forests

MDPI

Article

An Efficient and Light Transformer-Based Segmentation Network for Remote Sensing Images of Landscapes

Lijia Chen [1], Honghui Chen [2], Yanqiu Xie [1], Tianyou He [1], Jing Ye [1] and Yushan Zheng [1,*]

[1] College of Landscape Architecture, Fujian Agriculture and Forest University, Fuzhou 350002, China
[2] Department of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China
* Correspondence: zys1960@163.com

Abstract: High-resolution image segmentation for landscape applications has garnered significant attention, particularly in the context of ultra-high-resolution (UHR) imagery. Current segmentation methodologies partition UHR images into standard patches for multiscale local segmentation and hierarchical reasoning. This creates a pressing dilemma, where the trade-off between memory efficiency and segmentation quality becomes increasingly evident. This paper introduces the Multilevel Contexts Weighted Coupling Transformer (WCTNet) for UHR segmentation. This framework comprises the Mult-level Feature Weighting (MFW) module and Token-based Transformer (TT) designed to weigh and couple multilevel semantic contexts. First, we analyze the multilevel semantics within a local patch without image-level contextual reasoning. It avoids complex image-level contextual associations and eliminates the misleading information carried. Second, MFW is developed to weigh shallow and deep features for enhancing object-related attention at different grain sizes from multilevel semantics. Third, the TT module is introduced to couple multilevel semantic contexts and transform them into semantic tokens using spatial attention. Then, we can capture token interactions and obtain clearer local representations. The suggested contextual weighting and coupling of single-scale patches empower WCTNet to maintain a well-balanced relationship between accuracy and computational overhead. Experimental results show that WCTNet achieves state-of-the-art performance on two UHR datasets of DeepGlobe and Inria Aerial.

Keywords: ultra-high-resolution image; segmentation quality; multilevel semantic contexts; transformer

Citation: Chen, L.; Chen, H.; Xie, Y.; He, T.; Ye, J.; Zheng, Y. An Efficient and Light Transformer-Based Segmentation Network for Remote Sensing Images of Landscapes. *Forests* **2023**, *14*, 2271. https://doi.org/10.3390/f14112271

Academic Editors: Giovanni D'Amico, Walter Mattioli and Gherardo Chirici

Received: 8 October 2023
Revised: 16 November 2023
Accepted: 17 November 2023
Published: 20 November 2023

1. Introduction

With the rapid progress in remote sensing technology, the acquisition of satellite images and the surge in data availability have unveiled new opportunities for the computer vision community. Ultra-high-resolution (UHR) imagery [1] (i.e., 2K, 4K, or even higher-resolution images) acquired by low-orbit satellites and unmanned aerial vehicles (UAVs) [2] has driven the development of remotely sensed imagery analysis since it allows for a more comprehensive characterization of the ground surface compared to ordinary sensor data. It encounters diverse imaging applications, including but not limited to high-resolution geospatial image analysis, urban planning, and land use and land cover (LULC) [3], as well as land resource management [1].

UHR image classification is often implemented based on semantic segmentation, which is the process of labeling each pixel in an image as a different semantic category [4]. Compared with local tasks of structured output such as anchor-based object detection and classification, semantic segmentation not only understands the location of each object or scene in the image but also effectively confirms the boundaries by labeling closed object categories pixels. However, generalized semantic segmentation models [4] that work on full-resolution images and perform intensive prediction are not suitable for UHR images. The escalating image resolution demands increased computational resources, given its dependence on expansive receptive domains and intricate deep features [5,6] or

graph modules [7–9]. This creates a challenging trade-off between memory efficiency and segmentation quality.

As for UHR image segmentation, recent popular practices are categorized into two types: (i)—the input is downsampled to smaller spatial dimensions before performing segmentation; (ii)—partitioned patches are segmented individually and their results are merged into high-resolution patches. The first scheme sacrifices segmentation quality to improve model efficiency. UHR images encompass numerous objects/regions characterized by substantial variations in scale and shape. The segmentation model must not only grasp the semantics of extensive image regions but also discern image details across diverse granularities.Therefore, examples of the the second scheme are widely respected, such as GLNet [10] and FCtL [11]. They conducted multiple predictions on the patches, which constrained the overall inference speed. To improve it, ISDNet [12] abandons deep branching inference in favor of direct downsampling while WSDNet [13] naturally integrates the multilevel discrete wavelet transform (DWT) and inverse discrete wavelet transform (IWT). However, such approaches do not inherently improve the pressure of inference and training due to hierarchical inference.

To alleviate this issue, this paper proposes the Multilevel Contexts Weighted Coupling Transformer (WCTNet), which effectively merges the localized results delineated by the hierarchical inference into high-resolution semantic masks. Contextual information is an effective guidance in analyzing semantic regions with large-size contrast [10], which are used to construct multiscale context regions for association and hierarchical reasoning in many approaches [10–14]. Considering the different scales of contexts in local patches are the multilevel semantics, we propose to analyze multilevel features within local patches. First, we introduce the Multilevel Feature Weighting (MFW) module to deal with the multilevel semantic contexts. MFW analyzes the relationship between shallow and deep features by weighting the features that highlight the attention regions related to objects at different grain sizes. With previous knowledge, shallow texture features have a prominent contribution to distinguishing boundaries. Additionally, the processing of single-scale patches eliminates misleading information from multiscale contexts while avoiding the complex process of context–region correlation. Second, we introduce a Token-based Transformer (TT) to couple multilevel semantic contexts. Adding or concatenating semantics cannot accurately capture the correlation between multilevel contexts due to the redundancy associated with high-level features. TT uses spatial attention to transform the contextual semantics into a set of semantic tokens. These tokens are then provided to the self-attention module of the transformer [15] to capture token interactions. The generated visual tokens produce clearer local representations to avoid the distraction of redundant information. The contextual weighting and coupling of single-scale patches, facilitated by deep sharing interactions, empower WCTNet to effectively strike a balance between accuracy and computational overhead.

