Squeezing Context into Patches: Towards Memory-Efficient Ultra-High Resolution Semantic Segmentation

Wang Liu[1] , **Puhong Duan**[2] , **Xudong Kang**[2*] and **Shutao Li**[1,2]
[1]College of Electrical and Information Engineering, Hunan University, China
[2]School of Robotics, Hunan University, China
liuwa@hnu.edu.cn, puhong_duan@hnu.edu.cn, xudong_kang@163.com, shutao_li@hnu.edu.cn

Abstract

Segmenting ultra-high-resolution (UHR) images poses a significant challenge due to constraints on GPU memory, leading to a trade-off between detailed local information and a comprehensive contextual understanding. Current UHR methods often employ a multi-branch encoder to handle local and contextual information, which can be memory-intensive. To address the need for both high accuracy and low memory usage in processing UHR images, we introduce a memory-efficient semantic segmentation approach by squeezing context information into local patches (SCPSeg). Our method integrates the processing of local and contextual information within a single-branch encoder. Specifically, we introduce a context squeezing module (CSM) designed to compress global context details into local patches, enabling segmentation networks to perceive broader image contexts. Additionally, we propose a super-resolution guided local feature alignment (LFA) technique to improve segmentation precision by aligning local feature relationships. This approach calculates similarities within sliding windows, avoiding heavy computational costs during the training phase. We evaluate the effectiveness of our proposed method on four widely used UHR segmentation benchmarks. Experimental results demonstrate that our approach enhances UHR segmentation accuracy without incurring additional memory overhead during the inference stage. The code is available at https://github.com/StuLiu/SCPSeg.

1 Introduction

Semantic segmentation is a foundational challenge in computer vision with applications ranging from land cover mapping in remote sensing to object extraction in medical imaging and scene parsing in autonomous driving. Deep learning approaches have shown impressive performance in recent years [Azad *et al.*, 2024; Badrinarayanan *et al.*, 2017]. However, the increasing resolution of images captured by modern sensors presents a significant obstacle to semantic segmenting ultra-high-resolution images (UHRSS) on memory-constrained edge devices [Zhu *et al.*, 2024]. Our focus is on developing an accurate and memory-efficient UHRSS method tailored to the limitations of edge equipment, offering a promising solution for this challenging scenario.

To address the above challenge, the downscaling paradigm is employed, which involves resizing the UHR images to a more manageable size before feeding them into the segmenter. However, this approach leads to coarse predictions due to the loss of fine-grained detail [Minaee *et al.*, 2022]. An alternative is the stride inference paradigm, which works by cropping image patches from the original UHR image and processing these patches individually. While this method preserves more detail, it suffers from insufficient contextual information. Ultimately, both strategies fail to strike a balance between detail preservation and context utilization.

In recent years, various advanced methods have been proposed to address the trade-off between local detail preservation and the need for global context understanding in parsing UHR images. These methods can be categorized into detail refinement [Huynh *et al.*, 2021; Cheng *et al.*, 2020], shallow and deep fusion [Guo *et al.*, 2022; Ji *et al.*, 2023b; Ji *et al.*, 2023a], and cropped global and local fusion [Li *et al.*, 2021; Ding *et al.*, 2022; Zhang *et al.*, 2024; Zhu *et al.*, 2024], as depicted in Figure 1 (a)-(d). In the detail refinement paradigm, a refining module is introduced to correct the coarse predictions obtained from down-scaled images based on either the detailed output [Huynh *et al.*, 2021] or global predictions [Cheng *et al.*, 2020]. However, this approach requires pre-processing the global predictions, leading to a high computational cost. The shallow and deep fusion paradigm utilizes a shallow branch to process the original UHR images and a deep segmenter to parse the down-scaled images. The cropped global and local fusion paradigm parses global patches by a shallow branch and the local patches by a deep branch. Despite these paradigms achieving better performance than generic ones, they still suffer from high memory consumption due to the multi-branch architecture.

In this study, we introduce a novel single-branch UHRSS approach that incorporates context information into patches, illustrated in Figure 1 (d). Our method processes images in a single branch akin to slide inference, compressing redundant context surrounding local patches using a Context Squeez-

*Corresponding Author

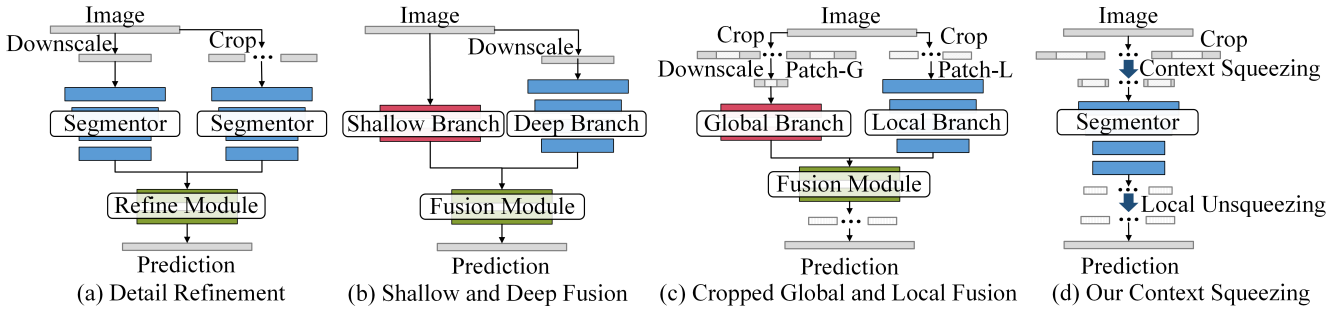

Figure 1: The architectures of the UHRSS methods. (a) Detail refinement methods. (b) Shallow and deep fusion methods. (c) Cropped global and local fusion methods. (d) Our context squeezing method with a single branch.

ing Module (CSM). By slightly downsizing local patches and squeezing in context data efficiently, our CSM achieves a balance between global context awareness and local detail retention. Additionally, we propose an auxiliary super-resolution decoder (SRD) assisted local feature alignment loss (LFA) to enhance segmenter detail features, aiding in preserving local intricacies when analyzing the squeezed global patch. This approach enhances context and detail information without imposing excessive memory demands during testing, elevating segmentation accuracy for UHR image processing compared to conventional methods.

