remote sensing

MDPI

Article

Encoding Contextual Information by Interlacing Transformer and Convolution for Remote Sensing Imagery Semantic Segmentation

Xin Li [1,2], Feng Xu [1,2,*], Runliang Xia [3], Tao Li [3], Ziqi Chen [4], Xinyuan Wang [1,2], Zhennan Xu [1,2] and Xin Lyu [1,2]

[1] College of Computer and Information, Hohai University, Nanjing 211100, China
[2] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China
[3] Information Engineering Center, Yellow River Institute of Hydraulic Research, Zhengzhou 450003, China
[4] Department of Earth System Science, Tsinghua University, Beijing 100084, China
* Correspondence: xufeng@hhu.edu.cn

Abstract: Contextual information plays a pivotal role in the semantic segmentation of remote sensing imagery (RSI) due to the imbalanced distributions and ubiquitous intra-class variants. The emergence of the transformer intrigues the revolution of vision tasks with its impressive scalability in establishing long-range dependencies. However, the local patterns, such as inherent structures and spatial details, are broken with the tokenization of the transformer. Therefore, the ICTNet is devised to confront the deficiencies mentioned above. Principally, ICTNet inherits the encoder–decoder architecture. First of all, Swin Transformer blocks (STBs) and convolution blocks (CBs) are deployed and interlaced, accompanied by encoded feature aggregation modules (EFAs) in the encoder stage. This design allows the network to learn the local patterns and distant dependencies and their interactions simultaneously. Moreover, multiple DUpsamplings (DUPs) followed by decoded feature aggregation modules (DFAs) form the decoder of ICTNet. Specifically, the transformation and upsampling loss are shrunken while recovering features. Together with the devised encoder and decoder, the well-rounded context is captured and contributes to the inference most. Extensive experiments are conducted on the ISPRS Vaihingen, Potsdam and DeepGlobe benchmarks. Quantitative and qualitative evaluations exhibit the competitive performance of ICTNet compared to mainstream and state-of-the-art methods. Additionally, the ablation study of DFA and DUP is implemented to validate the effects.

Keywords: semantic segmentation; Swin Transformer; local patterns and distant dependencies; feature aggregation; well-rounded context

Citation: Li, X.; Xu, F.; Xia, R.; Li, T.; Chen, Z.; Wang, X.; Xu, Z.; Lyu, X. Encoding Contextual Information by Interlacing Transformer and Convolution for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2022**, *14*, 4065. https://doi.org/10.3390/rs14164065

Academic Editors: Thien Huynh-The, Huang Wei and Sun Le

Received: 25 June 2022
Accepted: 16 August 2022
Published: 19 August 2022

1. Introduction

Remote sensing imagery (RSI) is collected by intermediate imaging sensors, commonly equipped on satellite, aircraft, and unmanned aerial vehicles (UAVs), observing ground objects without direct physical contact [1]. Therefore, an exhaustive semantic understanding of RSI impacts profoundly for downstream tasks, such as water resource management [2,3], land cover classification [4–6], urban planning [7–9], hazard assessment [10,11], and so forth. Striving to produce semantic labels of each pixel with a specific class, semantic segmentation [12], which is incipiently developed for natural image processing, has been implemented to remote sensing imagery with outstanding performance.

Conventional segmentation approaches principally deployed handcrafted features as the guidance for recognizing the pixels. Initially, classical methods, such as logistic regression [13] and distance measures [14], were taken because of stability and ease-of-use. Latterly, some superior models, such as support vector machine (SVM) [15], Markov random fields (MRFs) [16], random forest (RF) [17], and conditional random fields (CRFs) [18],

were developed to boost the classifier. However, despite introducing robust classifiers, the artificially selected features inherently constrain the overall performance, especially the unsatisfactory accuracy.

Recently, deep convolutional neural networks (DCNNs) have expressed salient benefits in computer vision tasks [19]. Concretely, DCNNs enable automatically derived features tailored for the targeted classification tasks, making such methods better choices for handling complicated scenarios. Then, the DCNNs intrigued the remote sensing community to study their application to RSI processing. As a result, various DCNN-based RSI interpretation methods were devised, demonstrating the flexibility and adaptability in understanding multi-sources and multi-resolution RSI [20,21]. Beyond simultaneously learning representation and training the classifier, as for semantic segmentation, fully convolutional networks (FCNs) were pioneeringly designed for semantic segmentation tasks [22]. Three stages are involved in FCNs: multiple convolutional layers, deconvolution, and fusion. Therefore, FCNs perform as an end-to-end trainable segmentation network. Subsequently, Badrinarayanan et al. [23] defined and extended the encoder–decoder architecture within the proposed SegNet, which remedies the transformation loss in shrinking and expanding feature maps. To further introduce the contextual information, several skip connections are incorporated in U-Net, enhancing the fidelity and distinguishability of learnt representations [24]. Nevertheless, ascribing to the fixed geometry structure of CNN, these methods are inherently limited by local receptive fields and short-range context. Moreover, RSI with a wide range, diverse objects, and jagged resolutions is more challenging than natural imagery.

After reviewing the studies, it can be concluded that context utilization is a feasible solution for boosting the discriminative ability of learnt representations. Two alternatives were introduced to aggregate rich contextual information, augmenting the pixel-wise representations in segmentation networks. First, different-scale dilated convolutional layers or pooling functions are appended in several works from the perspective of encoding multi-scale features. For RSI semantic segmentation, MLCRNet was proposed and reached preferable performance on ISPRS Potsdam and Vaihingen benchmarks with multi-level context aggregation [25]. Likewise, Shang et al. also designed a multi-scale feature fusion network based on atrous convolutions [26]. Du et al. devised a similar semantic segmentation network to map urban functional zones [27].

Another skillful manipulation is embedding attention modules designed to capture long-range dependencies. Attention is a behavioral and cognitive process of focusing selectively on a discrete aspect of information, whether subjective or objective, while ignoring other perceptible information, playing an essential role in human cognition and living beings' survival [28]. Benefitting from the attention mechanism, the network can focus on where more information lies, improving the representation of essential areas. Thus, the segmentation accuracy has risen significantly with the emerging attention-based methods [29]. For RSI, Li et al. [30] proposed dual attention and deep fusion strategies to address the problem of segmenting large-scale satellite RSI. Li et al. proposed a multi-attention network to extract contextual dependencies while retaining efficiency [31]. HCANet was proposed to hybridize cross-level contextual and attentive representations via the attention mechanism [32]. EDENet attentively learned edge distributions by designing a distribution attention module to inject edge information in a learnable fashion [33]. HMANet [34] adaptively captures the correlations that lie in space, channel, and category domains effectively. Lei et al. [35] proposed LANet to bridge the gap between high-level and low-level features by embedding the local focus with a patch attention module. In general, the attention mechanism has been proved to be superior in the task of the RSI field, helping the models recognize and accept the diverse intra-class variance and inconspicuous inter-class variance.

Transformers have recently revolutionized natural language processing (NLP) and computer vision (CV) tasks based on multi-head self-attention [36]. Semantic segmentation is inevitably ignited by transformers, which performed the capability of modeling rich interactions between pixels. SEgmentation Transformer (SETR) [37] treated the semantic

segmentation as a sequence-to-sequence prediction task, encoding an image as a sequence of patches for natural imagery. Therefore, the global context is modeled and contributes to generating dense predictions. Specifically, powerful capability makes the results state-of-the-art. SegFormer started from the efficiency optimization by comprising an encoder and avoiding decoder [38]. Meanwhile, to eliminate the computational costs and perpetuate the sufficient context, Swin Transformer (ST) has been presented [39]. The hierarchical architecture has the flexibility to model at various scales and has linear computational complexity concerning image size. Consequently, this model surpassed previous state-of-the-art dense prediction approaches by a large margin.

Predominantly, CNNs and transformers both blossomed and demonstrated exemplary performance. CNNs emphasize extracting local patterns, while transformers focus on global ones. From the essence perspective of a digital image, the 2D local structure is ubiquitous. For example, spatially neighboring pixels are usually highly correlated. CNNs' local receptive fields, shared weights, and spatial subsampling force the capture of local patterns. Transformers tokenized the image and sequentially computed the long-range dependencies (also known as global context) to compensate information for inferencing pixels. Accordingly, uniquely forming RSI segmentation models by CNNs or transformers will lead to the deficiency of locality or globality. Detailedly, we summarize the current issues as two aspects.

(1) Pure transformers have demonstrated astounding performance on computer vision tasks with their impressive and robust scalability in establishing long-range dependencies. However, pure transformers flatten the raw image into single-dimensional signals for high-resolution remote sensing imagery, breaking the inherent structures and missing countless details. After revisiting convolutions, CNNs enable the learning of the locality that supplies complementary geometric and topologic information from low-level features fundamentally. Therefore, it is necessary to sufficiently coalesce the convolved local patterns and attentive affinity, which greatly enriches local and distant contextual information, strengthening the distinguishability of learnt representations.
(2) Apart from encoding distinguishable pixel-wise representations, the decoder also plays a vital role in recovering feature maps while preserving the fundamental features. However, the transform loss is inevitable as the network deepens. The existing decoders deploy multiple stages to constantly enlarge the spatial resolution from former layers by bilinear upsampling. Moreover, the shallow layers' low-level features that contain valuable clues for predictions are not aggregated with relevant decoded ones.

Attempting to confront the deficiencies mentioned above, ICTNet is devised to properly encode contextual dependencies in the encoder stage and recover learnt feature maps without information sacrifice. Explicitly speaking, we creatively propose a context encoder (CE) by interlacing CNNs and ST modules for producing demanding representations, which possess plentiful global and local contextual details. Along with the crafted loss-free decoder, which involves data-dependent upsampling (DUP) [40] modules and multi-scale feature fusion strategy, the feature maps that support the dense prediction are eventually generated with dependable inference clues. In a nutshell, three contributions are summarized as follows:

(1) To leverage long-range visual dependencies and local patterns simultaneously, a typical encoder that interlaces convolution and transformer hierarchically is proposed to produce fine-grained features with high representativeness and distinguishability. This design of gradually integrating convolved and transformed representations facilitates the network to exploit the advantages of convolutions in extracting low-level features, strengthening locality, and the advantages of transformers in modeling distant visual dependencies at multiple scales.
(2) Striving to recover the features losslessly and efficiently, in the decoder stage, DUP followed by fusing with a corresponding encoded feature map is devised as the basic unit for constantly expanding spatial resolution. Instead of multiple convolutions

and upsampling operations, one-step matrix projection refers to DUP enabling well-preserved details while enlarging spatial size with an arbitrary spatial ratio.