The contributions of this paper can be summarized as follows:

- This paper proposes a novel local patch segmentation network WCTNet. The proposed model avoids the complex process of context region association while eliminating the misleading information brought by multiscale contexts.
- MFW is proposed for weighting multilevel semantic contexts. Relationships between shallow and deep features are captured to highlight object-related regions of attention at different grain sizes. Further, TT module is introduced to couple multilevel semantic contexts. Spatial attention is applied to transform multilevel contexts into compact semantic tokens and self-attention is used to capture the correlations between tokens.
- Experimental results demonstrate the superiority of WCTNet, which achieves state-of-the-art performance on two UHR datasets, including DeepGlobe [16] and Inria Aerial [17].

2. Related Work

2.1. Semantic Segmentation

The rapid development of deep learning [18–22] has significantly improved semantic segmentation, which requires fine-grained labeling of the image at the pixel level. Fully convolutional networks (FCN) [4] were the first CNN model to be used for semantic segmentation. Most of the subsequent generalized models are based on FCN to make improvements. U-Net [23] enriched the response features by the cross-layer fusion of multilevel features and jump connections. Similar to U-Net, Vijay et al. constructed codec networks to map low-level coded features to full input features to achieve pixel-by-pixel classification [24,25]. However, these models have very high GPU memory requirements for UHR images. They rely on large receiver domains and fine-grained deep features [5] or graph modules [7].

Researchers further balanced computational cost and performance [26–30]. ENet [26] and ICNet [28] reduced GPU memory through model compression. However, they are not effective on UHR images that contain much complex detail, since model compression does not maintain the complex feature representation of the original model. BiseNetV2 [29] designed bilateral aggregation and enhanced training strategies to improve performance. However, the feature representation is sacrificed by relying on small receiver domains and feature channel cuts. Additionally, knowledge distillation had been used to balance the accuracy and speed of segmentation models [31]. And then, the scheme of merging local patches gradually became popular for UHR images. GLNet [10] constructed dual branches to aggregate global and local information to improve the correlation. CPNet [32] designed a contextual prior layer to model intra- and interclass correlations. FCtL [11] proposed a location-aware context fusion to capture the location correlation of local patches and contexts. ISDNet [12] constructed bilateral models to feed shallow and deep branches with different sizes of inputs. WSDNet [13] applied natural integration of multilevel discrete wavelet transforms to release the computational burden.

2.2. The Fusion of Contextual Information

Contextual information includes image-level and semantic-level context. The context in visual tasks plays a key role in encoding local spatial neighborhoods or even nonlocal information [10,33,34]. Segmentation can be refined by integrating high-level and low-level features to capture semantic context at different grain sizes [35–38]. RefineNet [39] used multipath refinement blocks to fuse multiscale semantic contexts. The Laplace pyramid was used to refine boundaries reconstructed from low-resolution graphs [40]. The Feature Pyramid Network (FPN) [41] aggregated multilevel features after upsampling in a top-down manner. ParseNet [42] used global pooling to aggregate different levels of context for scene parsing. DeepLab [33] used dilated convolution and atrous spatial pyramids to expand the sense field. Multilevel context information has been used to aggregate global context and high-resolution details [39–41]. Image-level semantics were proposed to enhance the relevance of local and global semantics. GLNet [10] fused global and local contexts to enhance aggregation. CPNet [32] embedded a priori into the network to model intra- and inter-class dependencies. FCtL [11] proposed a location-aware context fusion to capture the positional relevance of local patches and contexts. ISNet [12] integrated image-level and semantic-level contexts to improve pixel representation.

Unlike previous work, we fuse contextual information in two ways. We consider semantic contexts since image-level contexts carry more redundant information compared to semantic contexts. The semantic context on localized patches can effectively balance accuracy and complexity. A dilemma arises between the complex feature processing used for aggregate image-level context and the slight performance gain. First, MFW is used to weigh shallow and deep features to highlight the attention of the region associated with objects in the image at different grain sizes, which can enhance training efficiency. Second, the self-attention mechanism in the transformer can capture the dependencies between different positions in a sequence, especially for long sequences. Therefore, this mechanism can be used to deal with UHR images to capture the correlation between feature

sequences without introducing image-level context. The performance can be improved while mitigating redundant feature processing during local patch aggregation.

3. Materials and Methods

3.1. Datasets

DeepGlobe [16]: This dataset contains 803 UHR images (2448 × 2448 pixels). The images are randomly divided into training, validation, and test sets in a 8:1:1 ratio. The dense annotation contains seven types of landscape regions: cyan for "city"; yellow for "agriculture"; purple for "rangeland"; green for "forest"; blue for "water"; white for "barren"; and one of the seven categories not considered in the challenge, called the "unknown" region.

Inria Aerial [17]: This dataset covers a diverse range of urban landscapes, from dense metropolitan areas to high mountains. It contains 180 UHR images (from five cities) with 5000 × 5000 pixels. We randomly divided the images into training, validation, and test sets in a ratio of 8:1:1, respectively. Each image is annotated using a binary mask for constructed/nonconstructed regions.