The key contributions of this work are as follows:

- We introduce SCPSeg, a method for semantic segmentation of UHR images that efficiently incorporates context into local patches within a single branch.
- Specifically, we propose a context squeezing module (CSM) to merge global context understanding with local detail retention effectively.
- Furthermore, we employ a multi-task learning strategy and propose a local feature alignment loss to enhance detail preservation within the segmenter.
- Through extensive experiments, SCPSeg demonstrates an exceptional balance between memory efficiency and segmentation accuracy, outperforming conventional semantic segmentation methods without increasing memory demands during testing.

2 Related Works

2.1 Generic Semantic Segmentation

Semantic segmentation has made notable advancements in recent years with the development of convolutional neural networks [He *et al.*, 2016] and vision transformers [Xie *et al.*, 2021]. Most of these methods follow an encoder-decoder architecture [Ronneberger *et al.*, 2015] inspired by the first fully convolutional networks for semantic segmentation [Long *et al.*, 2015]. To enhance segmentation accuracy, several strategies are commonly adopted, including increasing the receptive field [Zhao *et al.*, 2017; Chen *et al.*, 2018], integrating global and local features [Lin *et al.*, 2017; Fu *et al.*, 2019], and maintaining high-resolution branches [Yu *et al.*, 2018; Sun *et al.*, 2019; Zhao *et al.*, 2018]. In efforts to reduce computational costs, numerous real-time semantic segmentation methods have been developed [Pan *et al.*, 2023; Xu *et al.*, 2023]. These methods typically feature a lightweight encoder [Fan *et al.*, 2021; Sandler *et al.*, 2018], an efficient segmentation head [Li *et al.*, 2020; Xie *et al.*, 2021], and sometimes an auxiliary training decoder for detailed prediction [Fan *et al.*, 2021; Zhao *et al.*, 2018]. However, these generic semantic segmentation methods face challenges related to high memory consumption when processing UHR images.

2.2 Semantic Segmentation for UHR Images

Despite the potential of employing slide inference and image down-scaling techniques to alleviate the substantial memory overhead associated with parsing Ultra-High-Resolution (UHR) images, these methods often suffer from decreased accuracy due to the loss of contextual information and detail blurring. To address the challenge of achieving both high accuracy and efficient memory utilization for UHR image segmentation, a set of specialized methods has been developed, primarily falling into two categories: detail refinement [Xia *et al.*, 2016; Cheng *et al.*, 2020; Huynh *et al.*, 2021; Kirillov *et al.*, 2020] and global-local fusion [Zhao *et al.*, 2018; Chen *et al.*, 2019; Li *et al.*, 2021; Guo *et al.*, 2022; Ji *et al.*, 2023b; Ji *et al.*, 2023a]. Detail refinement approaches typically involve segmenting high-resolution scenes in multiple stages. For example, the Cascade Segmentation Refinement model [Cheng *et al.*, 2020] generates detailed results by parsing coarse predictions with high-resolution images. Likewise, MagNet [Huynh *et al.*, 2021] parses high-resolution image patches by integrating low-resolution outputs from preceding stages. However, this iterative refinement paradigm makes them computationally intensive and impractical for real-time applications. Global-local fusion often segments high-resolution images using a multi-branch encoder, where a local branch extracts detailed information, and global patches collect contextual details. For example, recent approaches such as ISDNet [Guo *et al.*, 2022], FCtL [Li *et al.*, 2021], and GINet [Zhu *et al.*, 2024] use lightweight shallow CNNs to extract detailed features in the high-resolution branch and an efficient image encoder for context features in the low-resolution branch. Nevertheless, the incorporation of multi-branches introduces additional computational and memory