(3) Concerning the variants and heterogeneity of aerial and satellite imagery, extensive experiments are conducted on three typical semantic segmentation benchmarks of remote sensing imagery, including ISPRS Vaihingen [41], ISPRS Potsdam [42], and DeepGlobe [43]. Quantitative and qualitative evaluations are compared and analyzed to validate the effectiveness and superiority. Furthermore, the ablation study is implemented to verify the efficacy of incorporating the transformer and the designed decoder.

The rest of the paper is organized as follows. Section 2 presents the related works of semantic segmentation of RSI and transformers. Section 3 introduces the devised overall network architecture and each sub-module associated with the formulation. Section 4 collects and compares the results on three representative RSI datasets to validate the performance of the proposed ICTNet, followed by necessary discussions. Finally, Section 5 draws the conclusions and points out the future directions.

2. Related Works

2.1. Attention-Based Semantic Segmentation for RSI

The attention mechanism derives from the human perception, which congenitally perceives surrounding scenes or areas as foveal vision. This selective visual attention endows human beings to recognize salient objects in a sophisticated scene and receive supportive information. Thus, the visual attention paradigm makes the network intellectually and unevenly learn the features. For example, the SE (Squeeze-and-Excitation) block is devised to utilize channel-wise correlations for recalibration weights of each channel [44]. As the foregoer, non-local neural networks were proposed to induce the spatial and channel dependencies simultaneously and self-attentively [45]. At that time, the concurrent CBAM [46], OCNet [47], and DANet [48] similarly validated the effectiveness and efficiency of the self-attention mechanism. These designs have promoted the results on several natural image benchmarks.

Afterward, the success of the self-attention mechanism has been transferred into the RSI semantic segmentation task, well distinguishing easily-confused objects and edges. For example, Teerapong et al. [49] embedded a channel attention block to every stage of the backbone to refine the leant features, leading to remarkable improvements in segmenting aerial images. Cui et al. [50] exploited the spatial relations based on geo-objects in RSI via the spatial attention mechanism. Moreover, SCAttNet [51] was proposed to polish the encoder features with two parallel branches for spatial and channel correlations injection. Concerning the geo-homogeneity of superpixels in RSI, Li et al. presented HCANet [32] that hybridizes the multi-level elements to enhance the local representation without distorting the original semantics of pixels. Likewise, LANet [35] dealt with local regions as the semantic-concerted objects, then applied to calculate object-wise dependencies. Homoplastically, multiple attention modules were composed to guarantee the extraction of sufficient contextual dependencies. More recently, the attention modules were favorably embedded to capture edge priors and boost the delineation of boundaries. In addition to the introduction of edge priors, Yang et al. [52] deployed a multipath encoder with regard to multipath inputs, followed by a multipath attention-fused module to fuse multipath features. Afterward, a refinement attention-fused block is presented to fuse high-level and low-level features. The constructed AFNet achieves state-of-the-art performance on benchmarks.

In addition to natural images, RSI is always acquired from a high-altitude angle, and the observed objects are complex and cross a wide range with the diverse visual surface. While the attention modules are transplanted to RSI, the segmentation performance is inevitably improved with the rich contextual information. Notably, the attention mechanism is also the prototype of transformers, which extends the context model capability to the next level.

2.2. Transformers for Semantic Segmentation

Transformers, an attention-based encoder–decoder architecture, have revolutionized computer vision tasks, including semantic segmentation. Resorting the competitive modeling capability, visual transformers have conveyed a milestone and ignited a new paradigm.

Transformer-based vision networks are started with a vision transformer (ViT) [53]. ViT is the pioneering work that first achieved a CNN-free architecture and formed of a self-attention mechanism. Furthermore, the non-overlapping image patches are tokenized as feature embedding for image classification in this architecture. Consequently, an impressive speed–accuracy tradeoff has been realized compared to CNN variants. Progressively, DeiT [54] incorporated several training strategies to boost efficiency with a smaller dataset. More recently, SETR [55] deployed the first pure segmentation transformer architecture concerning the dense prediction task, encoding an image as a sequence of patches and the position code. The global context is thus modeled in every layer. Moreover, Segmenter [55] was built on the ViT and extended it to semantic segmentation. Additionally, this method designs a simple point-wise linear decoder to the patch encodings. Interestingly, Liu et al. conceived the Swin Transformer [39], which utilizes a shifted window along the spatial dimension to model global and boundary features. Furthermore, patch partition and merging operations were innovatively formulated and embedded in the transformer. As a result, the computational complexity is immensely reduced while reaching state-of-the-art accuracy in multiple vision tasks. Transformers assume minimal prior knowledge about the structure of the problem as compared to their convolutional and recurrent counterparts in deep learning [56]. Considering that global information is also essential for vision tasks, a proper adaption of the transformer should be useful to overcome the limitation of CNNs. Transformers entirely rely on self-attention to capture the long-range global relationships and have achieved brilliant successes.

To our knowledge, long-range visual dependencies and low-level local patterns have identical statuses in RSI semantic segmentation. This is because RSI covers objects with diverse spatial sizes and visual expressions. Hence, collaboratively learning and aggregating structural and contextual information is imperative for remote sensing imagery semantic segmentation. In this study, we intensely explore the properties of the Swin Transformer and investigate the fusion with CNN reasonably.

2.3. Transformers in Remote Sensing

Stemming from the success of multi-head self-attention in long-range dependency modeling, transformers have been widely applied to the remote sensing field. Because non-local dependencies are encompassed in RSI ubiquitously, competitive performance has been established with the incorporation of transformers.

For instance, Bazi et al. [57] first investigated a vision transformer for RSI classification, comprehensively evaluating the performance of ViT. Correspondingly, they explored several data augmentation strategies to alleviate the large scale of data acquisition. As to the task of scene classification, Zhang et al. [58] proposed a remote sensing transformer that links the CNNs and transformers via an intermediate CNN + Transformer block. Moreover, for RSI super-resolution, Lei et al. [59] developed a transformer-based enhancement network that leverages transformers to exploit multi-scale level features. These proposed methods have verified the effectiveness of advanced features with extracted contextual information modeled by transformers.

Semantic segmentation of RSI conducts the precise interpretation target, demanding discriminative and distinguishable pixel-wise representations for dense predictions. Transformers can exploit the contextual information and inject it into the learnt representations by CNNs flexibly, paving a promising way to enhance existing segmentation approaches further. Motivated by this observation, we profoundly analyze the Swin Transformer and CNN backbones, and then propose an interlaced structure that takes advantage of convolved and attentive features for refinement. The following section will detailedly present the topological framework and formal inference of the proposed method.

3. The Proposed Method

In this section, the details of the proposed method are presented and discussed. Before the analysis of overall framework, the directly related preliminaries are introduced. Then, the proposed ICTNet and sub-modules are illustrated and formally described.

3.1. Revisiting Swin Transformer

The Swin Transformer [39] explores a general-purpose CNN-free backbone for computer vision. High-resolution images essentially present a grave problem of efficiency while using standard multi-head self-attention. Moreover, the representations are vulnerable to visual entities' scales and variations compared to text. Henceforth, ST creatively introduces shifted windows to limit self-attention computations to the settled local windows. Quantitively, the computational complexity goes to linear instead of quadratic. With the shift of windows, the encoding context goes beyond local to a broader range of the image by the cross-window connection. The overall architecture pursues encoder–decoder fashion and hierarchically models the dependencies with different spatial sizes.

As illustrated in Figure 1, we first revisit the Swin Transformer Block (STB) in this study, which initially consists of two successive ST blocks. Initially, we denote the features of the former block with linear embedding or patch merging. Formally, the procedure can be represented as:

$$\hat{z}^l = \text{W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1}, \tag{1}$$

$$z^l = \text{MLP}\left(\text{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l, \tag{2}$$

$$\hat{z}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(z^l\right)\right) + z^l, \tag{3}$$

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}, \tag{4}$$

where $\hat{z}^l$ represents the output features of W-MSA module, and z^l is the output of MLP module. W-MSA and SW-MSA denote window-based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

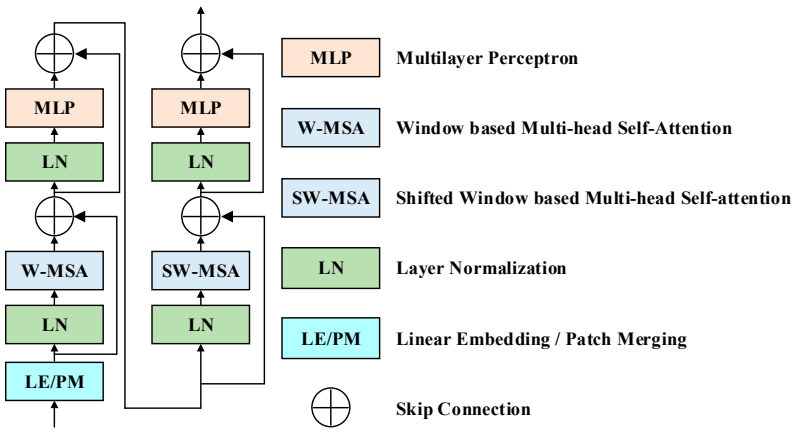

Figure 1. Illustration of Swin Transformer Block.

Similarly, self-attention embedded in this architecture is formed as follows:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{5}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ denote the query, key, and value branches respectively. M^2 represents the number of patches in a window, and d is the dimension. Additionally, B comes from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$.

3.2. ICTNet

In this subsection, the ICTNet is given with corresponding explanations. We first present the overall framework, and then detail the designed encoded feature aggregation module and decoded feature aggregation module.

3.2.1. Overall Framework

As illustrated in Figure 2, the ICTNet inherits encoder–decoder architecture while partial submodules are restructured. A novel encoder that aggregates the features from STBs and CBs is deployed to produce the representations with well-rounded contextual information and local patterns. Correspondingly, the decoder utilizes DUP to enlarge feature spatial size and designs a simple yet efficient feature aggregation module to enrich available clues. In this subsection, we shall first describe the ICTNet formally.