3.2. Pipeline

The main concept of this paper follows the design of merging the segmentation of local patches into panoramic high-resolution results, as shown in Figure 1, given an ultra-high-resolution image I with width W and height H. We uniformly divide it into N localized patches $I_k(k = [1, ..., N], I_k \in I)$ with width w and height $h(w < W, h < H)$ along the row and column axes. Next, the proposed WCTNet performs the segmentation of each localized patch. Then, we merge the localized results into one segment as the final high-resolution segmentation mask. WCTNet contains two modules i.e., MFW and TT, whose core idea is to utilize multilevel semantic context to analyze regions of high-size contrast within local patches. In particular, we only analyze multilevel features within a local patch region. The processing of single-scale patches eliminates misleading information from multiscale contexts while avoiding the complex process of context–area correlation. Figure 2 illustrates the framework of WCTNet. First, local patches of width w and height h are fed into the backbone network for feature extraction. The local patches are encoded as four-level multiscale semantic contexts f_1, f_2, f_3, f_4. Second, MFW weights and fuses multilevel semantic contexts to capture the relationship between shallow features and deep features. Third, we introduce TT to couple the output of MFW considering that the transformer can effectively capture the dependencies between different positions in the sequence. Correlations between multilevel contexts are further captured. The generated visual tokens produce clearer local representations to avoid the interference of redundant information. Finally, the enhanced semantic contexts are fused and up-sampled to obtain local segmentation masks. Next, we present the details of MFW and TT, respectively.

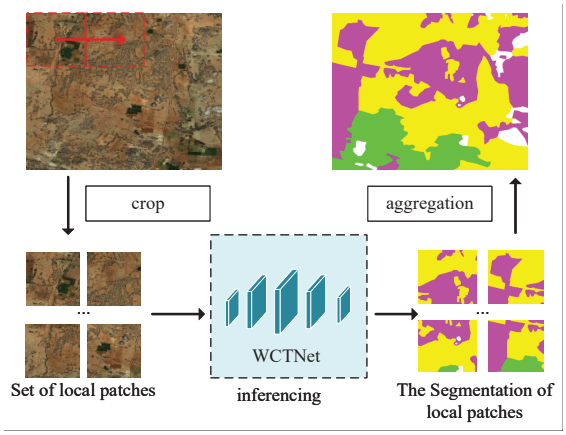

Figure 1. The main process of UHR image segmentation, which includes local block collection cropped from the image; Local patch inference based on WCTNet; And combination of the local segmentation results into a high-resolution mask.

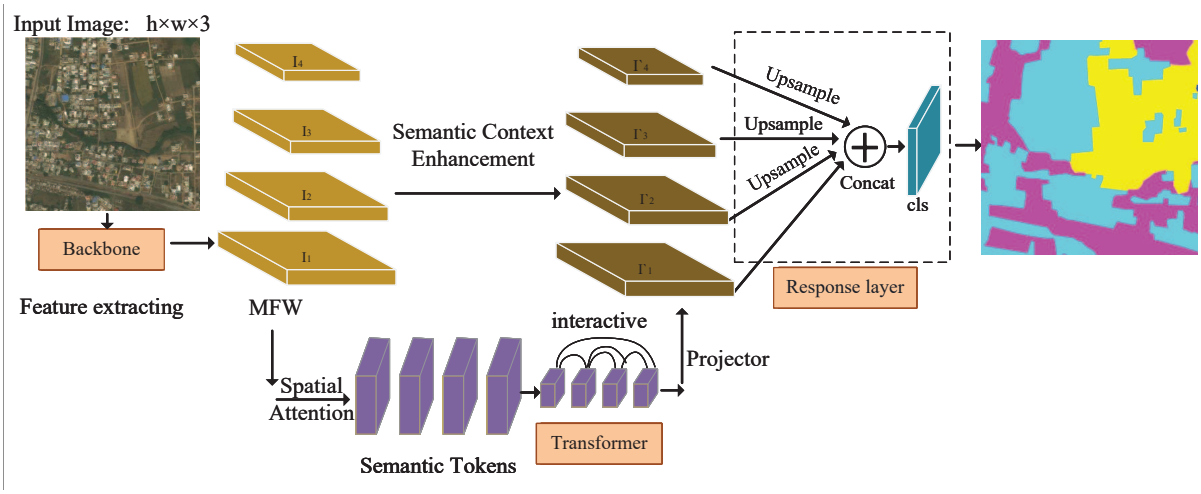

Figure 2. The pipeline of the WCTNet: MFW weights and fuses four levels of multiscale semantic contexts extracted by the backbone network. TT couples the output of MFW for semantic context enhancement. The enhanced semantic context is fused and upsampled in the response layer to obtain the local segmentation mask.

3.3. MFW Module

Image-level contexts at different scales within a localized patch are essentially multilevel semantics. The integration of multilevel semantic contexts aims to amalgamate features at various resolutions, imparting unequal contributions to the response output. Proper feature fusion can improve the efficiency of the model. MFW incorporates an extra weight for the semantic context, enabling the network to discern and learn the significance of each input. Figure 3 shows the structure of MFW, which weights multilevel semantic context f_1, f_2, f_3, f_4 and generates I_1, I_2, I_3, I_4. The weights consists of three 2D sets, (w_1, w_2, w_3); and three 3D sets, (w_4, w_5, w_6), normalized by:

$$\sum_{d=0}^{1}(w_i{}^d) + \epsilon = 1, \tag{1}$$

and

$$\sum_{d=0}^{2}(w_j{}^d) + \epsilon = 1, \tag{2}$$

where $w_i{}^d$ represents the i-th 2D weight, $i = 1, 2, 3$. $w_j{}^d$ represents the j-th 3D weight, $j = 4, 5, 6$. $\epsilon = 0.0001$ is a small value to avoid numerical instability. We provide a concrete example to illustrate the procedure of weighted features. Here, we elucidate the generation process of the feature I_1 when weighted by other features f_1, f_2, f_3, f_4 in Figure 3:

$$F_0 = conv(\text{swish}(Upsample(H_2)w_0^{*0} + I_0 w_0^{*1} + Upsample(I_1)w_0^{*2})), \tag{3}$$

where

$$H_2 = conv(\text{swish}(I_1 w_2{}^0 + Upsample(H_1)w_2{}^1)), \tag{4}$$

where $conv(\cdot)$ represents a convolutional op with $conv_{1\times1}(\cdot)$ and $conv_{3\times3}(\cdot)$ for feature processing, and $conv_{n\times n}(\cdot)$ is a depth separable convolutional op with a convolution kernel size of $n \times n$. $Upsample(\cdot)$ is an upsampling op for resolution matching; swish$(\cdot)$ is the activation function. Then, the weighted semantic contexts I_1, I_2, I_3, I_4 are fed into TT for relevance enhancement.

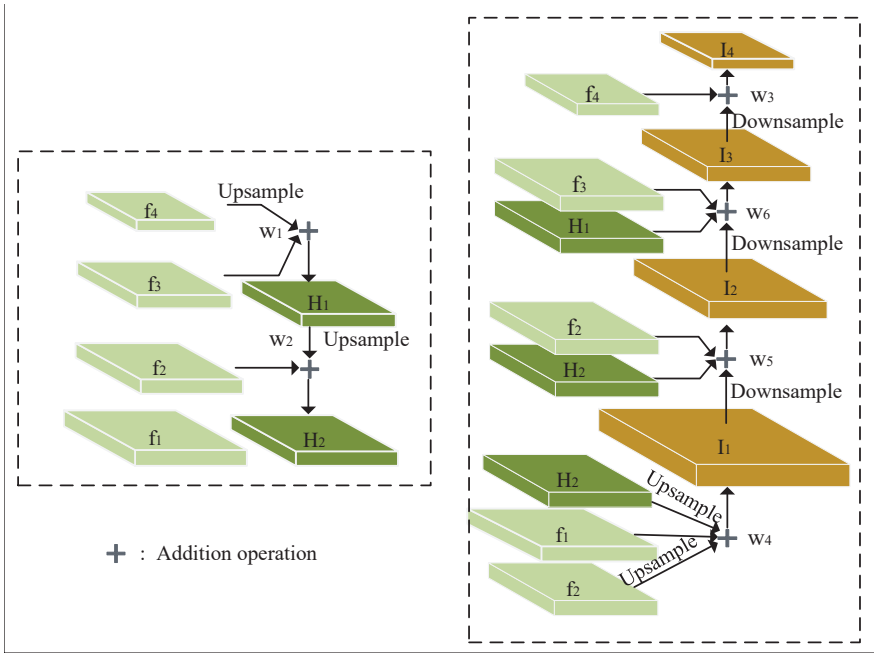

Figure 3. The pipeline of the MFW. Multilevel semantic contexts f_1, f_2, f_3, f_4 are weighted to generate H_1, H_2. Six sets of feature maps are weighted to generate I_1, I_2, I_3, I_4. The weights consist of three 2D sets (w_1, w_2, w_3) and three 3D weight sets (w_4, w_5, w_6), respectively.

3.4. TT Module

We introduce TT to couple multilevel semantic contexts. A transformer [15] was proposed to capture dependencies between different positions in a sequence. To reduce computation complexity, we directly model in the semantic context instead of directly applying the transformer to the pixel sequences of image patches [43]. Additionally, running in a semantic markup space enables contextually informed attention to rich image information.

First, to capture the relevance of multilevel features using self-attention, we adopt a filter-based tokenizer [44] to transform the semantic context into a compact set of semantic tokens. The filter-based tokenizer restructures the visual features using convolutional feature embedding. Additionally, point-wise convolution is employed to map multilevel visual features to independent semantic tokens. This operation avoids introducing additional computational costs in the embedding phase by reducing the number of parameters of the model. Formally, let $X \in \mathbb{R}^{H_{in}W_{in} \times C}$ (height H_{in}, width W_{in}, channels C) denote the input feature from multilevel semantics. We map each feature point $X_p \in \mathbb{R}^C$ to one of L semantic groups using point-wise convolutions. In each group, the visual tokens $T \in \mathbb{R}^{L \times C}$ can be obtained by spatially pooling as:

$$T = SoftMax_{H_{in}W_{in}}(XW_g)^T(XW_c), \tag{5}$$

where $W_g \in \mathbb{R}^{C \times L}$ forms semantic groups from X, $SoftMax_{H_{in}W_{in}}(\cdot)$ translates X activated by W_g into spatial attention, and W_c represents a point-wise convolution for the input feature.

Second, visual markers are inputted into the self-attentive module of the transformer to capture token interactions. This involves utilizing input-dependent weights by designing and supporting visual tokens with variable meanings, encompassing a broader range of possible concepts with fewer tokens. We use the core concept of the self-attention mechanism to capture visual tokens dependencies i.e., by learning the relationships between queries, keys, and values and assigning weights based on the similarity of the query to the keys. This mechanism endows the model with the ability to allocate attention between different locations, enabling the model to effectively capture long dependencies

and important information in the input tokens. We employ two repeated transformer encoders to effectively model interactions among these visual tokens by:

$$T_{out} = LayerNorm(T'_{out} + W_1(\sigma(W_2 T'_{out}))), \quad (6)$$

and

$$T'_{out} = LayerNorm(T + SoftMax_L((Tk_e)(Tq_e)^T)(Tv_e)), \quad (7)$$

where $T_{out}, T'_{out} \in \mathbb{R}^{L \times C}$ are the visual tokens, W_1, W_2 are point-wise convolutions, and $\sigma(\cdot)$ is an activation function $ReLU(\cdot)$. $SoftMax_L(\cdot)$ translates these activations into a token of attention. $LayerNorm(\cdot)$ represents the layer normalization. Self-attention is computed by a compatibility function of the q_e with the corresponding k_e as a weighted average of v_e, in which q_e, k_e, v_e are learnable weights queries, keys, and values, respectively. The generated visual tokens T_{out} produce clearer localized representations to avoid the interference of redundant information.