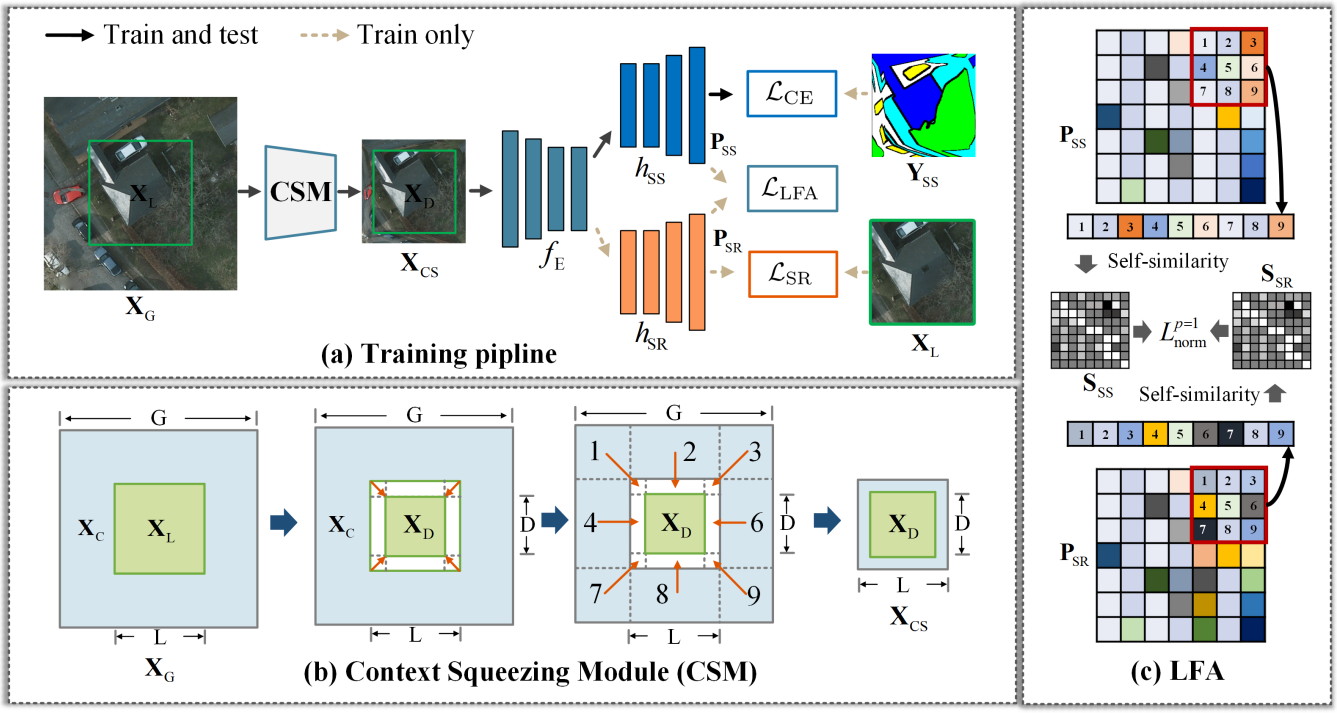

Figure 2: The overview of the proposed SCPSeg for RSISS. (a) The training pipeline. (b) The illustration of the proposed context squeezing module. (c) The illustration of the proposed local feature alignment loss.

costs, posing challenges for deployment on edge devices.

3 Method

This section outlines the architecture and functionalities of the proposed SCPSeg and its components in detail.

3.1 Overview of the Proposed Method

This work presents a novel ultra-high-resolution segmentation approach, named SCPSeg, as illustrated in Figure 2. SCPSeg effectively integrates global context and local details within a single-branch encoder f_{E}, achieving high accuracy with limited memory consumption. The fundamental concept behind SCPSeg involves compressing global context information into local patches, enabling the segmentation network to sense a broader image field. To facilitate this, we employ a context squeezing module (CSM) that strategically aggregates context information into local patches, ensuring effective utilization of both macro-level context and micro-level details. Furthermore, to enhance the preservation of local details, we introduce a super-resolution guided local feature alignment technique (LFA), that helps the segmentation decoder yield detailed predictions. Additionally, the single-branch encoder processes ultra-high-resolution (UHR) images in a sliding window manner. A series of global patches $\mathbf{X}_{\mathrm{G}}$ are cropped from the origin UHR image $\mathbf{X}_{\mathrm{UHR}}$ and fed into the encoder. The overall architecture is optimized using a multi-task learning strategy, which combines segmentation loss $\mathcal{L}_{\mathrm{CE}}$, super-resolution loss $\mathcal{L}_{\mathrm{SR}}$, and a novel local feature alignment loss $\mathcal{L}_{\mathrm{LFA}}$ to ensure comprehensive training and achieve optimal performance.

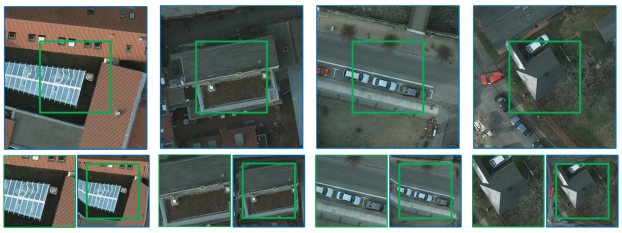

Figure 3: The illustration of several context-squeezing examples on the ISPRS Potsdam datasets. The green boxes indicate the local images. For each example, the top image is the original context image, and the bottom ones are the local patch and squeezed image.

3.2 Context Squeezing Module (CSM)

The CSM is the cornerstone of our proposed SCPSeg framework. Its primary function is to compress and integrate global context information into local patches through a series of interpolations. The CSM operates on the principle of context compression, where redundant pixels in the global context are removed. This process involves the following steps:

Given a square-like global patch $\mathbf{X}_{\mathrm{G}}$ with size of G where its centric is the local patch $\mathbf{X}_{\mathrm{L}}$ with size of L. We first resize the local patch to a down-scaled local patch $\mathbf{X}_{\mathrm{D}}$:

$$\mathbf{X}_{\mathrm{D}} = \nabla\left(\mathbf{X}_{\mathrm{L}}\right) \tag{1}$$

where $\nabla()$ represents the resize operation. Secondly, we squeeze the context information (the blue part in Figure 2 (b)):

$$\mathbf{X}_{\mathrm{C}}^{\mathrm{p1}} = \nabla\left(\mathrm{stack}\left(\mathbf{X}_{\mathrm{C}}^{1}, \mathbf{X}_{\mathrm{C}}^{3}, \mathbf{X}_{\mathrm{C}}^{7}, \mathbf{X}_{\mathrm{C}}^{9}\right)\right) \tag{2}$$

