electronics

Article

ResU-Former: Advancing Remote Sensing Image Segmentation with Swin Residual Transformer for Precise Global–Local Feature Recognition and Visual–Semantic Space Learning

Hanlu Li [1,†], Lei Li [2,†], Liangyu Zhao [1] and Fuxiang Liu [1,*]

[1] Key Laboratory of Dynamics and Control of Flight Vehicle, Ministry of Education, Beijing Institute of Technology, Beijing 100081, China; 3220230033@bit.edu.cn (H.L.); zhaoly@bit.edu.cn (L.Z.)
[2] Aerospace Tianmu (Chongqing) Satellite Science and Technology Co., Ltd., Chongqing 400000, China; lilei@casichttm.com
[*] Correspondence: wwffllff@bit.edu.cn
[†] These authors contributed equally to this work.

Abstract: In the field of remote sensing image segmentation, achieving high accuracy and efficiency in diverse and complex environments remains a challenge. Additionally, there is a notable imbalance between the underlying features and the high-level semantic information embedded within remote sensing images, and both global and local recognition improvements are also limited by the multi-scale remote sensing scenery and imbalanced class distribution. These challenges are further compounded by inaccurate local localization segmentation and the oversight of small-scale features. To achieve balance between visual space and semantic space, to increase both global and local recognition accuracy, and to enhance the flexibility of input scale features while supplementing global contextual information, in this paper, we propose a U-shaped hierarchical structure called ResU-Former. The incorporation of the Swin Residual Transformer block allows for the efficient segmentation of objects of varying sizes against complex backgrounds, a common scenario in remote sensing datasets. With the specially designed Swin Residual Transformer block as its fundamental unit, ResU-Former accomplishes the full utilization and evolution of information, and the maximum optimization of semantic segmentation in complex remote sensing scenarios. The standard experimental results on benchmark datasets such as Vaihingen, Overall Accuracy of 81.5%, etc., show the ResU-Former's potential to improve segmentation tasks across various remote sensing applications.

Keywords: semantic segmentation; transformer; balance between visual and semantic space; enhancement of both global and local aspects

check for updates

Citation: Li, H.; Li, L.; Zhao, L.; Liu, F. ResU-Former: Advancing Remote Sensing Image Segmentation with Swin Residual Transformer for Precise Global–Local Feature Recognition and Visual–Semantic Space Learning. *Electronics* **2024**, *13*, 436. https://doi.org/10.3390/electronics13020436

Academic Editor: Hyunjin Park

Received: 11 December 2023
Revised: 4 January 2024
Accepted: 16 January 2024
Published: 20 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the continuous expansion of remote sensing data and advancements in computer algorithms, there is a growing need to enhance the capabilities of existing models in the field of remote sensing to effectively capture both semantic information and intricate detailed features [1,2]. Semantic segmentation techniques are employed to assign a semantic category to each individual pixel in an image, and accurate pixel-level prediction methods are particularly relevant in the domain of remote sensing images, which often involve multi-scale complex scenes [3].

In complex remote sensing scenes [4], the visual space of remote sensing images is affected by phenomena such as same spectrum different objects or same object different spectra [5], leading to greater spectral differences among similar land features and spectral overlap among different objects. This results in increased intra-class variance and decreased inter-class variance [6], which confuses the image details with high-level semantic information, making it difficult to solve the problem solely through expert visual recognition. Traditional algorithms, such as color clustering [7], are unable to explore the deeper-level high-level semantic information behind the image, which limits their understanding of both local and global image features and reduces task efficiency. Therefore, it is necessary to adopt artificial intelligence machine learning algorithms to identify the advanced information contained in remote sensing images.

Artificial intelligence algorithms are capable of extracting and analyzing specific explicit features, abstracting and summarizing them into high-level semantic information, and systematizing the process of extracting and abstracting specific features, thereby enhancing efficiency and accuracy. However, this process may result in the loss of feature details. The primary task of visual space is to segment and locate feature details, while the primary task of semantic space is to summarize and learn high-level conceptual information. This is where the contradiction lies.

Based on the above problems, researchers begin to explore. Firstly, the addition of a self-attention mechanism [8] in the big model, like ChatGpt [9] and Pangu-Weather [10], which can capture the dependencies of global information [11,12], makes the big model become a phenomenal presence in the processing of natural language or in the weather forecasting domain, and researchers also apply self attention to the image processing domain [8,13]. Borrowing from Transformer [14], researchers introduced Transformer into the image vision field, and the vision transformer was proposed by the Google team for accomplishing the image recognition task [15]. Taking 2D image blocks with positional embedding as input and pre-training on a large dataset, VIT's performance [16] is comparable to that of CNN-based methods, but the computational requirements are enormous and it is restricted only to the image classification domain, unable to solve more downstream tasks. Later comes the Swin Transformer model; Swin Transformer is a method based on the self-attention mechanism and has good global perception, which is able to take into account both the global information and local relationships of the input data, which makes it more effective in dealing with long-global-distance dependencies in images [17]. At the same time, Swin Transformer introduces the rowing window operation, which can help to extract the local features, just like CNN does through the layer design of the convolution operation, and reduce the calculation amount. However, at the same time, the rowing window operation limits the size of its input features. And Swin-Unet [18] is applied to medical image segmentation; however, according to the limitations of its image processing that requires structural curing, as well as the finite number of samples to be processed, it can only be used as a medical solution in the medical field.

To improve the above mentioned crucial problems in the intelligent processing of remote sensing images in complicated scenarios, such as the recognition of multi-level and multi-scale local features, the imbalance between the underlying attributes and high-level semantic information, the lack of long-term semantic information, and the massive amount of required sample data [1], this paper proposes the ResU-Former to address this contradiction. From the perspective of the network's application effects, the ResU-Former enhances the capabilities of both global semantic relationship exchange and local feature recognition, in a way balancing these two aspects.