Third, we fuse the output of two repeat transformers with the semantic context to refine the pixel-array representation and supplement pixel-level information, given by:

$$X_{out} = X + SoftMax_L((XW_q)(TW_k{}^T))(TW_v), \quad (8)$$

where $X_{out} \in \mathbb{R}^{H_{in}W_{in} \times C}$ is the output semantic. W_q, W_k, W_v are learnable weights used to compute queries, keys, and values. We then obtain the augmented semantic context $\{I'_1, I'_2, I'_3, I'_4\}$ from X_{out}.

3.5. Response Layer

The enhanced semantic context $\{I'_1, I'_2, I'_3, I'_4\}$ is fed into the response layer to generate soft output. First, $\{I'_2, I'_3, I'_4\}$ are upsampled to the resolution of $\{I'_1\}$ and concatenate together, as:

$$I_{out} = I'_1 \bigoplus Upsample(I'_2) \bigoplus Upsample(I'_3) \bigoplus Upsample(I'_4), \quad (9)$$

where I_{out} is the output semantics and $\bigoplus$ represents the concatenation.

Second, the output semantics I_{out} generates soft output S_{out} in the segmentation header as:

$$S_{out} = Upsample(conv_{1\times1}(\sigma(BN(conv_{3\times3}(I_{out}))))), \quad (10)$$

where the soft output $S_{out} \in \mathbb{R}^{hw \times class}$ (the number of category $class$). σ represents the activation function ReLU and the BN is the batch normalization.

3.6. Implementation Details

The implemented WCTNet utilizes PyTorch 2.0.1. All experiments are conducted on a high-performance server equipped with 2 RTX-3060 GPUs. The model undergoes training on 2 GPUs and evaluation on 1 GPU. ResNet50 [45] was introduced as a backbone for multilevel feature extraction. The initial learning rate is set to 5×10^{-5} and decayed by a multilearning rate strategy, i.e., multiplied by $(1 - \frac{iter}{total\ iter})^{0.9}$ after each iteration. During the training process, we first pre-train 50 epochs in DeepGlobe to warm up WCTNet, and then 50 epochs of fine-tuning on Inria Aerial and DeepGlobe in turn. In the inference process, we follow the benchmark algorithm FCtL [11] using the test time augmentation (TTA) technique with rotation and flipping.

Evaluation Metrics: Cross-union (mIoU), frames per second (FPS), memory (Mem), and model complexity (FLOPs) are used to investigate validity and inference speed.

4. Results

4.1. Comparison with the States of the Art

We compare the differences in UHR segmentation between the proposed WCTNet and the state-of-the-art methods on the DeepGlobe and Inria Aerial datasets. The models

listed are the most popular and representative ones in recent times, which referenced in WSDNet [13]. The inference paradigm follows local inference with merging, i.e., merging inference results from multiple temporally localized patches.

DeepGlobe Table 1 lists the quantitative results on the DeepGlobe dataset. We can see that WCTNet strikes a good balance between mIoU, model complexity, memory, and FPS, as compared to generic and UHR models. Specifically, the overall inference speed of GLNet [10] and FCtL [11] is very low due to multiscale patch inference. As compared to ISDNet [12], WCTNet removes the heavy RAF module, and thus has faster inference speed, from 27.7 to 35.2. Compared to the state-of-the-art WSDNet [13], WCTNet performs better in terms of both mIoU and inference speed. Also note that our method is more than 270 times faster than FCtL [11]. Moreover, we show the qualitative comparison results in Figure 4. It can be observed that our model is able to identify striped areas (e.g., rivers) and large areas (e.g., agriculture) due to the correlation between local contexts and contexts. Superior to previous models, it is clear that our results are more detailed and closer to the ground truth for both large and small regions.

Table 1. Comparison with states of the art on DeepGlobe test set.

Model	Mem (M)	FPS	mIoU
U-Net [23]	5507	3.54	38.4
FCN [4]	5227	7.91	68.8
ICNet [28]	2557	5.3	40.2
DeepLab [33]	3199	4.44	63.5
BiseNet [29]	1801	14.2	53.0
GLNet [10]	1865	0.17	71.6
FCtL [11]	3167	0.13	72.8
ISDNet [12]	1948	27.7	73.3
WSDNet [13]	1876	30.3	74.1
Ours	1314	35.2	75.2

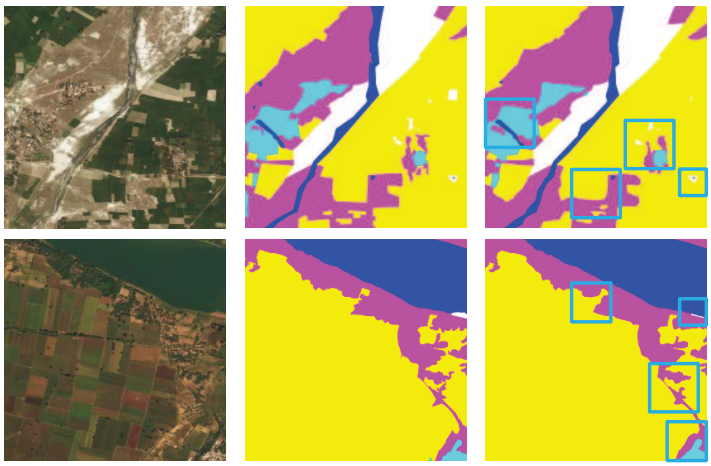

Figure 4. The visualization of the WCTNet outputs in DeepGlobe. From left to right: (**a**) input, (**b**) ground truth, (**c**) segmentation of WCTNet. Blue boxes indicate some distinct areas of fine-grained segmentation. This shows that WCTNet is able to effectively capture the fine-grained regions without being affected by feature downsampling.