$$\mathbf{X}_{\mathrm{C}}^{\mathrm{p2}} = \nabla\left(\mathrm{stack}\left(\mathbf{X}_{\mathrm{C}}^{2}, \mathbf{X}_{\mathrm{C}}^{4}, \mathbf{X}_{\mathrm{C}}^{6}, \mathbf{X}_{\mathrm{C}}^{8}\right)\right) \tag{3}$$

where $\mathrm{stack}\,()$ indicates the tensor stack operation. Finally, we concatenate the squeezed context $\mathbf{X}_{\mathrm{C}}^{\mathrm{p1}}$, $\mathbf{X}_{\mathrm{C}}^{\mathrm{p2}}$, and the scaled local patch $\mathbf{X}_{\mathrm{L}}^{\mathrm{cs}}$ to get the final squeezed patch $\mathbf{X}_{\mathrm{CS}}$:

$$\mathbf{X}_{\mathrm{CS}} = \mathrm{cat}\left(\mathbf{X}_{\mathrm{C}}^{\mathrm{p1}}, \mathbf{X}_{\mathrm{C}}^{\mathrm{p2}}, \mathbf{X}_{\mathrm{L}}^{\mathrm{cs}}\right) \tag{4}$$

where $\mathrm{cat}\,()$ indicates the tensor concatenate operation. The squeezed context and local patch $\mathbf{X}_{\mathrm{CS}}$ has the same size as the origin local patch $\mathbf{X}_{\mathrm{L}}$.

The final squeezed patches contain both highly squeezed global context and slightly squeezed local details, providing a comprehensive representation of the image. The effectiveness of the CSM lies in its ability to compress context information without losing critical details. In this way, the CSM ensures that the segmenter can perceive a larger image field while maintaining high accuracy. This approach is particularly beneficial for UHR images, where both global context and local details are essential for accurate segmentation.

3.3 Local Feature Alignment (LFA)

To further enhance the preservation of local details, we introduce a super-resolution guided local feature alignment technique. This technique leverages the high-resolution details features in the super-resolution decoder (SRD) h_{SR} to guide the alignment of local features in the segmentation decoder h_{SS}. The primary goal is to ensure precise alignment of local features, leading to more accurate segmentation outputs.

Super-resolution Decoding

For the auxiliary super-resolution task, we utilized several convolutions and de-convolutions to construct the SRD h_{SR}. It decodes the encoded local features and recovers the original high-resolution local patch:

$$\mathbf{P}_{\mathrm{SR}} = h_{\mathrm{SR}}\left(f_{\mathrm{E}}\left(\mathbf{X}_{\mathrm{CS}}\right)\right) \tag{5}$$

where f_{E} is the image encoder. Then, a super-resolution loss is computed between the predicted high-resolution local patch $\mathbf{P}_{\mathrm{SR}}$ and the real high-resolution local patch $\mathbf{X}_{\mathrm{L}}$:

$$\mathcal{L}_{\mathrm{SR}} = \frac{1}{\mathrm{L}\times\mathrm{L}}\sum\nolimits_{i=0}^{\mathrm{L}\times\mathrm{L}} \left\|\mathrm{ccrop}\left(\mathbf{P}_{\mathrm{SR}}^{i}\right) - \mathbf{X}_{\mathrm{L}}^{i}\right\|^{p=1} \tag{6}$$

where $\|\cdot\|^{p=1}$ is the L1 norm. $\mathrm{ccrop}\,()$ indicates the center-crop function for cut out the features of $\mathbf{X}_{\mathrm{L}}$.

Local Feature Aligning

The features of the SRD contain rich structural information of the original image, although they do not explicitly map the categories. Since computing similarity over all the feature pixels will lead to heavy memory and computational cost, we can effectively model the pixel relationship within sliding windows with a size of $k \times k$. These relationships can implicitly deliver semantic information, thus benefiting the task of semantic segmentation.

Firstly, the segmentation features $\mathbf{P}_{\mathrm{SS}}$ and super-resolution features $\mathbf{P}_{\mathrm{SR}}$ are cropped by the slide windows. Then, we compute the self-similarity matrix within the sliding window:

$$\mathbf{S}^{i,j} = \left(\frac{\mathbf{P}^{i}}{\|\mathbf{P}^{i}\|^{p=2}}\right)^{\mathrm{T}} \cdot \left(\frac{\mathbf{P}^{j}}{\|\mathbf{P}^{j}\|^{p=2}}\right) \tag{7}$$

where $\mathbf{S}^{i,j}$ indicates the similarity between pixel i and pixel j within a sliding window. Besides, $\mathbf{P}_{\mathrm{SS}} = h_{\mathrm{SS}}\left(f_{\mathrm{E}}\left(\mathbf{X}_{\mathrm{CS}}\right)\right)$. The decoded features $\mathbf{P}$ can be $\mathrm{crop}(\mathbf{P}_{\mathrm{SS}})$ and $\mathrm{crop}(\mathbf{P}_{\mathrm{SR}})$. At last, to optimize the alignment process, we introduce a local feature alignment loss to align the features by minimizing the similarity matrixes between slide windows at the same location:

$$\mathcal{L}_{\mathrm{LFA}} = \frac{1}{k\times k}\sum\nolimits_{i=0}^{k}\sum\nolimits_{j=0}^{k}\left\|\mathbf{S}_{\mathrm{SR}}^{i,j} - \mathbf{S}_{\mathrm{SS}}^{i,j}\right\|^{p=1} \tag{8}$$

3.4 Optimization

The final objective function is computed as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + w_1 \times \mathcal{L}_{\mathrm{SR}} + w_2 \times \mathcal{L}_{\mathrm{LFA}} \tag{9}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss for segmentation. $\mathcal{L}_{\mathrm{SR}}$ is the super-reolution loss defined in Equation.6. $\mathcal{L}_{\mathrm{LFA}}$ is the local feature alignment loss defined in Equation.8. w_1 and w_2 are set as 0.5 and 0.5, making the gradients of the loss components ranges comparable.