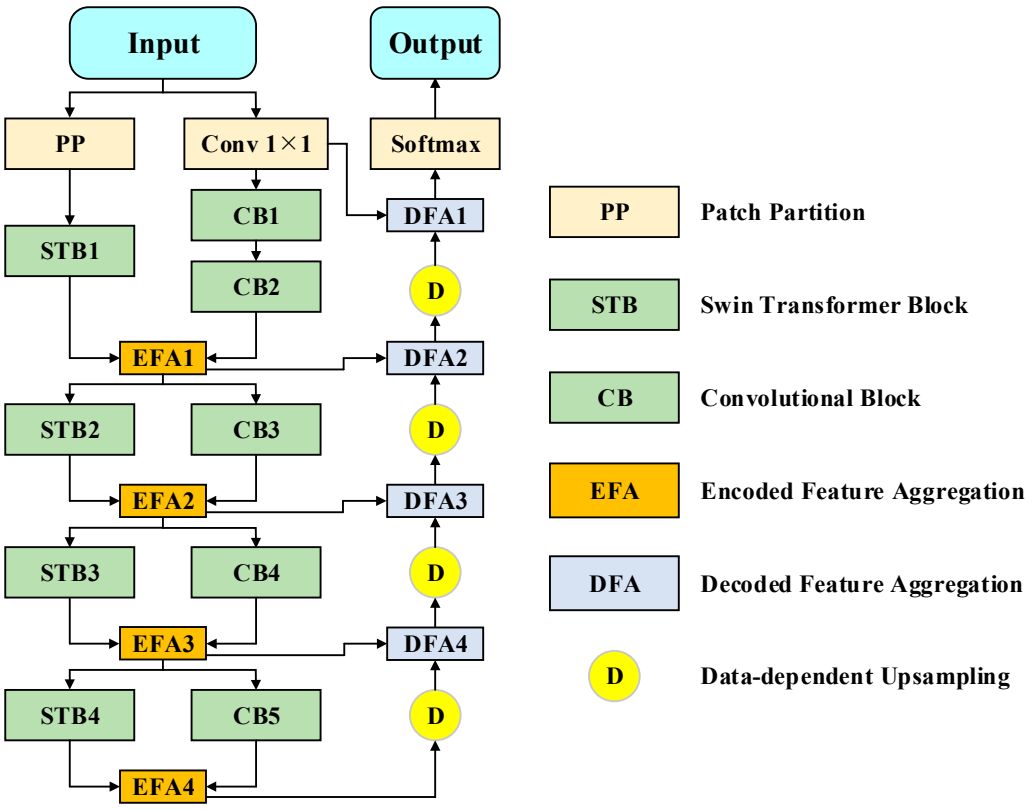

Figure 2. The Framework of ICTNet.

Firstly, a concrete introduction of the encoder is presented. Let the input image with the size of $I \in \mathbb{R}^{H \times W \times C}$, H and W represent the height and width, C is the channels of the raw image. Patch partition initially tokenized the image to s size of $\frac{H}{4} \times \frac{W}{4} \times 48$. The term $F_{st}(i)$ defines the output feature of ith STB, and $1 \leq i \leq 4$. Hereafter, $F_{st}(1) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 256}$ is obtained by STB1. Meanwhile, two CBs produce the convolved feature maps with $F_{cb}(1) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 128}$ and $F_{cb}(2) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 256}$, where s denotes the output feature of jth CB, and $1 \leq j \leq 5$. Afterward, the first encoded feature aggregation module (EFA) is embedded. The aggregated feature by EFA is denoted as $F_e(i)$,

$$F_e(i) = EFA(F_{st}(i), F_{cb}(i+1)), \tag{6}$$

where $EFA(\bullet)$ means feature fusion procedure by EFAs, integer $1 \leq i \leq 4$. To keep the consistency, $F_e(i)$ maintains the same size with $F_{st}(i)$ and $F_{cb}(i+1)$. Due to the feature aggregation, $F_e(4) \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 2048}$ is endowed with the desired contextual information. Instead of fusing output features at the end of the transformer and convolutional back-

bones, this fashion enriches available information at various scales. Totally, four STBs, five CBs, and four EFAs constitute an interlaced visual style that allows the encoder to take advantage of transformer and convolutional features, which have been proven to be beneficial for inference.

Next, the decoder is reformed to recover the features, preferably without information missing. First of all, DUP [40] replaces the bilinear upsampling with matrix projection. DUP was proposed by Tian et al., revealing that the ground truth mask preserves enough mutual dependent structural information and can be compressed with the arbitrary ratio losslessly. Then, the learnable transformation matrix is formed as the projection coefficient for adjusting the spatial size of feature maps. Moreover, the matrix is unceasingly tuned by minimizing the following loss function:

$$L_D(i) = \underset{M_c}{\operatorname{argmin}} \| g_{i-1} - M_c(F_d(i)) \|^2, \tag{7}$$

where M_c is the transformation matrix, g_{i-1} represents the compressed ground truth mask with the same spatial size of $F_d(i-1)$, and $M_c(F_d(i))$ means DUP of $F_d(i)$. Besides the deployment of DUP, the decoded feature aggregation module (DFA) is designed to fuse the features in the decoder stage. Let the output features of the encoder be $F_e(4) \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 2048}$, the $M_c(F_e(4)) \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 2048}$ is produced by DUP. Specifically, the channels are not shrunk in this step. Formally, DFA is given by

$$F_{df}(i) = DFA(M_c(F_e(i)), F_e(i)), \tag{8}$$

where $F_{df}(i)$ is the output of the ith DFA, $F_e(i)$ comes from the corresponding EFA. Finally, the decoded features are predicted by Softmax function.

Overall, the proposed encoder extracts and fuses distant and local dependencies simultaneously by interlaced STB-CB architecture. This design helps the network endow the features with these dependencies at various scales, enriching the contextual clues for inference. Moreover, the decoder skillfully deploys DUP to reduce transformation loss. Moreover, DFA is designed and embedded hierarchically to keep the fidelity and consistency of features. In Sections 3.2.2 and 3.2.3, the details of EFA and DFA are given.

3.2.2. Encoded Feature Aggregation

In this subsection, the schematic diagram of EFA is discussed. Transformers flatten the raw image into single-dimensional signals for high-resolution remote sensing imagery, breaking the inherent structures and missing countless details. Then, the further transformation is unable to recover or regenerate such original features. Therefore, the geometric and topologic information are necessary to be aggregated with transformer-produced self-attentive features. CB convolves the local pixels with fixed receptive field (similar to the concept of window in STB) size and stride, and the interaction between two windows is ignored. Each convolution operation is independent to the next step (stride). In contrast, STB achieves a shifted window attention by two consecutive self-attention layers. In the first layer of STB, a regular window partitioning scheme is adopted, and self-attention is computed within each window. In the second layer of STB, the window partitioning is shifted, resulting in new windows. The self-attention computation in the new windows crosses the boundaries of the previous windows in the first layer, providing connections among them.

To aggregate the features from STBs and CBs, inspired by skip connections, we design a simple yet efficient workflow. Concerning the multi-scale variances of ground objects, we deploy four EFAs to make the multi-scale features well-extracted and aggregated.

As illustrated in Figure 3, $F_{st}(i)$ and $F_{cb}(i+1)$ are of same spatial size and dimensions. First of all, a concatenation operator is applied followed by a 1×1 convolution:

$$F_{sc}(i) = Conv_{1\times1}(Concat(F_{st}(i), F_{cb}(i+1))), \tag{9}$$

where $F_{sc}(i)$ is the preliminary fusion feature from STB and CB, $Conv_{1\times1}(\bullet)$ represents a 1×1 convolution, $Concat(\bullet)$ denotes concatenation, and $1 \leq i \leq 4$.

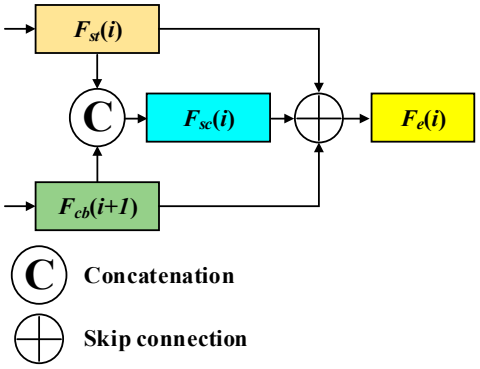

Figure 3. The Pipeline of EFA.

Then, a skip connection is deployed to further aggregation of the features, formally:

$$F_e(i) = SC(F_{st}(i), F_{cb}(i+1), F_{sc}(i)), \tag{10}$$

where $F_e(i)$ is the refined feature by EFA and $SC(\bullet)$ is the skip connection. Intuitively, $F_e(i)$ keeps the same size as $F_{st}(i)$ and $F_{cb}(i+1)$.

The devised EFA is easy and simple, however, the direct efficiency is presented. Prior to fusion, the local patterns and contextual dependencies are well-extracted. The information is carried on $F_{st}(i)$ and $F_{cb}(i+1)$, respectively. Therefore, concatenation and skip connection lend sufficient supports to reach the goal of feature refinement.

3.2.3. Decoded Feature Aggregation

In this subsection, the workflow of DFA is illustrated in Figure 4. Similarly, the skip connection plays the core role in feature aggregation.

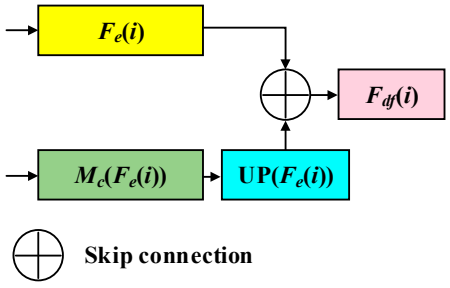

Figure 4. The Pipeline of DFA.

First of all, the input features are $F_e(i)$ and $M_c(F_e(i))$. $F_e(i)$ refers to the output of ith EFA. $M_c(F_e(i))$ comes from the upsampled feature by the former DUP operator. Then, a 1×1 convolution is applied to compress the channel dimensions of $M_c(F_e(i))$. Formally:

$$UP(F_e(i)) = Conv_{1\times1}(M_c(F_e(i))), \tag{11}$$

where $UP(F_e(i))$ is with the same size of $F_e(i)$. At last, $F_{df}(i)$ can be generated by the following equation:

$$F_{df}(i) = SC(F_e(i), UP(F_e(i))), \tag{12}$$

where $SC(\bullet)$ is the skip connection.

As previously discussed, the contextual information is well-rounded in the encoder stage. Further refinement aims at preserving details and clues. Thus, the proposed DFA is devised and embedded in the decoder, resulting in a well-preserved context and representation.

4. Experiments and Discussion

In this section, extensive experiments are conducted on three benchmarks. We first present the experimental settings, data descriptions, and numerical metrics. Next, the qualitative and quantitative evaluations are compared. Moreover, the ablation study of DFA is implemented.

4.1. Settings

In this section, we present the settings of experiments. First of all, three benchmarks are introduced. Nets, the implement details, and comparative methods are presented. Finally, the numerical metrics are given.

4.1.1. Datasets

Three benchmarks are used to evaluate the performance, including aerial and satellite imagery. The details are given as follows.

