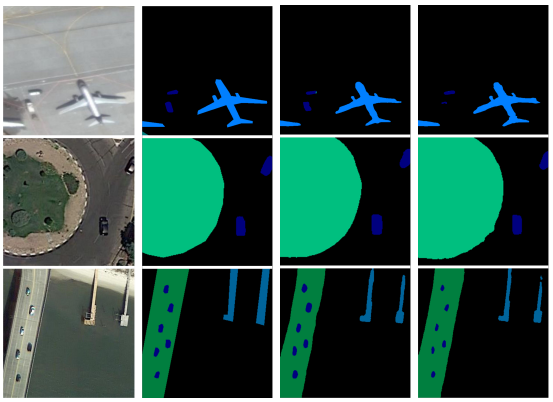
$$Out = Conv(SE(Cas_{out^{2'}} \oplus \cdots \oplus Cas_{out^{6'}}) \quad (7)$$

where $\oplus$ denotes channel concatenation, *SE* denotes implementation of the SE module and *Conv* denotes 1×1 convolution layer.

4 EXPERIMENTS

4.1 Experiment Setting

Datasets. We mainly utlize iSAID dataset (Waqas Zamir et al., 2019) for analysis and evaluation in this work. iSAID dataset consists of 2,806 HSR remote sensing images, in which all of the them are three-channels RGB images. Specifically, iSAID provides 655,451 instances annotations over 15 categories including, *large vehicle, small vehicle, ship, swimming pool, helicopter, plane, store bank, baseball diamond, tennis court, basketball court, ground track field, bridge, roundabout, soccer ball field, harbor*. Mimicking the same manner of PFNet (Li et al., 2021), the original HSR remote sensing images are augmented into 896×896 small images through cropping operations. Afterwards, the augmented images are configured for benchmarking, in which 28,029 images for training and 9,512 images for validation.



(a) Input (b) Label (c) W-PFNet (d) PFNet

Figure 6: **Visualization of preliminary study between PFNet (Li et al., 2021) and W-PFNet**. (a) is the original input image, (b) is the segmentation ground-truth, (c) is output of W-PFNet and (d) is the output of the PFNet model with the Swin-T backbone. For various scale objects, W-PFNet generates more accurate and smoother boundaries prediction than PFNet.

Meanwhile, to verify the generality of our proposal, we further extend the comparison experiment on ISPRS Vaihingen and Potsdam datasets.

Implementation Details and Metrics. We use 4*16GB Tesla V100 GPUs for distributed data parallel (DDP) training. Meanwhile, the inference (evaluation) is implemented with batch-size = 1 on single gpu. In addition, during the training process, batch-size is set to 4 for each GPU. We utilize AdamW as the optimizer by setting learning rate (LR) as 0.00005, betas as (0.9, 0.99) and weight decay as 0.01. Moreover, we adopt Cross-Entropy (CE) loss for loss function computation. In order to evaluate the performance of models and demonstrate numeric results, we adopt mean Intersection over Union (mIoU) and mean F1-score (mFscore) as the metrics in preliminary, comparison and ablation studies.

4.2 Preliminary Study

Table 1: **Results of Preliminary Study.** By replacing the points selection and point based affinity function of PFNet (Li et al., 2021) with window partition and window wise MSA respectively, W-PFNet (Ours) achieves higher mIoU with lower parameters. Furthermore, W-mac (Ours) outperforms other cross-level interaction methods.

Method	mIoU(%)↑	Parameter(M) ↓
PFNet (Li et al., 2021)	67.87	36.63
W-PFNet	68.36	32.10
W-mac	68.57	35.33

In previous affinity learning-based method (Fu et al., 2019, Zheng et al., 2020), semantic information between two different level features $P_i, P_j \in \mathbb{R}^{h\times w\times C}$ of FPN are fused though the following equation,

$$P_i' = A(P_i, P_j)P_i + P_i \tag{8}$$

where A denotes affinity functions and outputs affinity matrix $\in \mathbb{R}^{hw\times hw}$, and P_i' denotes the enhanced feature. In particular, the recent Pointflow (Li et al., 2021), abbreviated as PFNet, designs a salient point selection scheme to reduce the background noise and computation while propagating the contextual information in a top-down manner. The process is as follows,