4 Experiments

4.1 Datasets

We conduct experiments on four UHRSS datasets to comprehensively evaluate the effectiveness of our proposed method.

ISPRS Potsdam. This dataset is a land cover mapping dataset collected in an urban area. It contains 38 tiles with the size of 6000×6000. The *train*, *val*, and *test* sets are split to 18, 6, and 14 tiles following [Zhang *et al.*, 2024]. The bands of this dataset include red, green, blue, near-infrared, and DSM information.

BLU. This dataset is collected in urban and rural areas. It contains 4 tiles with the size of 15680×15680. We crop the original URH data into nonoverlapped tiles with resolution 2048×2048 following [Ding *et al.*, 2022]. It is split into 192, 28, and 32 tiles for training, validating, and testing, respectively.

DeepGlobe. This is a land cover mapping dataset collected in both urban and rural areas. It contains 803 tiles with the size of 2448×2448. We split it into 455, 142, and 206 subsets for training, validating, and testing following previous work [Chen *et al.*, 2019].

Inria Aerial. Inria Aerial is a building extraction dataset collected in urban areas. It contains 180 UHR images with 5000×5000 pixels. We split it into 126, 27, and 27 subsets for training, validating, and testing following previous work [Chen *et al.*, 2019].

4.2 Evaluation Metrics

We utilize the mean intersection-of-union (mIoU), mean F1 (mF1), and accuracy (Acc) to assess the effectiveness. The GPU memory cost is monitored by the "gpustat" tool.

Methods	mIoU↑	mF1↑	Acc↑	Mem↓
Generic Methods				
FCN-8s†	83.8	91.0	89.4	2447
DeepLabv3+	83.9	91.1	89.7	5122
PSPNet	82.3	90.0	89.3	6289
ST-UNet	84.5	90.1	-	-
BiTSRS	75.5	-	83.6	-
UNetFormer	84.6	91.6	89.7	4644
GLOTS	82.8	-	-	-
CF-Net	84.1	90.9	89.6	4278
LANet†	83.7	90.7	89.5	2642
UHR Methods				
GLNet	84.0	90.4	85.6	2663
WiCoNet†	84.1	91.2	89.8	2014
TCNet†	85.0	91.7	90.3	2445
SCPSeg (ours)†	86.7	92.7	91.0	**1818**
SCPSeg (ours)	**87.6**	**87.5**	**91.7**	1834

Table 1: Experiment results on ISPRS Potsdam *test* set. † indicates that FCN-8s with a ResNet50 encoder is used as the basic segmenter.

Methods	mIoU↑	mF1↑	Acc↑	Mem↓
Generic Methods				
FCN-8s†	70.1	81.9	86.5	2447
DeepLabv3+	68.2	80.4	86.3	5122
PSPNet	70.4	82.2	86.7	6289
ST-UNet	68.2	80.4	-	-
CF-Net	70.7	82.3	86.9	4278
LANet	70.4	82.1	86.5	2642
UHR Methods				
GLNet	70.5	82.2	-	2663
WiCoNet†	71.0	82.5	87.0	2014
TCNet†	71.6	82.9	87.4	2445
SCPSeg (ours)†	71.1	82.6	87.2	**1818**
SCPSeg (ours)	**71.9**	**83.5**	**87.9**	1834

Table 2: Experiment results on BLU *test* set. † indicates that FCN-8s with a ResNet50 backbone is utilized as the basic segmenter.

4.3 Implementation Details

We utilize Deeplabv3+ with a ResNet-18-d8 as the basic segmenter. All the experiments are conducted in a single Nvidia RTX4090. The encoders were pre-trained on the ImageNet-1K dataset. SGD is used to optimize the neural networks. The learning rate is initially set to 0.01 and decayed by a cosine learning rate policy after each iteration. The training iteration number is set to 40000. The slide window size k in LFA is set to 7. For the ISPRS Potsdam, BLU, and Inria Aerial datasets, we set $\mathrm{G} = 512$, $\mathrm{L} = 256$, and $\mathrm{D} = 192$, respectively. The training batch size is set to 16. For DeepGlobe datasets, we set $\mathrm{G} = 1024$, $\mathrm{L} = 512$, and $\mathrm{D} = 384$, respectively. The training batch size is set to 8.