1. ISPRS Vaihingen Benchmark

The ISPRS Vaihingen dataset is acquired from an aerial platform, observing the town of Vaihingen in Germany. The spatial resolution is 9 cm. Six categories are labeled, including impervious surfaces, buildings, low vegetation, trees, cars, and clutter. Specifically, clutter is served as a background class and ignored in accuracy evaluation. The publicly available data consist of 33 true orthophoto (TOP) tiles with an average spatial size of 2494 $\times$ 2064. Three spectral bands, red (R), green (G), and near-infrared (NIR), are involved in forming the false-color image. There are 14 images of the earlier shared data for training and 2 images for validation. Moreover, the other 17 images are for test. An example of this dataset is illustrated in Figure 5.

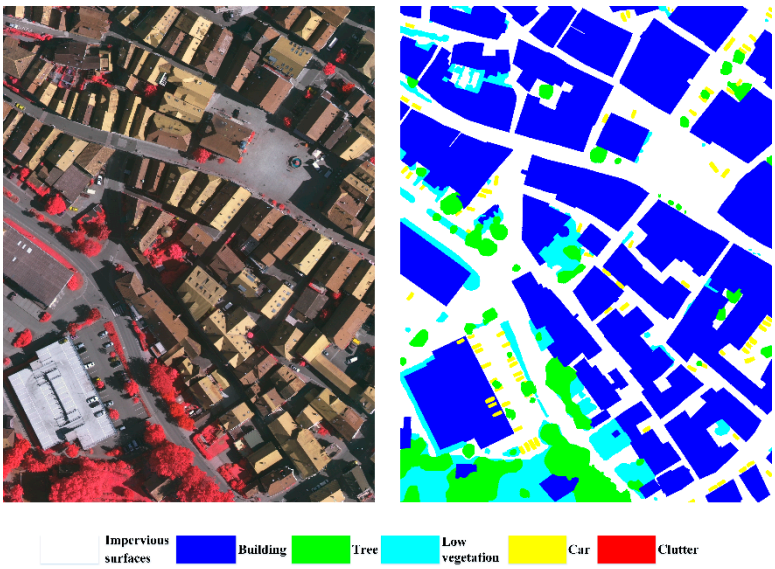

Figure 5. A Sample of the ISPRS Vaihingen Benchmark.

2. ISPRS Potsdam Benchmark

The ISPRS Potsdam dataset is acquired from an aerial platform with a spatial resolution of 5 cm. Ground truth contains the same six categories as the ISPRS Vaihigen benchmark. Four bands, R, G, blue (B), and NIR, are available, in which R, G, and B bands form the raw input imagery. The spatial size of each image is 6000 $\times$ 6000. There are 26 images for training, 4 images for validation and 8 images for test. An example of this dataset is illustrated in Figure 6.

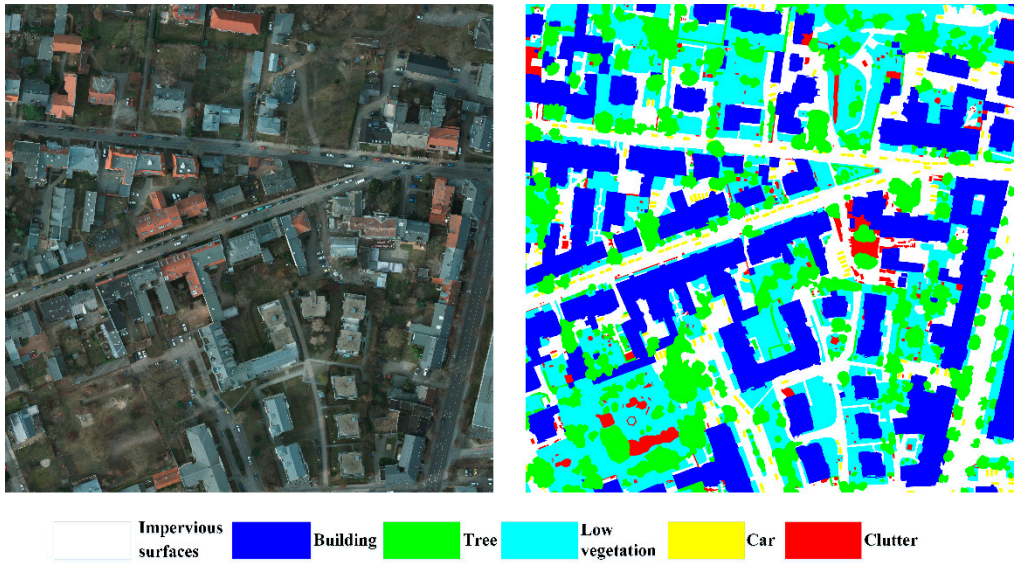

Figure 6. A Sample of the ISPRS Potsdam Benchmark.

3. DeepGlobe Land Cover Dataset

The DeepGlobe Land Cover Classification Dataset is acquired from a satellite with a spatial resolution of 0.5 m, formed of R, G, and B bands. The associated ground truth labels seven categories. They are urban land, agriculture land, rangeland, forest land, water, barren land, and unknown. A wider range and more complex scenarios are exhibited than aerial imagery. A total of 1146 scenes with a spatial size of 2448 $\times$ 2448 can be used. A total of 803 images are trained, 171 images are validated, and 172 images are for test. An example of this dataset is illustrated in Figure 7.

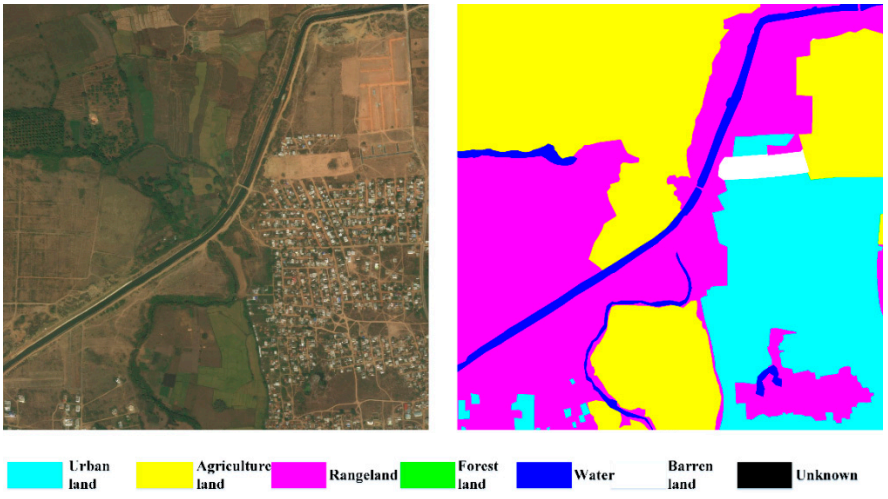

Figure 7. A Sample of the DeepGlobe Benchmark.

4.1.2. Implement Details

The experiments are implemented using PyTorch with NVIDIA Tesla V100-32GB under Linux OS. The hyperparameters are listed in Table 1. The CBs are referred to ResNet 101 under U-Net architecture. STBs come from Swin-S, which is a variant of Swin Transformer.

Table 1. Parameter settings.

Items	Settings
CB Backbone	ResNet 101
STB Backbone	Swin-S
Batch size	16
Learning strategy	Poly decay
Initial learning rate	0.002
Loss Function	Cross-entropy
Optimizer	SGD
Max epoch	500
Sub-patch size	256×256
Data augmentation	Rotate 90, 180, and 270 degrees, horizontally and vertically flip

Moreover, several methods are compared. FCN-8s [22], SegNet [23], U-Net [24], DeepLab V3+ [60], CBAM [46], DANet [48], ResUNet-a [61], SCAttNet [51], and HCANet [32] are re-produced under the same environments and parameter settings. Specifically, DeepLab V3+ and ResUNet-a have multiple versions. In this study, we adopt rate = 2 and rate = 4 to the last two blocks, respectively, for output stride = 8 of DeepLab V3+. As for ResUNet-a, the version of d6 cmtsk is re-produced (see [61] for more details).

4.1.3. Numerical Metrics

Two commonly used numerical metrics, OA (Overall Accuracy) and F1-score, are used. Formally:

$$\mathrm{OA} = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{14}$$

$$precision = \frac{TP}{TP + FP} \tag{15}$$

$$recall = \frac{TP}{TP + FN}, \tag{16}$$

$$\text{Mean F1-score} = \frac{\sum_{i=1}^{n} F_1(i)}{n}, \tag{17}$$

where TP denotes the number of true positives, FP denotes the number of false positives, FN denotes the number of false negatives, TN denotes the number of true negatives, n is the number of class number. In addition, Mean F1-score is an average of class-wise F1-score.

4.2. Compare to State-of-the-Art Methods

As outlined before, our intention was to enrich the contextual information with local patterns and long-range dependencies. Since ICTNet is a segmentation network that inherits the advantages of the Swin Transformer and conventional convolution encoder, the numerical evaluation of segmentation accuracy is investigated experimentally.

4.2.1. Results on Vaihingen Benchmark

As reported in Table 2, the numerical results are collected and compared. It is apparent that both the mean F1-score and OA of ICTNet reveal a significant improvement with the comparative methods. The mean F1-score reaches more than 92%, performing surprising correctness and completeness in segmentation. Moreover, the OA peaks at more than 90%, leading to a more than 1% increase compared to HCANet and ResUNet-a. Compared to FCN-8s and SegNet, the OA sharply rises by a margin of more than 20%. With the incorpo-

ration of the attention mechanism, CBAM and DANet further enhance the representations with long-range dependencies, yielding a remarkable amelioration that about an 8% increase in OA is obtained compared to U-Net. ResUNet-a hybridizes multiple strategies and multi-task inference, considerably amplifying the encoder–decoder network. The complex procedure and massive parameters realize a state-of-the-art performance with a 91.54% mean F1-score and 88.90% OA. Apart from general metrics, the class-wise F1-score is also presented. Distinctly, almost all the class-wise F1-scores of ICTNet are the highest, except for low vegetation. We suppose that the uncertainty of the initialization of neural networks makes this happen. HCANet provides a hybrid context fusion pipeline, adjusting to the visual patterns of trees. Specifically, only 0.31% is degraded. This is acceptable for overall evaluation. In summary, the numerical results indicate the robust learning capability of discriminative representation and certainly predict the dense labels.

Table 2. Results of the ISPRS Vaihingen test set. The class-wise F1-score, mean F1-score, and OA are presented.