$$\bar{P}_l = A(\beta(P_l), \beta(P_{l-1}))\beta(P_l) + P_l \tag{9}$$

in which $P_l, P_l - 1 \in \mathbb{R}^{h\times w\times C}$ are two adjacent level feature of FPN, β denotes the point selection scheme of PFNet and $\bar{P}_l$ denotes the enhanced feature. We argue that the point selection strategy weaken the prediction of boundary pixels. Therefore, following the structure of PFNet, we design a preliminary experiment on iSAID dataset to test our hypothesis while verifying the superiority of window wise MSA. The process is replacing the point-selection part of PFNet with a window partition as well as replacing the point-based affinity function A with 4×4 size window-wise MSA, the details are as follows,

$$\tilde{P}_l = w_r(MSA(w_p(P_l), w_p(P_{l-1}))w_p(P_l)) + P_l \tag{10}$$

in which w_p and w_r denotes window partition and its reverse process, respectively, and $\tilde{P}_l$ denotes the enhanced feature through cross-level interaction. We dub the model as W-PFNet for simplicity. To guarantee the fair comparison, we use the same Swin Transformer Tiny, i.e., Swin-T as the backbone. The numerical results of the preliminary study shown in Table.1 demonstrate that the W-PFNet outperfrom the original PFNet with lower parameters. Furthermore, through expanding the details of the prediction mask, the W-PFNet outputs more accurate and smoother boundary predictions on various scale objects, as shown in Figure. 6.

Benefiting from the merits of window-wise MSA, we further extend this preliminary experiment to verify another central idea: without cross-level interaction, better feature representation can be exploited at each FPN level individually. The process is to utilize the aforementioned 4×4 size window-wise MSA to further extract spatial contextual information at each FPN output, which is shown as follows,

$$\tilde{P}_l{}' = w_r(MSA(w_p(P_l), w_p(P_l))w_p(P_l)) + P_l \tag{11}$$

in which $\tilde{P}_l{}'$ denotes the enhanced feature without cross-level interaction. We dub this model as W-mac because its structure could be regarded as using only one cascade stage in *mac* module. The result is

Table 2: Comparison experiment on iSAID dataset.

Method	Backbone	mIoU(%) ↑	mFscore(%) ↑
PSPNet (Zhao et al., 2017)	Res-50 (He et al., 2016)	60.30	73.04
UperNet (Xiao et al., 2018)	Res-50	63.15	75.87
DpLbv3+ (Chen et al., 2018)	Res-50	61.60	74.37
PointRend (Kirillov et al., 2020)	Res-50	66.35	77.20
PFNet (Li et al., 2021)	Res-50	66.85	77.59
MAC	Res-50	**67.24**	**79.33**
PSPNet	Swin-T (Liu et al., 2021)	60.75	73.81
UperNet	Swin-T	66.87	77.43
DpLbv3+	Swin-T	67.32	79.31
PointRend	Swin-T	67.97	79.83
DANet (Fu et al., 2019)	Swin-T	61.00	74.11
PFNet	Swin-T	67.87	79.79
SegFormer (Xie et al., 2021)	MixViT (Xie et al., 2021)	66.20	78.60
MAC	Swin-T	**69.06**	**80.68**

also shown in Table. 1, in which W-mac outperforms the cross-level interaction method. Therefore, we argue for further exploring at each individual FPN level rather than cross-level interaction.

4.3 Comparison Study

General Result. To evaluate the performance of the proposed MAC, the comparison study is shown in Table. 2. Farseg (Zheng et al., 2020) and PFNet (Li et al., 2021) benchmark several CNN-based segmentation methods on iSAID datasets (Waqas Zamir et al., 2019), in which the original PFNet adopt ResNet-50, a.k.a, Res-50 (He et al., 2016) as the backbone. In this comparison experiment, we additionally replace the Res-50 backbone with Swin-T (Liu et al., 2021) to explore the effectiveness of the Transformer backbone in aerial images. Meanwhile, we extend more Transformer-based segmentation methods on iSAID. Mimicking the same manner, we further conduct the comparison experiments on ISPRS Vaihingen and Potsdam datasets to verify the generality of our work, the results are shown in Table. 3.

The optimal performance achieved by the proposed MAC-Swin-T is **+2.2%** mIoU and **+3.1%** mFscore higher than that of the SoTA PFNet-Res-50. Meanwhile, with the same Swin-T as the backbone for fair comparison, our proposed MAC outperforms the previous State-of-the-art (SoTA) method PFNet by **+1.19%** mIoU and **+0.89%** mFscore. In addition, more detailed visualization results demonstrated in Figure. 7 show the superiority of MAC on handling scale-variation and acquiring precise information.