4.4 Comparing with Other Methods

In this section, we compare our method with other state-of-the-art ones including FCN-8s [Long *et al.*, 2015], DeepLabv3+ [Zhao *et al.*, 2017], PSPNet [Zhao *et al.*, 2017], ST-UNet [He *et al.*, 2022], BiTSRS [Liu *et al.*, 2023a], UNetFormer [Libo *et al.*, 2022], GLOTS [Liu *et al.*, 2023b], CF-Net [Peng *et al.*, 2022], LANet [Ding *et al.*, 2021], GLNet [Chen *et al.*, 2019], WiCoNet [Ding *et al.*, 2022], TCNet [Zhang *et al.*, 2024], U-Net [Ronneberger *et al.*, 2015], ICNet [Zhao *et al.*, 2018], BiseNetv1[Yu *et al.*, 2018], STDC [Fan *et al.*, 2021], CascadePSP [Cheng *et al.*, 2020], PointRend [Kirillov *et al.*, 2020], MagNet [Huynh *et al.*, 2021], ISDNet [Guo *et al.*, 2022], FCtL [Li *et al.*, 2021], WSDNet [Ji *et al.*, 2023b], and GINet [Zhu *et al.*, 2024] on ISPRS Potsdam, BLU, DeepGlobe, and Inria Aerial, in terms of mIoU (%), mF1 (%), OA (%), Mem (M). Some of these methods are generic semantic segmentation methods denoted as "Generic Methods" and the other methods are specially designed for UHR images, denoted as "UHR Methods". The generic methods tagged by "*" are tested in the slide inference mode.

Results on the ISPRS Potsdam Dataset. We compare our SCPSeg with the aforementioned methods on the ISPRS Potsdam testing set. Due to the high inter-class similarity between buildings and impervious surfaces, this dataset poses significant challenges. Table 1 presents the performance of each semantic segmentation method. The experiments demonstrate that our method outperformed all others in terms of mean Intersection over Union (mIoU), mean F1 score (mF1), and overall accuracy (Acc). Specifically, we surpass the UHRSS methods (GLNet, WiCoNet, and TCNet) by substantial margins, clearly demonstrating the effectiveness of our segmentation approach and the improvements in performance. Additionally, we evaluate memory usage during the testing phase, and the experimental results indicate that our approach achieves the highest scores among all models in this regard. With such impressive performance, our method strikes an excellent balance between accuracy and memory cost.

Results on the BLU Dataset. Table 2 presents the performance metrics for both generic and UHRSS methods on the BLU dataset. In all evaluation scenarios, our approach demonstrated the highest accuracy coupled with the lowest memory usage. Overall, the UHRSS methods consistently outperformed the generic methods, highlighting the advantages of utilizing extensive contextual information in semantic segmentation tasks. Notably, our method surpassed the performance of the cropped global and local fusion approach, WiCoNet and TCNet, indicating the effectiveness of our context-squeezing strategy, which achieves superior results with lower memory requirements compared to this architecture.

Results on the DeepGlobe Dataset. As shown in Table 3, we first compare our method with the above-mentioned methods on the DeepGlobe testing set. The WSDNet achieves the best performance because it encodes the whole image which contains all the context information. Our method achieves comparable results to other UHRSS methods. With such impressive performance, our method is economical in the mem-

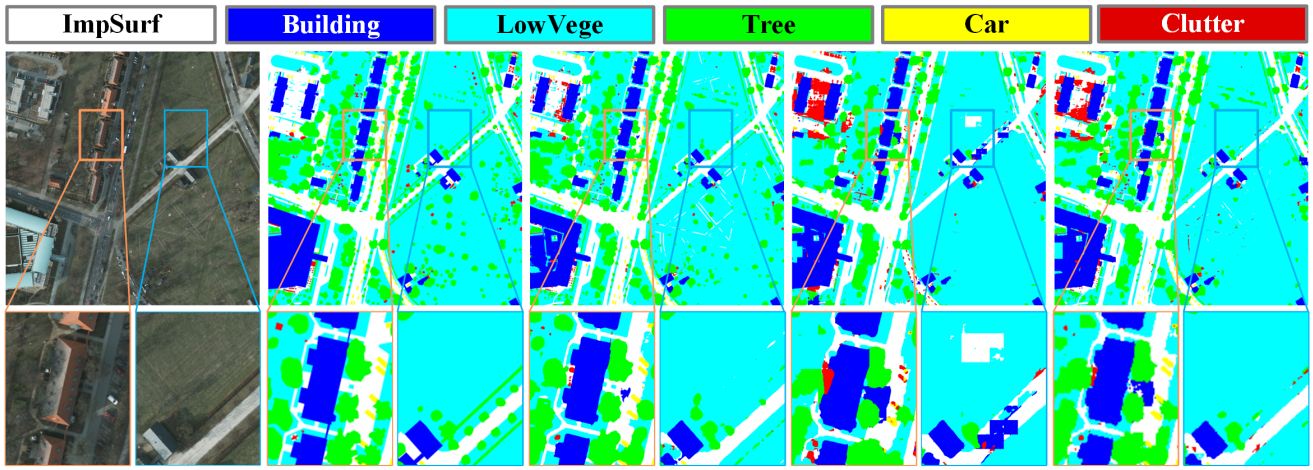

Figure 4: Illustration of the ISPRS Potsdam validating set, compared with the other methods. '*' indicates the slide inference mode.

Methods	mIoU↑	mF1↑	Acc↑	Mem↓
Generic Methods				
U-Net	38.4	-	-	5507
U-Net∗	37.3	-	-	949
ICNet	40.2	-	-	2557
PSPNet	56.6	-	-	6289
DeepLabv3+	63.5	-	-	3199
DeepLabv3+∗	63.1	-	-	1279
FCN-8s	68.8	79.8	86.2	5227
FCN-8s∗	71.8	82.6	87.6	1963
BiseNetv1	53.0	-	-	1801
DANet	53.8	-	-	6812
STDC	70.3	-	-	2580
UHR Methods				
CascadePSP	68.5	79.7	85.6	3236
PPN	71.9	-	-	**1193**
PointRend	71.8	-	-	1593
MagNet	72.9	-	-	1559
GLNet	71.6	83.2	88.0	1865
ISDNet	73.3	84.0	88.7	1948
FCtL	73.5	83.8	88.3	3167
WSDNet	**74.1**	**85.2**	**89.1**	1876
GINet	73.5	-	-	4086
SCPSeg (ours)	74.0	84.0	88.1	2124

Table 3: Experiment results on DeepGlobe *test* set. '*' indicates the slide inference mode when testing the generic methods.

ory cost, attaining a satisfactory balance between effectiveness and memory efficiency.