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	Mean F1	OA
FCN-8s [22]	84.08	73.61	65.09	79.97	38.76	68.30	66.67
SegNet [23]	87.21	75.37	67.73	82.81	42.58	71.14	69.45
U-Net [24]	84.67	86.13	69.49	82.71	41.89	72.98	71.24
DeepLab V3+ [60]	87.99	87.80	72.72	85.55	47.37	76.29	74.47
CBAM [46]	92.38	87.95	78.81	89.46	57.60	81.24	79.30
DANet [48]	91.57	90.37	80.01	88.15	58.50	81.72	79.78
ResUNet-a [61]	92.98	95.59	85.54	89.36	91.87	91.07	88.90
SCAttNet [51]	89.59	90.77	80.45	80.73	70.87	82.48	80.52
HCANet [32]	94.29	96.20	83.33	**92.38**	88.86	91.01	88.84
ICTNet	**94.69**	**96.70**	**86.04**	92.07	**92.18**	**92.34**	**90.14**

Visualization of random samples from the test set is plotted in Figure 8. All distributed objects have their latent patterns, in which locality and distant affinity are involved. The better the extraction and utilization of local features and long-range context, the higher the accuracy. Unquestionably, ICTNet segments most pixels with high consistency with ground truth. The main parts and edges of objects are well-distinguished visually. Whether with discrete-distributed small objects or consistent areas, ICTNet exhibits satisfactory scalability and adaptability.

The results of the ISPRS Vaihingen benchmark indicate that ICTNet enables well-rounded context together with the convolved local patterns. Furthermore, qualitative and quantitative evaluations vigorously support the superiority of ICTNet.

4.2.2. Results on Potsdam Benchmark

The ISPRS Potsdam benchmark has a finer spatial resolution and observes the Potsdam city of Germany. This benchmark provides more images that could train a model with better performance. Surprisingly, the mean F1-score of 93.00% and OA of 91.57% are showcased by ICTNet according to Table 3. Although ResUNet-a scores 91.47% for OA, there is an increase of 0.1% by ICTNet. Specifically, only the F1-score of impervious surfaces is 0.1% behind ResUNet-a. As to other RSI-specific methods, ICTNet precedes remarkably. Compared to conventional encoder–decoder variants, including FCN-8s, SegNet, U-Net, and DeepLab V3+, a large margin of OA can be observed. Especially for cars, ICTNet recognizes the objects with an F1-score of 96.70%, which is more than twice that of FCN-8s. The complete results reveal that aggregating local patterns and long-range dependencies efficiently benefits the learned representations. Whether the ground objects are sparse or dense, ICTNet lends a solid foundation for feature refinement and pixel-wise predictions. To sum up, the quantitative evaluations on the ISPRS Potsdam benchmark validate the outstanding performance of ICTNet.

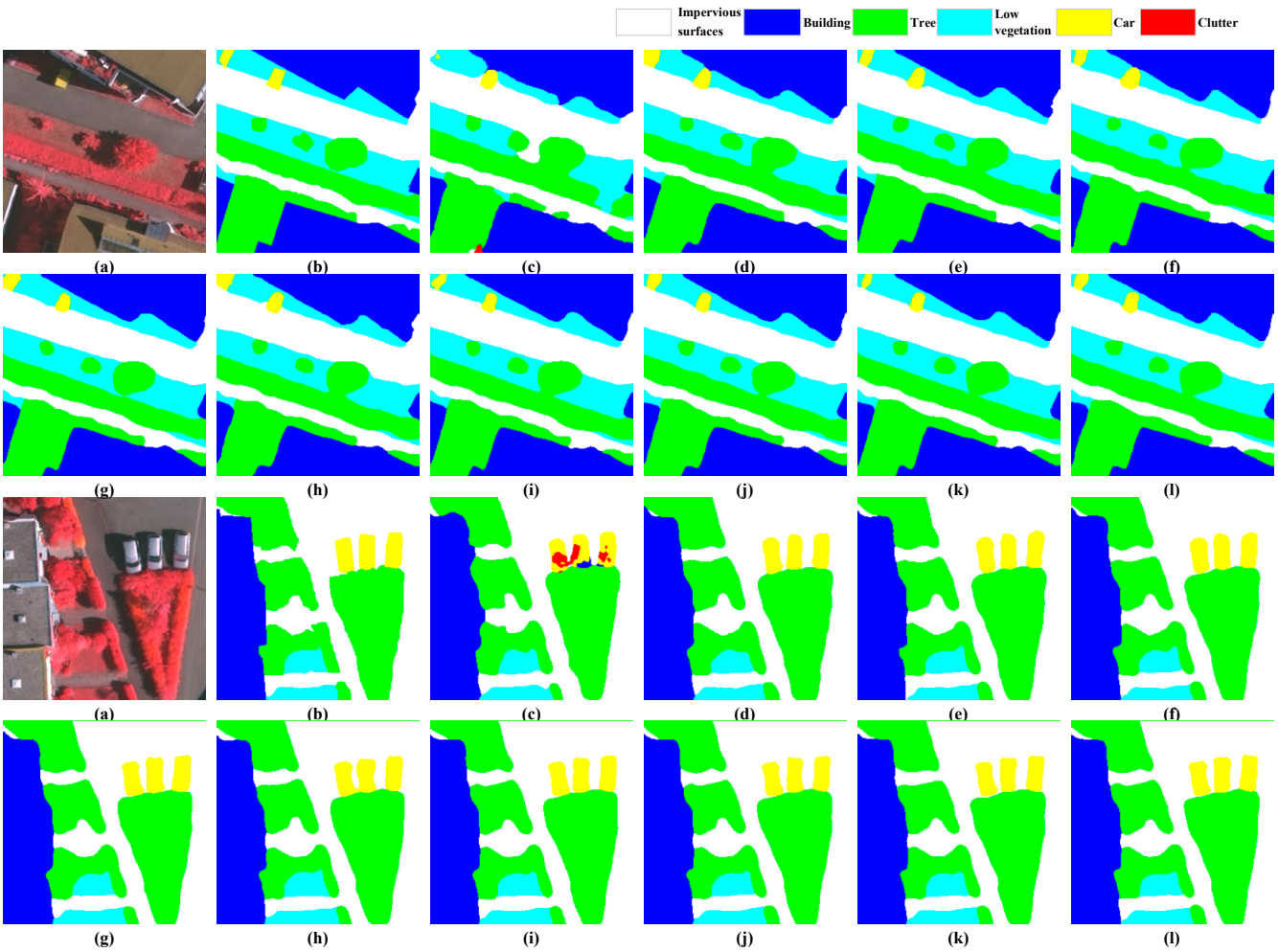

Figure 8. Visualizations of predictions of the test set of the ISPRS Vaihingen benchmark. (**a**) Input image, (**b**) ground truth, (**c**) FCN-8s, (**d**) SegNet, (**e**) U-Net, (**f**) DeepLab V3+, (**g**) CBAM, (**h**) DANet, (**i**) ResUNet-a, (**j**) SCAttNet, (**k**) HCANet, (**l**) ICTNet.

Table 3. Results of the ISPRS Potsdam test set. The class-wise F1-score, mean F1-score, and OA are presented.

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	Mean F1	OA
FCN-8s [22]	84.82	74.26	65.67	80.68	39.10	68.91	67.82
SegNet [23]	85.52	85.75	70.20	83.55	41.71	73.35	72.26
U-Net [24]	88.77	88.58	73.37	86.30	47.79	76.96	75.81
DeepLab V3+ [60]	87.44	90.20	81.03	81.03	89.51	85.84	84.48
CBAM [46]	90.67	95.69	84.54	85.44	86.55	88.58	88.25
DANet [48]	91.07	96.40	83.93	83.73	93.58	89.74	88.36
ResUNet-a [61]	**93.88**	97.30	88.05	88.76	96.60	92.92	91.47
SCAttNet [51]	91.87	97.40	85.24	87.05	92.78	90.87	89.06
HCANet [32]	92.88	96.90	87.25	88.15	93.88	91.81	90.67
ICTNet	93.78	**97.50**	**88.15**	**88.86**	**96.70**	**93.00**	**91.57**

Figure 9 plots the predicted results of two random samples from the test set. Similar trends are observed with the ISPRS Vaihingen benchmark. Although different coverages and spatial resolution make the two benchmarks heterogonous, the visual features are potentially related. Specifically, ICTNet has good transferability and is robust to various data

properties. Consequently, the test results are in good agreement with ground truth labels. Moreover, a majority of areas predicted by ICTNet are almost identical to referred labels.

The results indicate that ICTNet is well-suited for aerial imagery with desired performance. Furthermore, numerical and visual comparisons validate that comprehensive context can boost segmentation accuracy.

4.2.3. Results on DeepGlobe Benchmark

Apart from ISPRS benchmarks, the DeepGlobe benchmark is acquired from a satellite. Visually, the covered range and spatial resolution are different. Therefore, a more heightened capability of the segmentation network is required. Table 4 provides the results of the test set of the DeepGlobe benchmark. There was a lower accuracy than ISPRS benchmarks, deriving from the heterogeneous data property. Even so, ICTNet owns the minimum-amplitude degradation. A total of 86.95% of pixels are well-distinguished correctly by ICTNet, while others are below 82%. FCN-8s only correctly classified about 63% of pixels. With the introduction of atrous convolution and pyramid spatial pooling, DeepLab V3+ rises about 4% of OA. However, long-range contextual information plays a pivotal role in satellite imagery. Therefore, the perceptible field of DeepLab V3+ is scarce. Interestingly, the attention mechanisms alleviate this drawback by few computations, capturing long-range dependencies effectively. As can be seen, CBAM and DANet strikingly level up the OA to more than 77%. Compared to SCAttNet and HCANet, which are the RSI-specific attention-based segmentation networks, our ICTNet enacts a great success. Over 5% improvement of OA is displayed with the two counterparts. In a nutshell, ICTNet expresses competitive capability.

The difficulty of segmenting satellite imagery lies in the easy-prone pixels. These pixels are always found around contours with low certainty. Moreover, the uncorrected classified pixels may lead to more correlated pixels being misclassified. The diversity of ground objects is much more than a scene of an aerial platform. Hence, the predicted results are slightly coarse. However, a striking prediction by ICTNet is illustrated in Figure 10. Compared to conventional and state-of-the-art methods, the segmentation quality is excellent. Easily confused edges are well segmented, and the inner consistency of extensive areas is meritorious.

Table 4. Results of the DeepGlobe test set. The class-wise F1-score, mean F1-score, and OA are presented.