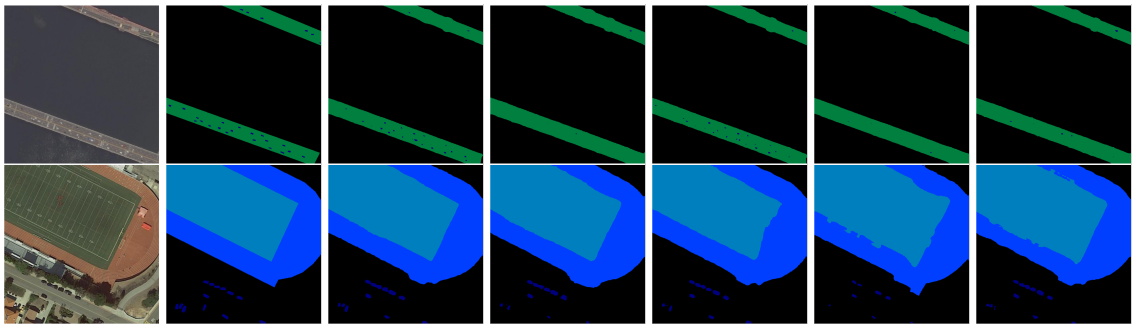
Moreover, as illustrated in Table.3, both of MAC-Res-50 and MAC-Swin-T are **+3%** mIoU higher than baseline on ISPRS Vaihingen dataset. However, compared to SoTA performance on iSAID and Vaihingen, MAC only achieves *SoTA-comparable* performance on ISPRS Potsdam dataset. This is because the scale-variation in Potsdam is imperceptible, strengths of MAC cannot be brought into play. By conducting extensive experiments, we prove the effectiveness of Transformer-based methods on aerial image tasks. We hope this work provides a new benchmark for aerial image segmentation.

Table 3: Comparison with the SoTA methods on ISPRS Vaihingen (left, mIoU-V) and Potsdam (right, mIoU-P) datasets.

Method	Backbone	mIoU-V(%) ↑	mIoU-P(%) ↑
PSPNet	Res-50	65.10	73.90
UperNet	Res-50	66.90	74.30
DpLbv3+	Res-50	64.30	74.10
DANet	Res-50	65.30	74.10
PointRend	Res-50	65.90	72.00
PFNet	Res-50	70.40	**75.40**
MAC	Res-50	**73.37**	73.87
PSPNet	Swin-T	71.58	73.75
UperNet	Swin-T	72.96	74.85
DpLbv3+	Swin-T	72.67	74.18
DANet	Swin-T	71.89	73.79
PointRend	Swin-T	72.53	74.36
PFNet	Swin-T	73.00	75.06
MAC	Swin-T	**73.06**	**75.10**

Table 4: Module Ablation Results on iSAID dataset.

Swin-T	FPN	Cas-WSA	LCA	mIoU(%) ↑
✓	-	-	-	67.35
✓	✓	-	-	67.66
✓	✓	✓	-	68.69
✓	✓	-	✓	68.68
✓	-	✓	✓	68.55
✓	✓	✓	✓	**69.06**



(a) Input Image (b) Groundtruth (c) MAC (d) PFNet (e) DpLbv3+ (f) PiontRend (g) SegFormer

Figure 7: Visualization results on iSAID dataset (Waqas Zamir et al., 2019), in which MAC outperforms other methods including, PFNet (Li et al., 2021), DpLbv3+ (Chen et al., 2018), PointRend (Kirillov et al., 2020), SegFormer (Xie et al., 2021). Except for SegFormer, all of the other models adopt Swint-Transformer Tiny (Swin-T) as backbone. The visualization of segmentation results shows that MAC not only generate accurate tiny objects predictions, but also achieve finer segmentation on large objects boundaries region.

Table 5: Cascade Ablation on iSAID.

Swin-FPN-LCA	+Cas1	+Cas2	+Cas3	mIoU_*DT*(%) ↑	mIoU_*RL*(%) ↑	mIoU_*IS*(%) ↑	mIoU(%) ↑
✓	-	-	-	57.79	63.68	74.25	68.68
✓	✓	-	-	58.50	64.50	73.33	68.65
✓	-	✓	-	58.29	63.41	73.77	68.57
✓	-	-	✓	**59.01**	63.36	74.01	68.89
✓	✓	✓	-	58.30	63.55	73.81	68.62
✓	✓	-	✓	57.92	63.28	**74.73**	68.86
✓	✓	✓	✓	58.65	**63.86**	74.44	**69.09**