Results on the Inria Aerial Dataset. We present a comparison using the Inria Aerial test dataset in Table 4. This dataset is particularly challenging, featuring images with approximately 25 million pixels, around four times that of the DeepGlobe dataset. The substantial size of this UHR dataset places significant demands on memory usage for semantic segmentation tasks. Our experimental results demonstrate that SCPSeg significantly outperforms other UHRSS methods across all accuracy metrics while maintaining superior memory efficiency.

Methods	mIoU↑	mF1↑	Acc↑	Mem↓
Generic Methods				
DeepLabv3+	55.9	-	-	5122
FCN-8s	69.1	81.7	93.6	2447
STDC	72.4	-	-	7410
UHR Methods				
CascadePSP	69.4	81.8	93.2	3236
GLNet	71.2	-	-	2663
ISDNet	74.2	84.9	95.6	4680
FCtL	73.7	84.1	94.6	4332
WSDNet	75.2	86.0	96.0	4379
GPWFormer	76.5	86.2	**96.7**	4710
GINet	76.6	-	-	4086
SCPSeg (ours)	**78.6**	**88.0**	95.8	**1818**

Table 4: Experiment results on Inria Aerial *test* set.

4.5 Ablation Study

In this section, we investigate the proposed modules and demonstrate their effectiveness. All the ablation studies are performed on the ISPRS Potsdam validating set.

Effectiveness of the SCPSeg Components. We conduct experiments to verify the effectiveness of the proposed SCPSeg based on various basic segmenters, such as FCN-8s [Long *et al.*, 2015], Deeplabv3+ [Chen *et al.*, 2018], PSPNet [Zhao *et al.*, 2017], SegFormer [Xie *et al.*, 2021], TopFormer [Zhang *et al.*, 2022], DDRNet [Pan *et al.*, 2023], PIDNet [Xu *et al.*, 2023]. The experiment results show that the segmen-

Basic Segmenter	FCN-8s	Deeplabv3+	PSPNet	SegFormer	TopFormer	LANet	DDRNet	PIDNet
w/o SCPSeg	84.5	86.5	84.7	84.9	84.7	84.9	85.2	86.1
w SCPSeg	86.2	87.8	85.8	85.3	86.1	86.1	86.3	87.6

Table 5: Ablation study for the effectiveness of incorporating our method.

Segmenter	CSM	SRD	LFA	mIoU	Mem
FCN-8s (ResNet50)				84.5	1818
	✓			85.6	1818
	✓	✓		85.9	1818
	✓	✓	✓	**86.2**	1818
Deeplabv3+ (ResNet18)				86.5	1834
	✓			87.4	1834
	✓	✓		87.7	1834
	✓	✓	✓	**87.8**	1834

Table 6: Ablation study for the key components of our SCPSeg on the ISPRS Potsdam validating set.

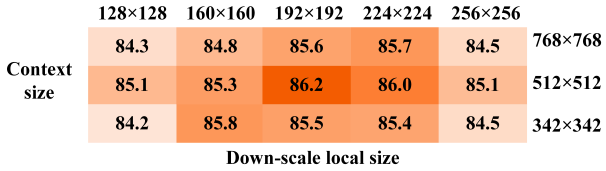

Figure 5: Illustration of the different context size and down-scaled local size versus mIoU on the ISPRS Potsdam validating set.

tation accuracy (mIoU) improves around 1.2% after incorporating the proposed method, which indicates the effectiveness of our SCPSeg. We also verify the effectiveness of different components in SCPSeg, as shown in Table 6. Firstly, incorporating CSM leads to considerable performance gain, indicating the effectiveness of the context-squeezing paradigm. Secondly, the auxiliary SRD improves the performance of the segmentation task slightly, which proves that super-resolution guided multi-task learning is useful for parsing down-scaled local images. Thirdly, a considerable improvement is gained after the utilization of LFA because the local feature alignment loss helps transfer mutual knowledge from different decoders. Lastly, the memory footprint is not increased, indicating our method is memory-efficient in the testing stage.

Influence of Window Size. We conduct experiments to analyze the influence of the context window size G and down-scaled local size D. As Figure 5 shows, the best performance gains when the context size is set to 512×512 and the down-scaled local size is set to 192×192. As context size increases, the performance decreases because squeezing too much context information into a fixed area results in context degradation. Moreover, leveraging a few context information improves performance only marginally. As the down-scaled local size increases, the squeezed context information decreases, which results in a decrease in performance.

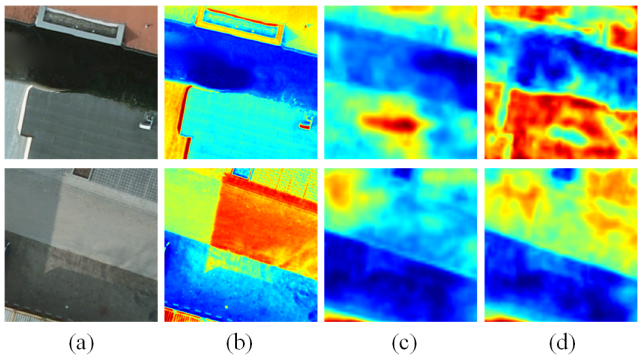

Figure 6: The feature visualization of the semantic segmentation and super-resolution branches. (a) Input images. (b) Features of the super-resolution branch. (c) Features of the segmentation branch before aligning. (d) Features of the segmentation branch after aligning.