Inria Aerial: This has an image pixel size of 5000 × 5000 as the UHR dataset. It annotates only one class of buildings. As shown in Table 2, compared to each of the

baselines, WCTNet achieves a better balance on all metrics. In particular, WCTNet achieves an mIoU of 78.9% and an FPS of 11.2, outperforming the state-of-the-art methods [10–13]. Moreover, WCTNet is simple to train and occupies only 1354 M with a local patch size of 512 and a batch size of 1. The results of the qualitative comparison are presented in Figure 5. We can see that WCTNet highlights regions of attention in an image with different granularities and generates clearer local representations. It can produce clear segmentation while reducing training and reasoning stress.

Table 2. Comparison with states of the art on Inria Aerial test set.

Model	Mem (M)	FPS	mIoU
FCN [4]	2447	1.90	38.4
DeepLab [33]	5122	1.67	55.9
GLNet [10]	2663	0.05	71.2
FCtL [11]	4332	0.04	73.7
ISDNet [12]	4680	6.90	74.2
WSDNet [13]	4379	7.80	75.2
Ours	1354	11.20	78.9

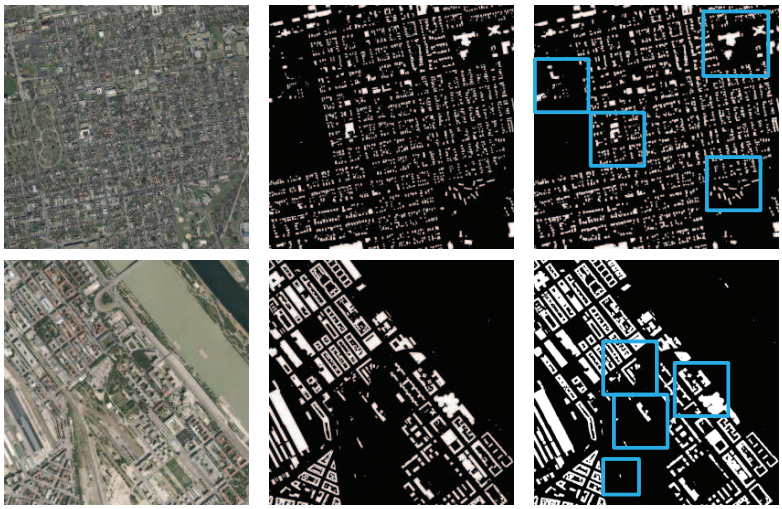

Figure 5. The visualization of the WCTNet outputs in Inria Aerial. From left to right: (**a**) input, (**b**) ground truth, (**c**) segmentation of WCTNet. As shown in the blue box, the proposed WCTNet is able to maintain the accurate segmentation of the fine-grained edge, thus ensuring the integrity of the merged results.

4.2. Ablation Study

This section examines the proposed MFW and TT modules and settings and demonstrates their effectiveness. In all ablation studies, we experiment on the DeepGlobe test set and all models are trained without any external dataset.

The effectiveness of MFW: To evaluate the performance of the proposed MFW, we compare the MFW with two feature networks, namely "FPN" and "PAN". "FPN" stands for feature pyramid network [41] and "PAN" stands for feature network using "2FPEMs + FFM" [46]. Table 3 shows the evaluated values of metrics for different feature networks. We can see that "PAN" outperforms "FPN" in terms of mIoU values with less memory usage and FLOPs. In addition, MFW improves the mIoU by 3.9% and 2.7% over "FPN" and "PAN", respectively, and achieves a faster inference speed of 35.2 FPS. The qualitative

results are shown in Figure 6. Compared with "FPN" and "PAN", background noise can be handled more effectively with MFW.

Table 3. The ablation experiment of feature networks.

Feature Network			Param (M)	FLOPs (G)	Mem (M)	FPS	mIoU
FPN	PAN	MFW					
✓			37.87	50.23	1484	19.0	71.3
	✓		28.53	48.77	1422	27.8	72.5
		✓	25.12	40.86	1314	35.2	75.2