Methods	Urban Land	Agriculture Land	Forest Land	Water	Barren Land	Rangeland	Mean F1	OA
FCN-8s [22]	65.29	66.87	50.39	64.82	69.01	57.88	62.38	63.16
SegNet [23]	67.01	67.32	47.75	67.03	72.82	62.93	64.14	65.58
U-Net [24]	65.92	68.65	56.78	69.83	74.68	71.92	67.96	68.83
DeepLab V3+ [60]	68.97	71.82	59.41	73.06	78.14	75.24	71.11	72.34
CBAM [46]	73.36	79.77	63.43	76.74	81.15	79.87	75.72	77.65
DANet [48]	75.26	79.64	65.03	77.91	81.63	81.30	76.79	78.41
ResUNet-a [61]	80.82	83.15	72.46	76.57	83.33	82.46	79.80	80.20
SCAttNet [51]	79.57	83.20	67.28	81.50	87.15	85.90	80.77	81.79
HCANet [32]	78.75	80.23	68.54	80.20	85.06	81.81	79.10	81.85
ICTNet	**87.62**	**90.14**	**78.56**	**85.18**	**92.34**	**91.40**	**87.54**	**86.95**

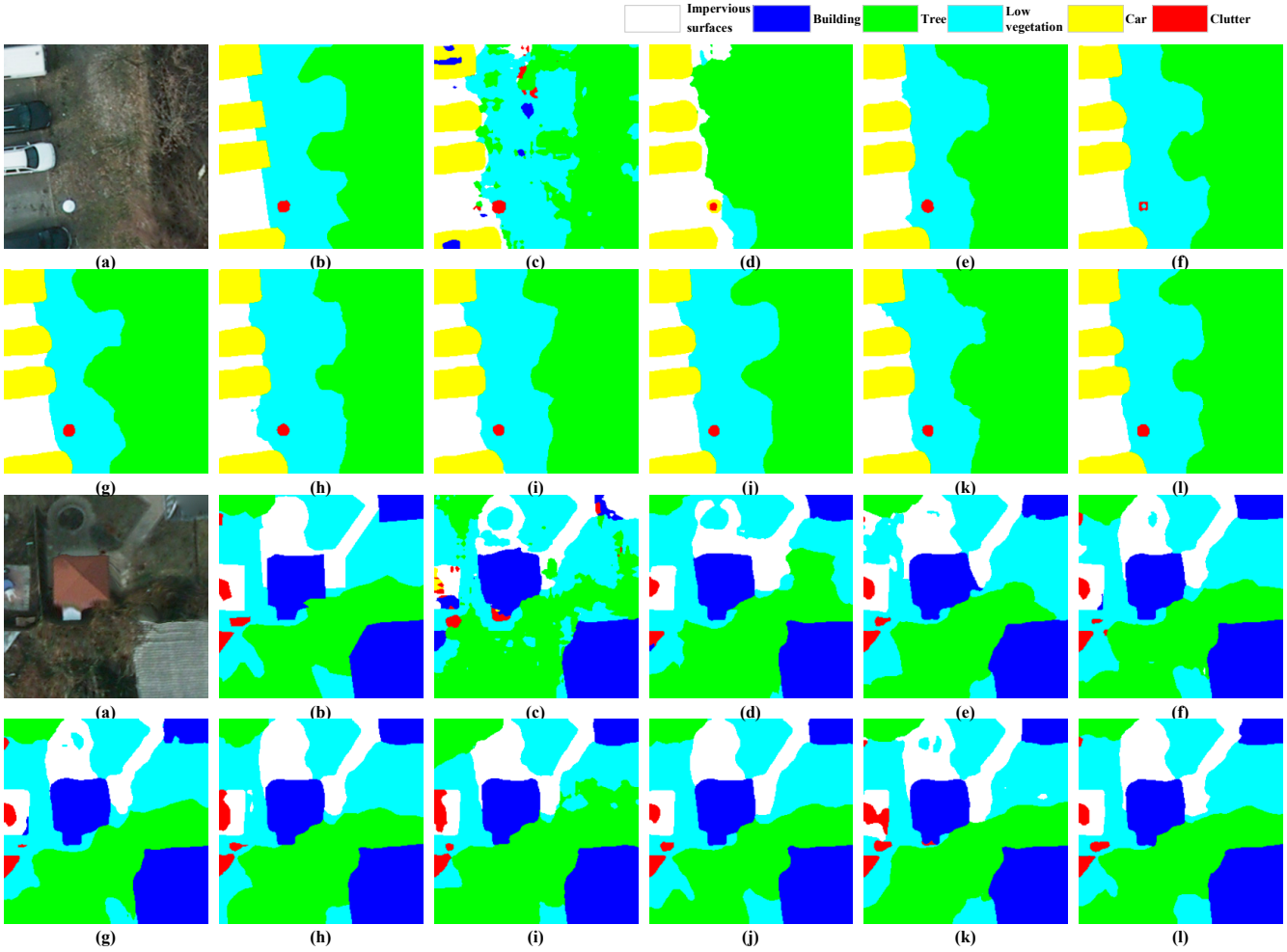

Figure 9. Visualizations of predictions of the test set of the ISPRS Potsdam benchmark. (**a**) Input image, (**b**) ground truth, (**c**) FCN-8s, (**d**) SegNet, (**e**) U-Net, (**f**) DeepLab V3+, (**g**) CBAM, (**h**) DANet, (**i**) ResUNet-a, (**j**) SCAttNet, (**k**) HCANet, (**l**) ICTNet.

In addition to aerial imagery, the results of the DeepGlobe benchmark further evaluate the predominance of ICTNet. The encoded features by STBs and CBs contribute to aggregating available clues. Both numerical results and visualizations corroborate the efficacy and efficiency of ICTNet when segmenting satellite imagery.

4.3. Ablation Study of DFA

In this section, the ablation study of DFA is implemented. For simplicity, we denote ICTNet-S as the version that replaces DFAs with decoder blocks of SegNet. Therefore, the decoder can be considered a universal expansion path by gradually recovering spatial size. Ultimately, we trained these two models on three benchmarks for comparison.

As shown in Table 5, replacing DFAs with the decoder block that consists of convolutional layers implies a degree of attenuation in accuracy. For ISPRS benchmarks, about 2% of mean F1-score and 2.5% of OA are witnessed. With the coarser spatial resolution, the performance degradation of DeepGlobe drops from 86.95% to 80.01%, and about a 6% decrease in OA is examined. For more details of the training procedure, see Appendix A.

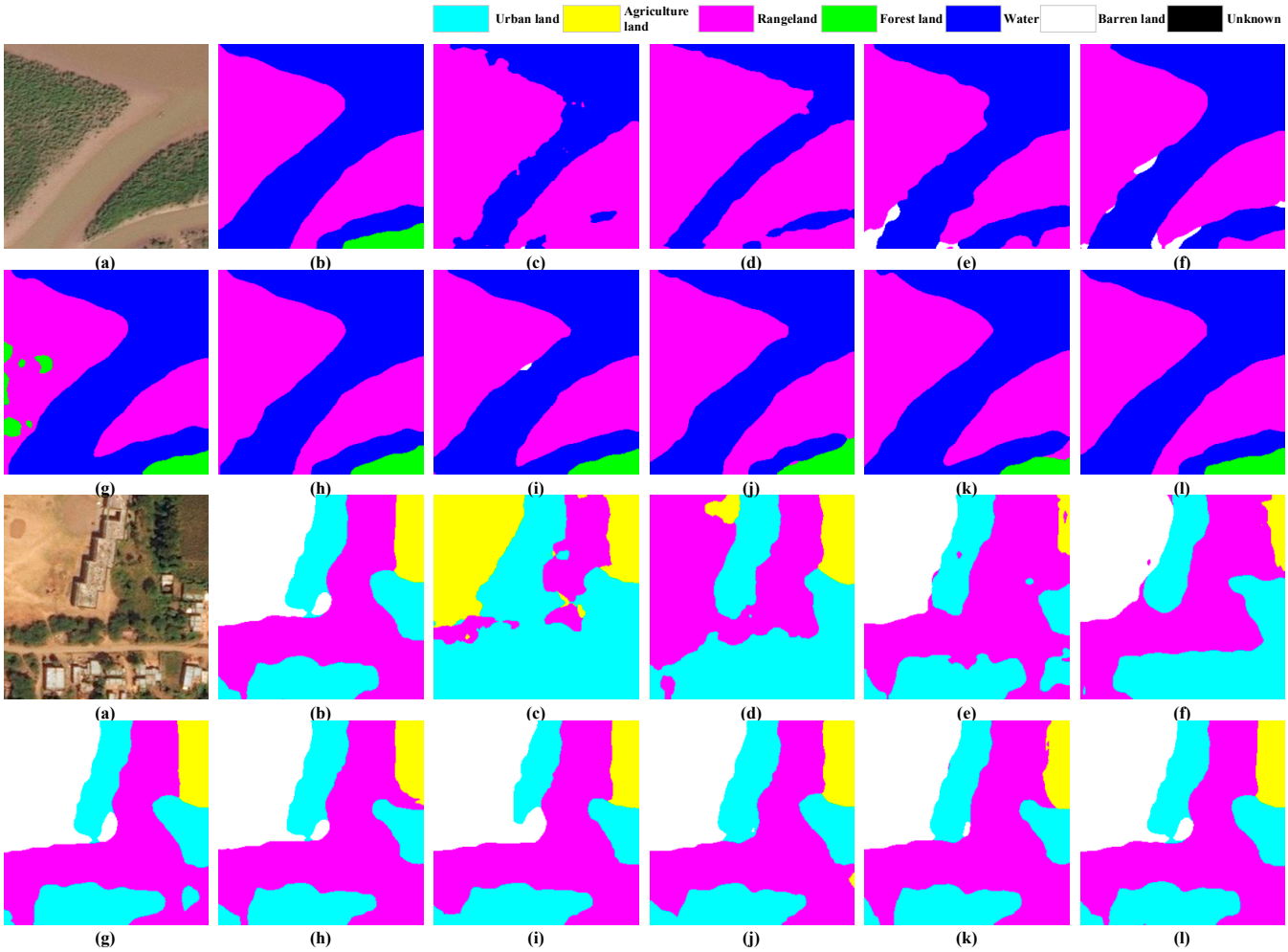

Figure 10. Visualizations of predictions of the test set of the DeepGlobe benchmark. (**a**) Input image, (**b**) ground truth, (**c**) FCN-8s, (**d**) SegNet, (**e**) U-Net, (**f**) DeepLab V3+, (**g**) CBAM, (**h**) DANet, (**i**) ResUNet-a, (**j**) SCAttNet, (**k**) HCANet, (**l**) ICTNet.

Table 5. Results of the ablation study of DFA. The accuracy is in the form of mean F1-score/OA.

Models	Vaihingen	Potsdam	DeepGlobe
ICTNet	92.34/90.14	93.00/91.57	87.54/86.95
ICTNet-S	90.26/87.66	90.64/88.79	76.51/80.01

In a word, the DFAs are used to reduce the transformation loss during spatial recovery. Moreover, injecting encoded features is beneficial for preserving some essential details. As a result, the numerical results lend fundamental support to this conclusion.