As the down-scaled local size decreases, the local images become blurred, which leads to poor segmentation performance. Therefore, a moderate context window size and a properly down-scaled local size are important to balance the global information and local details.

Visualization of Features. To show the effectiveness of the proposed local feature alignment, we visualize the features of super-resolution and semantic segmentation branches in Figure 6. By comparing (b) and (c) in Figure 6, we can easily find that the super-resolution branches contain more complete structure information of objects. The locally aligned semantic features in Figure 6 (d) have clearer edges than the unaligned features. It indicates that the local-consistent relationship between super-resolution features can be effectively transferred to the semantic segmentation branch, thus benefiting the task of semantic segmentation.

5 Conclusion

We proposed a memory-efficient semantic segmentation method, SCPSeg, for the UHR images. It leverages both the global context and local fine structure effectively to enhance the segmentation in the scenario of ultra-high resolution without sacrificing the GPU memory usage. Our work provides a bright new paradigm for parsing UHR images where squeezing redundant context information is useful. Extensive experiments indicate that our method provides a good trade-off between accuracy and memory efficiency. In the future, we plan to extend SCPSeg to other high-resolution vision tasks, such as object detection and instance segmentation, to further validate its generalizability. Besides, adaptive mechanisms to dynamically adjust memory usage based on image complexity will be investigated.

Acknowledgments

We sincerely thank the reviewers for their valuable and constructive comments, which have greatly improved this article. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFA0715203; in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 62221002; in part by the National Natural Science Foundation of China under Grant 62201207 and 62371185; in part by the Natural Science Foundation of Hunan Province under Grant 2023JJ40163; in part by the Science and Technology Innovation Program of Hunan Province under Grant 2023RC3124 and 2024RC1030.

References

[Azad *et al.*, 2024] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024.

[Badrinarayanan *et al.*, 2017] 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.

[Chen *et al.*, 2018] 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*, September 2018.

[Chen *et al.*, 2019] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[Cheng *et al.*, 2020] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[Ding *et al.*, 2021] 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, 2021.

[Ding *et al.*, 2022] Lei Ding, Dong Lin, Shaofu Lin, Jing Zhang, Xiaojie Cui, Yuebin Wang, Hao Tang, and Lorenzo Bruzzone. Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[Fan *et al.*, 2021] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9725, June 2021.

[Fu *et al.*, 2019] 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*, June 2019.

[Guo *et al.*, 2022] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Guannan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, and Ke Xu. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4351–4360, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[He *et al.*, 2022] 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.

[Huynh *et al.*, 2021] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16755–16764, June 2021.

[Ji *et al.*, 2023a] Deyi Ji, Feng Zhao, and Hongtao Lu. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation. In *International Joint Conference on Artificial Intelligence*, page 920–928, 2023.

[Ji *et al.*, 2023b] Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23621–23630, 2023.

[Kirillov *et al.*, 2020] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[Li *et al.*, 2020] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Proceedings of the European Conference on Computer Vision*, volume 12346, pages 775–793, 2020.

[Li *et al.*, 2021] Qi Li, Weixiang Yang, Wenxi Liu, Yuanlong Yu, and Shengfeng He. From contexts to local-

ity: Ultra-high resolution image segmentation via locality-aware contextual correlation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7252–7261, October 2021.

[Libo *et al.*, 2022] Wang Libo, Li Rui, Zhang Ce, Fang Shenghui, Duan Chenxi, Meng Xiaoliang, and M. Atkinson Peter. 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.

[Lin *et al.*, 2017] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, jul 2017.

[Liu *et al.*, 2023a] Yuheng Liu, Yifan Zhang, Ye Wang, and Shaohui Mei. Bitsrs: A bi-decoder transformer segmentor for high-spatial-resolution remote sensing images. *Remote Sensing*, 15:840, 2023.

[Liu *et al.*, 2023b] Yuheng Liu, Yifan Zhang, Ye Wang, and Shaohui Mei. Rethinking transformers for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2015.

[Minaee *et al.*, 2022] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[Pan *et al.*, 2023] Huihui Pan, Yuanduo Hong, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3448–3460, 2023.

[Peng *et al.*, 2022] Chengli Peng, Kaining Zhang, Yong Ma, and Jiayi Ma. Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[Sandler *et al.*, 2018] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[Sun *et al.*, 2019] 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 5686–5696, 2019.

[Xia *et al.*, 2016] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *Proceedings of the European Conference on Computer Vision*, pages 648–663, June 2016.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021.

[Xu *et al.*, 2023] Jiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19529–19539, June 2023.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[Zhang *et al.*, 2022] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, June 2022.

[Zhang *et al.*, 2024] Li Zhang, Zhenshan Tan, Guo Zhang, Wen Zhang, and Zhijiang Li. Learn more and learn usefully: Truncation compensation network for semantic segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, July 2017.

[Zhao *et al.*, 2018] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[Zhu *et al.*, 2024] Peng Zhu, Xiangrong Zhang, Xiao Han, Puhua Chen, Xu Tang, Xina Cheng, and Licheng Jiao. High-resolution remote sensing image segmentation with global-guided normalization and local affinity distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.