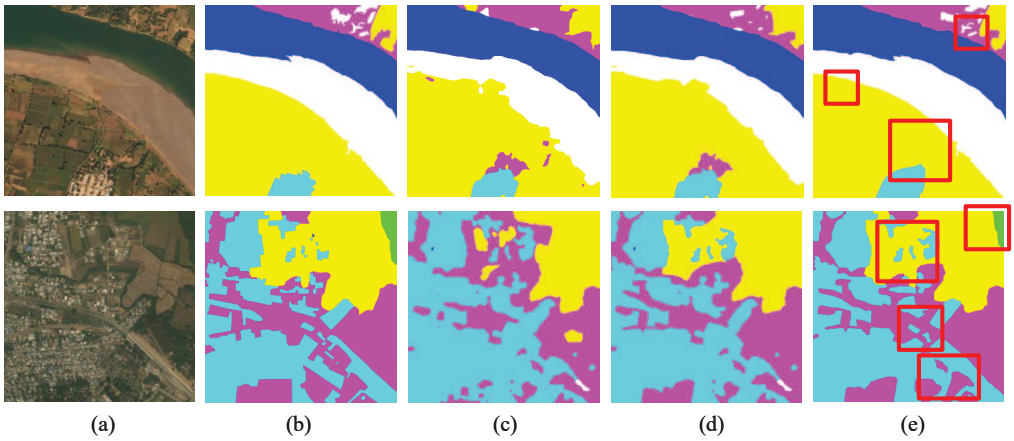

Figure 6. Qualitative results by feature networks. From left to right: (**a**) input, (**b**) ground truth, (**c**) visualization results of 'FPN' inferencing, (**d**) visualization results of 'PAN' inferencing, (**e**) visualization results of MFW inferencing. The red boxes mark where the segmentation maps change significantly. This finding is that the proposed MFW efficiently segments large regions while maintaining the capture and correction of fine-grained regional information.

The effectiveness of TT: Further, we design two sets of experiments to verify the effectiveness of TT module. First, as shown in Table 4, the model with TT achieves 75.2% mIoU performance gain and only 44 M memory increment over the model without TT, while the computational cost only increases 0.82 FLOPs. Second, the impact of the transformer in TT with different numbers of layers is given in Table 5. It is found that modeling two layers of the transformer can bring 3.1% performance gain to the model without extra training cost. Figure 7 shows the visualization results of segmentation. We observe that the model with TT can better focus on rich image information from the context as compared to that without the TT module. Moreover, our method can effectively capture fine-grained image details, as shown in the red dashed box in Figure 7.

Table 4. Ablation results of TT.

TT	Param (M)	FLOPs (G)	Mem (M)	FPS	mIoU
	24.82	40.04	1270	36.7	72.1
✓	25.12	40.86	1314	35.2	75.2

Table 5. Ablation results of the transformer layers in TT. "Num of layers" represents the number of the transformer layers.

TT Num of Layers	Param (M)	FLOPs (G)	Mem (M)	FPS	mIoU
0	24.82	40.04	1270	36.7	72.1
1	25.01	40.35	1292	35.9	74.5
2	25.12	40.86	1314	35.2	75.2
3	25.23	41.38	1336	34.2	75.3

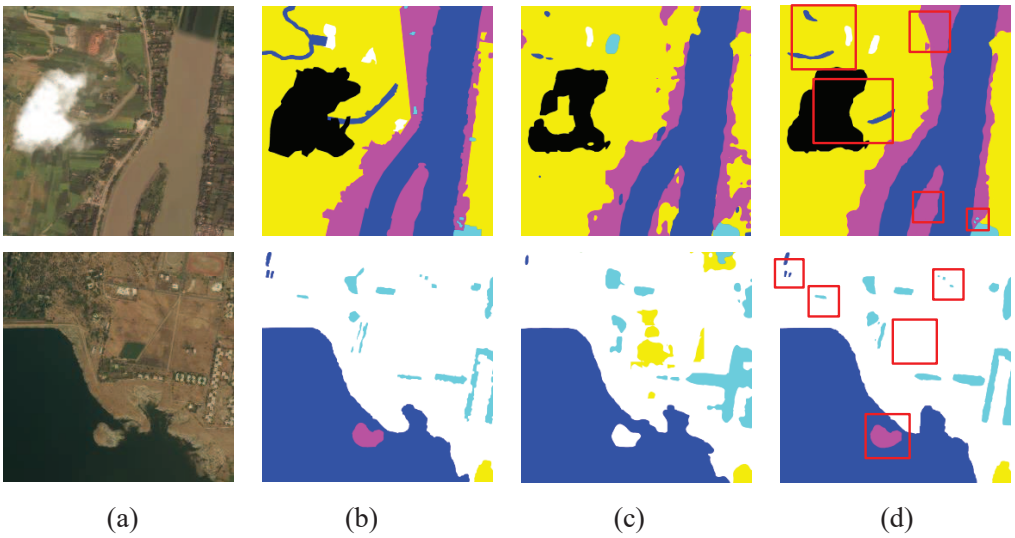

Figure 7. Qualitative results by the proposed TT module. From left to right: (**a**) input, (**b**) ground truth, (**c**) visualization results without inferencing of TT, (**d**) visualization results with inferencing of TT. As shown in the red box, the proposed TT allows the coherence of large region segmentation while maintaining the capture of fine-grained region information.

5. Discussion

UHR image segmentation has attracted a great deal of attention due to its various imaging applications. UHR image classification is usually based on semantic segmentation, which is the process of labeling each pixel in an image into a different semantic category. Semantic segmentation understands the position of each object or scene in an image while effectively recognizing the boundaries by labeling closed-object-class pixels. However, current approaches segment UHR images into standard blocks for multiscale local segmentation and hierarchical inference. This creates an imbalance in the trade-off between memory efficiency and segmentation quality. To solve this dilemma, the WCTNet proposed in this paper considers semantic context to reason about the localization results and merge them into high-resolution semantic masks, since image-level context suffers from redundant information. The semantic context on the localized patch can effectively balance the accuracy and complexity, where both the local and global features are exploited while avoiding additional pixel processing.

The proposed MFW weights shallow and deep features to highlight the attention of regions associated with objects of different granularity in the image. As shown in Figure 6, compared to popular feature networks, MFW can handle background noise efficiently. It can maintain the segmentation information of both large and fine-grained regions. Moreover, considering that the self-attention mechanism captures the dependencies between different positions in the sequence, we propose a TT module to capture the correlations between feature sequences without image-level context. Ablation experiments show that models with TT can better focus on rich image information from the context compared to models

without the TT (as shown in Table 4). In addition, the proposed TT allows for consistency of large region segmentation while maintaining the capture of fine-grained region information (as shown in Figure 7).