4.4. Ablation Study of DUP

The DUPs are incorporated in the decoder for losslessly recovering the spatial size of feature maps. Instead of pooling variants, DUP provides a matrix projection fashion to meet this target. Moreover, one-time matrix projection is affordable in the network. To analyze the effects of DUPs, we re-built two versions for the ablation study. ICTNet-B adopts the bilinear upsampling as the substitution of DUP. ICTNet-M embeds max unpooling for upsampling. As previously discussed, the DUPs induce less loss in expansion. We compare the performance on three benchmarks for comparisons.

As reported in Table 6, the mean F1-score and OA are collected on three benchmarks. Notably, ICTNet provides the most competitive performance compared with the other

two designs. ICTNet-B exhibits the disadvantages in the spatial expansion of feature maps. A gap of about 7% of OA on the ISPRS Vaihingen dataset is presented by ICTNet-B compared to ICTNet. Although max-pooling indices from the encoder stage help the decoder recover features with necessary guidance, the recovered results still suffer from inevitable missing information. ICTNet-M narrows the gap from 7% to about 3%. Likewise, the results from the other two benchmarks express that ICTNet is superior to ICTNet-B and ICTNet-M.

Table 6. Results of the ablation study of DUP. The accuracy is in the form of mean F1-score/OA.

Models	Vaihingen	Potsdam	DeepGlobe
ICTNet	92.34/90.14	93.00/91.57	87.54/86.95
ICTNet-B	85.50/83.04	88.56/86.35	74.02/83.06
ICTNet-M	90.13/87.52	90.88/89.28	75.19/84.37

4.5. Ablation Study of STB and CB

We adjust two versions of ICTNet to examine the effects of STB and CB. As shown in Figure 11, (a) removes STBs from the encoder while (b) removes CBs. Under the same parameter settings, we trained the two versions on three benchmarks to 500 epochs and collected the mean F1-score and overall accuracy of corresponding test sets. Notably, the experiments were deployed simultaneously on three servers equipped with Tesla V100-32GB with respect to three benchmarks.

As presented in Table 7, ICTNet demonstrates the most competitive performance, with STB-only second and CB-only the worst. As we all know, pixels cannot be reliably judged using only the information within the local receptive field, and STB-only inherits the non-local information aggregation capability of the self-attention mechanism, which provides more helpful clues for prediction, resulting in a significant improvement in accuracy compared to CB-only.

However, we find that the STB-only stretches the local region into a vector, losing a large amount of structural and textural information while incorporating the non-local information. Attempting to preserve such local information, convolution paves an effective way in nature. Therefore, an encoder incorporating CBs and STBs can extract and aggregate such information. Considering the scale variation, feature aggregations are performed at each stage of the encoder to ensure the interaction of non-local and local information at each scale.

In summary, the ablation study validates the efficacy and efficiency of interlacing STBs and CBs, making the encoded feature with sufficient contextual information.

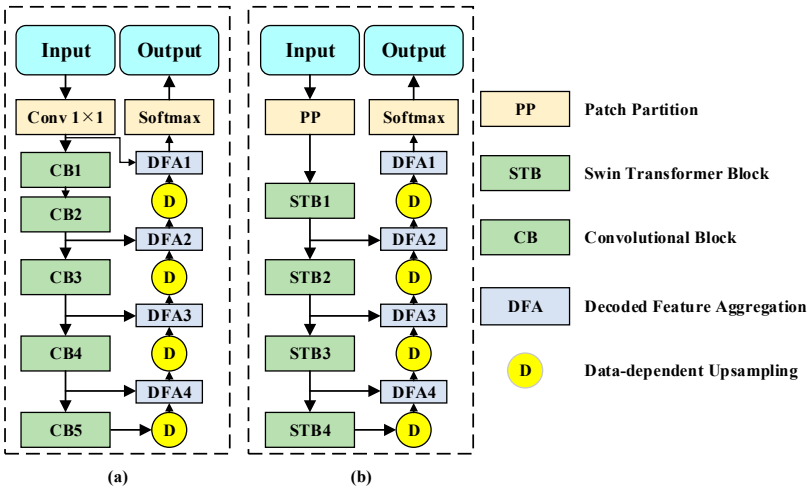

Figure 11. The pipelines of (**a**) CB-only encoder, (**b**) STB-only encoder.

Table 7. Results of the ablation study of STB and CB. The accuracy is in the form of mean F1-score/OA.

Models	Vaihingen	Potsdam	DeepGlobe
STB-only	81.55/79.61	89.55/88.17	76.63/78.25
CB-only	72.85/71.11	77.83/76.63	68.70/69.58
ICTNet	92.34/90.14	93.00/91.57	87.54/86.95

5. Conclusions

This study proposed a novel semantic segmentation network, ICTNet, to conquer the lack of contextual information in the encoder stage and reduce the transformation and recovery loss in the decoder stage. Primarily, we designed an interlaced fashion to take advantage of local patterns and long-range dependencies derived from CBs and STBs, respectively. Meanwhile, the EFAs are devised and embedded. Hence, multiple interactions are implemented in the decoder stage, enriching the aggregated representations with more available information. Symmetrically, the decoder inherits the gradual recovery style. DUPs associated with the proposed DFAs achieve a lossless feature expansion procedure. Consequently, the local patterns and distant correlations, which have been injected into the features, contribute to the inference with desired clues. The experimental results validate the efficacy and efficiency of ICTNet compared to mainstream and state-of-the-art networks. Moreover, the ablation study provides support for DFA and DUP. The findings of this study are promising. In the future, further investigations about the deep fusion of transformers and convolutional neural networks are encouraged.

Author Contributions: Conceptualization, X.L. (Xin Li) and F.X.; methodology, X.L. (Xin Li), F.X. and X.L. (Xin Lyu); software, X.L. (Xin Li), Z.C., X.W. and Z.X.; validation, X.L. (Xin Li), R.X. and T.L.; formal analysis, X.L. (Xin Li) and X.L. (Xin Lyu); investigation, X.L. (Xin Li) and R.X.; data curation, X.L. (Xin Li), Z.C., X.W., Z.X. and X.L. (Xin Lyu); writing—original draft preparation, X.L. (Xin Li) and F.X.; writing—review and editing, X.L. (Xin Lyu) and F.X.; visualization, X.L. (Xin Li) and Z.X.; supervision, F.X.; project administration, R.X., T.L. and X.L. (Xin Lyu); funding acquisition, R.X., T.L. and X.L. (Xin Lyu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program (Grant No. 2018YFC0407105), the Science Foundation for Distinguished Young Scholars of Henan Province (Grant No. 202300410539), the Science Foundation for Excellent Young Scholars of Henan Province (Grant No. 212300410059), the Major Scientific and Technological Special Project of Henan Province (Grant No. 201400211000), the National Natural Science Foundation of China (Grant No. 42104033, 51779100 and 51679103), the Central Public-interest Scientific Institution Basal Research Fund (Grant No. HKY-JBYW-2020-21 and HKY-JBYW-2020-07), the Project of Water Science and Technology of Jiangsu Province (Grant No. 2021080) and the Fundamental Research Funds for the Central Universities (Grant No. B210202080).

Data Availability Statement: Publicly available datasets were analyzed in this study. The download links are: https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx (accessed on 25 May 2021) and http://deepglobe.org/challenge.html (accessed on 9 January 2022).

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

Appendix A. Ablation Study of DFA

This appendix provides the training loss and mIoU of ICTNet and ICTNet-S on three benchmarks, supplying a complementary explanation of the effects of DFA. Network-V, Network-P, and Network-D represent the performance of relative networks on the ISPRS Vaihingen, ISPRS Potsdam, and DeepGlobe benchmarks.

As shown in Figure A1, ICTNet has lower training loss than ICTNet-S regarding the three benchmarks. Correspondingly, Figure A2 presents the training mIoU. Again, the accuracy steadily grows and keeps an intuitive lead by ICTNet.

With the supplementary details of training loss and mIoU, the merits of ICTNet are further supported. Especially, DFAs help realize a lossless decoder. Moreover, in line

with the DFAs, the contextual clues are well-preserved and transformed, contributing to identifying pixels.

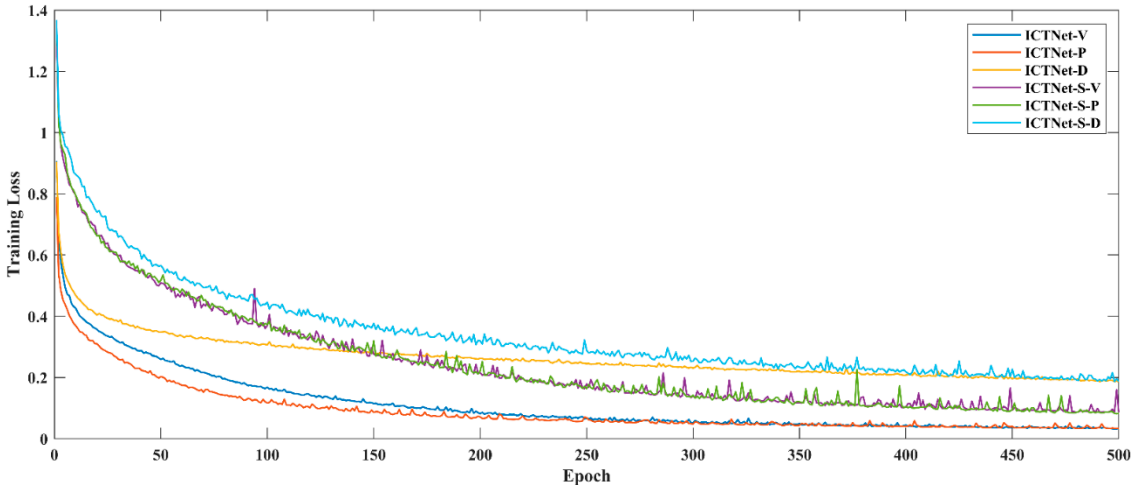

Figure A1. Training loss of ablation study of DFA.