4.4 Ablation Study

Based on iSAID (Waqas Zamir et al., 2019), we conduct two ablation experiments to respectively analyze the impact of each component in MAC (as shown in Table. 4) and the contribution of each *Cas* stage in W-MSA cascade (as shown in Table. 5). The ablation study emphasizes the essence of MAC, i.e., *to exploit more detailed scale information in the wide scale range covered by feature pyramid.*

General Ablation. The ablation results for each component of MAC are shown in Table. 4. By comparing line 2 with line 3&4, we find that either W-MSA cascade or LCA can achieve **+1%** mIoU performance based on the naive feature fusion of FPN, which further verifies that sufficient multi-scale information can be exploited and harmonized at each individual FPN level. Moreover, MAC attains the optimal **69.06** mIoU by resorting *mac*, i.e., the combination of W-MSA cascade and LCA. It is worth noting that the mIoU decreases without FPN in line 5, which is resultant because the following *mac* is a growing scrutiny of multi-scale feature fusion obtained through pyramid network structure.

Cascade Ablation. The ablation results for the cascade stage is shown in Table. 5. By keeping the rest of the model, i.e., Swin-FPN-LCA as baseline, we separate the three cascade stage in W-MSA cascade, i.e., Cas_i, where the window size of each stage is $k^i \times k^i = \{2\times 2, 4\times 4, 7\times 7\}$. To demonstrate the effectiveness of different stages on varied scale objects, we categorize dispersed tiny objects (*DT*), regularly shaped large objects (*RL*) and irregularly shaped objects (*IS*) based on the whole 15 classes in iSAID. The cascade ablation results illustrate that each stage of mac is designed to tackle varied scale objects. Specifically, through adding Cas_1, the results on *DT* and *RL* are improved while on *IS* are degraded. Moreover, such a deterioration are tempered by adding Cas_2 and Cas_3 gradually. Finally, by cascading the multi-scale WSA and LCA at each level of FPN, MAC can handle the great scale-variation problem and achieve the optimal.

5 CONCLUSIONS

In this work, we proposed Multi-scale Attention Cascade, a.k.a, MAC, to handle the three predominant issues in aerial image segmentation. On the basis of stacking consecutive multi-size window multi-head self-attention (W-MSA cascade) and local channel at-

tention (LCA), i.e. *mac*, at each level of Feature Pyramid Network (FPN), MAC overcomes the great scale variation and complex background. As a result, numeric and visualization results demonstrate MAC can output accurate predictions on both very tiny and extremely large objects, especially on the ambiguous boundary part. Extensive experiments shows the effectiveness and *state-of-the-art* performance of MAC.

6 DISCUSSIONS

With the triumph achieved by current deep learning and machine learning methods, human can extract the important geo-spatial information from aerial image. In addition, most of the methods are trained and tested in a single domain, i.e., clear weather with adequate illumination. Specifically, iSAID (Waqas Zamir et al., 2019) and ISPRS datasets are leveraged in this work, in which the majority of the data (high-resolution RGB images) is under the aforementioned comfortable condition.

However, the performance of deep learning model is prone to deterioration and even collapse due to domain shift, i.e., domain transferring from one to another. In particular, the changeable weather and illumination are problematic for the model trained under the common domain. Therefore, the data in adverse domain is the desideratum to improve the robustness of the model while such data like aerial images in low-illumination, snowy or foggy weather are difficult to acquire. Therefore, we plan to deploy the current deep learning-based image synthesis and style transfer methodology to augment the aerial image data with different weather and illumination conditions to enhance the model's ability for domain adaptation.

REFERENCES

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(12):2481–2495.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., and Sun, J. (2021). You only look one-level feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 13034–13043.

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

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3146–3154.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7132–7141.

Jin, Z., Yu, D., Song, L., Yuan, Z., and Yu, L. (2022). You should look at all objects. In *Proceedings of the European conference on computer vision (ECCV)*.

Kirillov, A., Girshick, R., He, K., and Dollár, P. (2019). Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 6399–6408.

Kirillov, A., Wu, Y., He, K., and Girshick, R. (2020). Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9799–9808.

Li, X., He, H., Li, X., Li, D., Cheng, G., Shi, J., Weng, L., Tong, Y., and Lin, Z. (2021). 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.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image seg-

mentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272.

Takikawa, T., Acuna, D., Jampani, V., and Fidler, S. (2019). Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5229–5238.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems (NeurIPs)*, 30.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., and Bai, X. (2019). isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 28–37.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2018). Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPs)*, 34:12077–12090.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2881–2890.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 6881–6890.

Zheng, Z., Zhong, Y., Wang, J., and Ma, A. (2020). 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 (CVPR)*, pages 4096–4105.