As compared with the state of the art, WCTNet effectively balances mIoU, model complexity, memory, and FPS. Qualitative comparison results (Figures 4 and 5) indicate that our model is able to recognize detailed geographic large and small regions, since the correlations between local and global environments can be captured. In addition, WCTNet highlights regions of interest in the image at different granularities and generates clearer local representations without additional training overhead.

6. Conclusions

To alleviate the stress on reasoning computation and training due to hierarchical reasoning, this paper proposes WCTNet for UHR segmentation, where the core pipeline is to reconstruct the global by merging the segmentation of local patches. In WCTNet, the MFW module is proposed to weigh shallow and deep features. In this way, attention to objects of different grain sizes is enhanced by multilevel semantics. It also avoids complex image-level contexts and eliminates the misleading information carried by multiscale contexts for processing single-scale patches. Moreover, we introduce TT to couple multilevel semantic contexts and turn them into semantic tokens via spatial attention. Token interaction information captured can produce clear local representations and avoid redundant interference information. The contextual weighting and coupling of single-scale patches enable WCTNet to strike a balance between accuracy and computational overhead. Rigorous experiments demonstrate that WCTNet achieves competitive mIoU performance compared to state-of-the-art methods while maintaining a high inference speed.

Author Contributions: Methodology, L.C.; Software, H.C.; Formal analysis, Y.X.; Investigation, T.H.; Resources, Y.Z.; Data curation, J.Y.; Writing—original draft, L.C.; Writing—review & editing, H.C.; Visualization, Y.X.; Supervision, Y.Z.; Project administration, T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fujian Provincial Regional Development Project (2015N3015); Fujian Provincial Science and Technology Innovation Team Project (Fujian Education Science and Technology 2018[49]).

Data Availability Statement: DeepGlobe: The data that support the findings of this study are openly available in [DEEPGLOBE] at [http://deepglobe.org/ (accessed on 25 June 2023)]. Inria Aerial: The data that support the findings of this study are openly available in [Inria Aerial Image Labeling Dataset] at [https://project.inria.fr/aerialimagelabeling/ (accessed on 25 June 2023)]. Code is available at https://github.com/giganticpower/WCTNet (accessed on 25 June 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
2. Mishra, H. Introduction To Satellite Remote Sensing. *GIS India* **1998**.
3. Naushad, R.; Kaur, T.; Ghaderpour, E. Deep transfer learning for land use and land cover classification: A comparative study. *Sensors* **2021**, *21*, 8083. [CrossRef] [PubMed]
4. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
5. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
6. Zheng, X.; Cui, H.; Xu, C.; Lu, X. Dual Teacher: A Semisupervised Cotraining Framework for Cross-Domain Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [CrossRef]
7. Hu, H.; Ji, D.; Gan, W.; Bai, S.; Wu, W.; Yan, J. Class-wise dynamic graph convolution for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–17.

8. Sun, Z.; Zhou, W.; Ding, C.; Xia, M. Multi-Resolution Transformer Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 165. [CrossRef]
9. Huang, Z.; Zhang, Q.; Zhang, G. MLCRNet: Multi-Level Context Refinement for Semantic Segmentation in Aerial Images. *Remote Sens.* **2022**, *14*, 1498. [CrossRef]
10. Chen, W.; Jiang, Z.; Wang, Z.; Cui, K.; Qian, X. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 8924–8933.
11. Liu, W.; Li, Q.; Lin, X.; Yang, W.; He, S.; Yu, Y. Ultra-high Resolution Image Segmentation via Locality-aware Context Fusion and Alternating Local Enhancement. *arXiv* **2021**, arXiv:2109.02580.
12. Guo, S.; Liu, L.; Gan, Z.; Wang, Y.; Zhang, W.; Wang, C.; Jiang, G.; Zhang, W.; Yi, R.; Ma, L.; et al. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 4361–4370.
13. Ji, D.; Zhao, F.; Lu, H.; Tao, M.; Ye, J. Ultra-High Resolution Segmentation with Ultra-Rich Context: A Novel Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2023; pp. 23621–23630.
14. Zheng, X.; Chen, W.; Lu, X. Spectral Super-Resolution of Multispectral Images Using Spatial–Spectral Residual Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5404114. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* .
16. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–24 June 2018; pp. 172–181.
17. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
19. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 833–841.
20. Ji, D.; Lu, H.; Zhang, T. End to end multi-scale convolutional neural network for crowd counting. In Proceedings of the Eleventh International Conference on Machine Vision (ICMV 2018), Munich, Germany, 1–3 November 2018 ; Volume 11041, pp. 761–766.
21. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160.
22. Feng, W.; Ji, D.; Wang, Y.; Chang, S.; Ren, H.; Gan, W. Challenges on large scale surveillance video analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 69–76.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
24. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
25. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
26. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
27. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-Scale Context Aggregation for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 701. [CrossRef]
28. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 405–420.
29. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
30. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
31. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2604–2613.
32. Yu, C.; Wang, J.; Gao, C.; Yu, G.; Shen, C.; Sang, N. Context prior for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12416–12425.
33. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]

34. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
35. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
36. Zheng, X.; Sun, H.; Lu, X.; Xie, W. Rotation-Invariant Attention Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 4251–4265. [CrossRef] [PubMed]
37. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
38. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 447–456.
39. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
40. Ghiasi, G.; Fowlkes, C.C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 519–534.
41. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
42. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
44. Zhang, S.; He, X.; Yan, S. Latentgnn: Learning efficient non-local relations for visual recognition. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7374–7383.
45. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
46. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8440–8449.