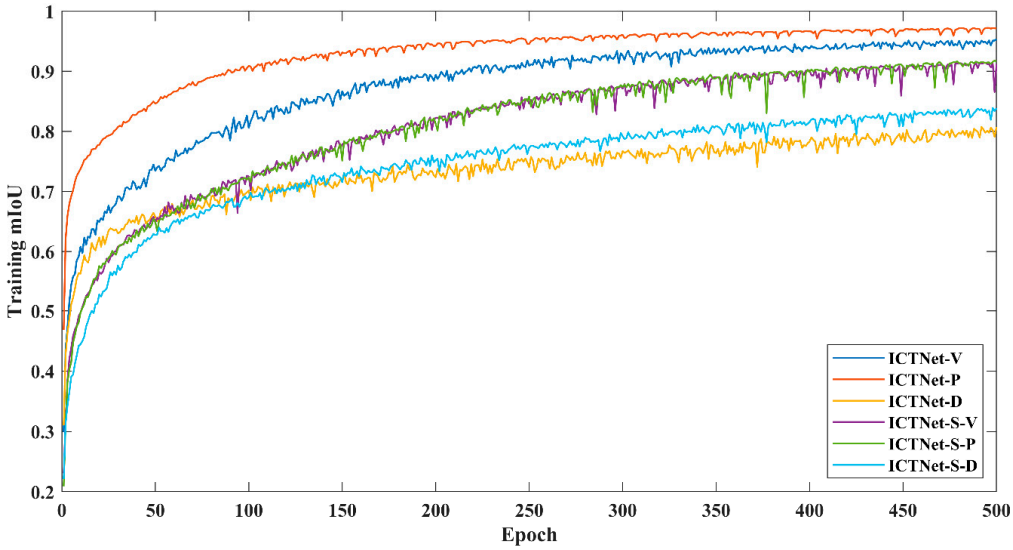

Figure A2. Training mIoU of ablation study of DFA.

Appendix B. Feature Maps of Different Encoder Stages of ICTNet

This appendix gives the feature's size and dimension with regard to different encoder blocks, making a complementary explanation of the proposed decoder. Normally, the input image has a size of $H \times W \times 3$. As listed in Table A1, the blocks' name refers to Figure 2, Section 3.2.

Table A1. Supplementary Details of ICTNet. The size and dimension of output with regard to the block are presented.

Stages	Block Name	Output Feature	Block Name	Output Feature
Stage 1	PP	$\frac{H}{4} \times \frac{W}{4} \times 48$	CB1	$\frac{H}{2} \times \frac{W}{2} \times 128$
	STB1	$\frac{H}{4} \times \frac{W}{4} \times 256$	CB2	$\frac{H}{4} \times \frac{W}{4} \times 256$
EFA1		$\frac{H}{4} \times \frac{W}{4} \times 256$		
Stage 2	STB2	$\frac{H}{8} \times \frac{W}{8} \times 512$	CB3	$\frac{H}{8} \times \frac{W}{8} \times 512$
EFA2		$\frac{H}{8} \times \frac{W}{8} \times 512$		
Stage 3	STB3	$\frac{H}{16} \times \frac{W}{16} \times 1024$	CB4	$\frac{H}{16} \times \frac{W}{16} \times 1024$
EFA3		$\frac{H}{16} \times \frac{W}{16} \times 1024$		
Stage 4	STB4	$\frac{H}{32} \times \frac{W}{32} \times 2048$	CB5	$\frac{H}{32} \times \frac{W}{32} \times 2048$
EFA4		$\frac{H}{32} \times \frac{W}{32} \times 2048$		

References

1. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
2. Caballero, I.; Roca, M.; Santos-Echeandía, J.; Bernárdez, P.; Navarro, G. Use of the Sentinel-2 and Landsat-8 Satellites for Water Quality Monitoring: An Early Warning Tool in the Mar Menor Coastal Lagoon. *Remote Sens.* **2022**, *14*, 2744. [CrossRef]
3. Li, X.; Lyu, X.; Tong, Y.; Li, S.; Liu, D. An object-based river extraction method via optimized transductive support vector machine for multi-spectral remote-sensing images. *IEEE Access* **2019**, *7*, 46165–46175. [CrossRef]
4. Wang, H.; Li, W.; Huang, W.; Nie, K. A Multi-Objective Permanent Basic Farmland Delineation Model Based on Hybrid Particle Swarm Optimization. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 243. [CrossRef]
5. Di Pilato, A.; Taggio, N.; Pompili, A.; Iacobellis, M.; Di Florio, A.; Passarelli, D.; Samarelli, S. Deep Learning Approaches to Earth Observation Change Detection. *Remote Sens.* **2021**, *13*, 4083. [CrossRef]
6. Wang, H.; Li, W.; Huang, W.; Niu, J.; Nie, K. Research on land use classification of hyperspectral images based on multiscale superpixels. *Math. Biosci. Eng.* **2020**, *17*, 5099–5119. [CrossRef]
7. Trenčanová, B.; Proença, V.; Bernardino, A. Development of Semantic Maps of Vegetation Cover from UAV Images to Support Planning and Management in Fine-Grained Fire-Prone Landscapes. *Remote Sens.* **2022**, *14*, 1262. [CrossRef]
8. Can, G.; Mantegazza, D.; Abbate, G.; Chappuis, S.; Giusti, A. Semantic segmentation on Swiss3DCities: A benchmark study on aerial photogrammetric 3D pointcloud dataset. *Pattern Recognit. Lett.* **2021**, *150*, 108–114. [CrossRef]
9. Liu, C.; Zeng, D.; Akbar, A.; Wu, H.; Jia, S.; Xu, Z.; Yue, H. Context-Aware Network for Semantic Segmentation Towards Large-Scale Point Clouds in Urban Environments. *IEEE Trans. Geosci. Remote Sens.* 2022, *early access*. [CrossRef]
10. Pham, H.N.; Dang, K.B.; Nguyen, T.V.; Tran, N.C.; Ngo, X.Q.; Nguyen, D.A.; Phan, T.T.H.; Nguyen, T.T.; Guo, W.; Ngo, H.H. A new deep learning approach based on bilateral semantic segmentation models for sustainable estuarine wetland ecosystem management. *Sci. Total Environ.* **2022**, *838*, 155826. [CrossRef]
11. Bragagnolo, L.; Rezende, L.; da Silva, R.; Grzybowski, J.M.V. Convolutional neural networks applied to semantic segmentation of landslide scars. *Catena* **2021**, *201*, 105189. [CrossRef]
12. Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
13. Csurka, G.; Perronnin, F. A Simple High Performance Approach to Semantic Segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Leeds, UK, 20 June 2008; pp. 1–10.
14. Chai, D.; Newsam, S.; Huang, J. Aerial image semantic segmentation using DCNN predicted distance maps. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 309–322. [CrossRef]
15. Saha, I.; Maulik, U.; Bandyopadhyay, S.; Plewczynski, D. SVMeFC: SVM Ensemble Fuzzy Clustering for Satellite Image Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 52–55. [CrossRef]
16. Zheng, C.; Wang, L. Semantic Segmentation of Remote Sensing Imagery Using Object-Based Markov Random Field Model with Regional Penalties. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1924–1935. [CrossRef]
17. Smith, A. Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. *J. Spat. Sci.* **2010**, *55*, 69–79. [CrossRef]
18. Liu, Y.; Piramanayagam, S.; Monteiro, S.T.; Saber, E. Semantic segmentation of multisensor remote sensing imagery with deep ConvNets and higher-order conditional random fields. *J. Appl. Remote Sens.* **2019**, *13*, 016501. [CrossRef]
19. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* 2021; *early access*. [CrossRef]
20. Sun, L.; Wu, Z.; Liu, J.; Xiao, L.; Wei, Z. Supervised spectral–spatial hyperspectral image classification with weighted Markov random fields. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1490–1503. [CrossRef]
21. Sun, L.; Ma, C.; Chen, Y.; Zheng, Y.; Shim, H.J.; Wu, Z.; Jeon, B. Low rank component induced spatial-spectral kernel method for hyperspectral image classification. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3829–3842. [CrossRef]

22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
23. 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]
24. 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 (MCCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
25. Huang, Z.; Zhang, Q.; Zhang, G. MLCRNet: Multi-Level Context Refinement for Semantic Segmentation in Aerial Images. *Remote Sens.* **2022**, *14*, 1498. [CrossRef]
26. Shang, R.; Zhang, J.; Jiao, L.; Li, Y.; Marturi, N.; Stolkin, R. Multi-scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 872. [CrossRef]
27. Du, S.; Du, S.; Liu, B.; Zhang, X. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. *Remote Sens. Environ.* **2021**, *261*, 112480. [CrossRef]
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, LA, USA, 4–9 December 2017; pp. 5999–6009.
29. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T.S. CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, *early access*. [CrossRef]
30. Li, X.; Xu, F.; Lyu, X.; Gao, H.; Tong, Y.; Cai, S.; Li, S.; Liu, D. Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images. *Int. J. Remote Sens.* **2021**, *42*, 3583–3610. [CrossRef]
31. Li, R.; Zheng, S.; Zhang, C.; Su, J.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
32. Li, X.; Xu, F.; Xia, R.; Lyu, X.; Gao, H.; Tong, Y. Hybridizing Cross-Level Contextual and Attentive Representations for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2021**, *13*, 2986. [CrossRef]
33. Li, X.; Li, T.; Chen, Z.; Zhang, K.; Xia, R. Attentively Learning Edge Distributions for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 102. [CrossRef]
34. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid multiple attention network for semantic segmentation in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
35. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 426–435. [CrossRef]
36. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, *early access*. [CrossRef]
37. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6877–6886. [CrossRef]
38. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), online, 6–14 December 2021; pp. 12077–12090.
39. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Nashville, TN, USA, 10–17 October 2021; pp. 10012–10022.
40. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, LA, USA, 15–20 June 2019; pp. 3126–3135.
41. ISPRS Vaihingen 2D Semantic Labeling Dataset. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html (accessed on 22 December 2021).
42. ISPRS Potsdam 2D Semantic Labeling Dataset. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html (accessed on 22 December 2021).
43. Ilke, D.; Krzysztof, K.; David, L.; 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 31th IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
44. Hu, J.; Shen, L.; Albanie, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
45. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the 31st Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
46. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
47. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object Context for Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 2375–2398. [CrossRef]

48. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, LA, USA, 16–20 June 2019; pp. 3141–3149.
49. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning. *Remote Sens.* **2019**, *11*, 83. [CrossRef]
50. Cui, W.; Wang, F.; He, X.; Zhang, D.; Xu, X.; Yao, M.; Wang, Z.; Huang, J. Multi-Scale Semantic Segmentation and Spatial Relationship Recognition of Remote Sensing Images Based on an Attention Model. *Remote Sens.* **2019**, *11*, 1044. [CrossRef]
51. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic Segmentation Network with Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 905–909. [CrossRef]
52. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
53. 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 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021. [CrossRef]
54. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers and Distillation through Attention. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021; pp. 10347–10357.
55. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 7262–7272.
56. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A Survey of Transformers. *arXiv* **2021**, arXiv:2106.04554.
57. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
58. Zhang, J.; Zhao, H.; Li, J. TRS: Transformers for Remote Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 4143. [CrossRef]
59. Lei, S.; Shi, Z.; Mo, W. Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
60. 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.* **2018**, *40*, 834–848. [CrossRef]
61. Foivos, D.; François, W.; Peter, C.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114.