The breakthrough lies in recognizing the evolutionary nature of information features, which aids in balancing features across different dimensions and improving the utilization of image pixels. The network architecture incorporates various structured designs, including the Swin Transformer structure with residual connections to capture global information, mining contextual relationships among pixel points and all the information the image itself contains, the U-shape structure to complement underlying features, and the cascade connection of feature maps to transfer contextual information. This enables the network to evolve the characteristic information. The Swin Transformer Residual Block is used as the fundamental unit for feature learning, providing the network with sufficient information for ingestion, extraction, reproduction, and learning. This unit establishes distant connections and dependencies between features, thereby uncovering a wider range of contextual semantic information and improving segmentation capabilities. Additionally, a scale adjustment module is introduced to address the constraint of input image feature size for the Swin Transformer. At a macro level, the network complements contextual information through skip connections, while, at a micro level, the scale adaptive block combines with the Swin Residual Transformer to consistently utilize and balance information features, enhancing their utilization rate and allowing for the relearning of prior information from low to high levels. Through this process, the network achieves the evolution of information features and enhances the accuracy both locally and globally, addressing the imbalance between local recognition and global semantic context exchange.

In conclusion, the contributions of this paper can be summarized as follows:

1. The integration of the Swin Transformer and Resnet to construct Swin Residual Transformer blocks, achieving local and global self-attention while suppressing the generation of degeneracy problems and gradient explosion.
2. The design of a scale adaptive block to solve the problem of insensitivity of the Swin Transformer module to input feature size.
3. A symmetric encoder–decoder architecture with skip connections is constructed. In the encoder, gradual convolutional downsampling increases the feature receptive field while decreasing the resolution; in the decoder, features are progressively upsampled back to the resolution at the time of original input.
4. The introduction of fussion loss effectively mitigates the issues of class imbalance by incorporating the Soft Cross Entropy Loss, while Lovasz Loss enhances the model's capability to delineate object contours with higher fidelity. This dual-objective loss function fosters a robust learning process that results in a superior segmentation performance.

2. Related Work

2.1. Unet

Unet is a classical deep learning convolutional neural network structure proposed by Ronneberger et al. in 2015 [19]. It adopts an encoder–decoder structure. Unet increases the receptive field by stacking a large number of convolutional layers and downsampling layers [20], and the high-level feature maps acquired through multi-layer convolutional operations help to segment the target recognition and the skip connections used, which joins the underlying detailed features of the encoder stage to the up-sampling part; this is conducive to the accurate localization of the target. In spite of the U-shaped structure and skip connections in the Unet network, that achieve a certain degree of balance between visual and semantic spaces, there are still some deficiencies in capturing details.

2.2. Swin Tramsformer and CNN

CNN is widely used in the field of image classification with its advantages of excellent local perception and parameter sharing [21]. However, it also has limitations; CNN will lose some details in the process of convolution and pooling, resulting in a lack of sufficient information to recover the image information. The features extracted by CNN are localized, resulting in a lack of contextual connections between pixels. Moreover, the information extracted from superficial and profound features is not the same, and CNN-based semantic segmentation methods fail to utilize this information efficiently.

Dosovitskiv et al. first proposed the transformer backbone network VIT for computer vision [15,22]. The experiments demonstrate that the Vision Transformer (VIT) performs self-attention computation on a global scale, leading to a significant increase in network parameters and a requirement of a large number of training samples. For this reason, Liu et al. proposed the Swin Transformer [17], which divides small windows for patches, calculates local self-attention within the window, and enhances the local features by shifting the window operation to interact with information between different windows [23]; at the same time, in order to be able to design the same hierarchical structure as a convolutional neural network for dense prediction tasks, it is proposed to merge neighboring patch blocks. The two major improvements of calculating local self-attention and merging patch blocks greatly reduce the number of parameters of the network, while maintaining the sensory field of the model, which reduces the difficulty of applying the transformer in the semantic segmentation of remote sensing images. Wang et al. [24] connected a pyramid pooling module (PPM) to the Swin Transformer to obtain rich edge and background information, and Shi et al. [25] simply combined an all-aware module (ALL-MLP) with the Swin Transformer to reduce the complexity when the Swin Transformer is extended to a semantic segmentation network. Yu et al. [26] proposed the combination of multi-scale moving windows with FPN for expanding the sensory field of the network. But feature maps at different scales contain different semantic information, and simply splicing and fusing them together may lead to a serious loss of contextual information; thus, the U-shaped architecture of ResU-Former is constructive.

2.3. Resnet

Traditional deep neural networks [27] are prone to the problem of gradient vanishing or gradient explosion as the number of layers increases, making the network difficult to train and also prone to degradation during the training process [28]. Resnet solves this problem by introducing residual connections. Residual connections allow information to propagate directly across layers in the network, making it easier for the network to learn the residual function. However, the residual block always contains convolutional layers, which is still limited for balancing the global semantic information according to the above theory.

3. Methodology

3.1. The Architecture of ResU-Former Net

The overall neural network ResU-Former Net in this paper is based on the designed Swin Residual Transformer block as the basic unit that supports the designed U-shaped architecture including an encoder, decoder, and skip connections. The Swin Residual Transformer block is based on the Swin Transformer and also uses operations such as convolutional downsampling, focusing on balancing both local and global information features. Meanwhile, the adaptive scale module is added to solve the problem that the Swin Transformer is not sensitive to the input size. The residual connections are embedded within the network for computational compression and solving problems such as gradient explosions.

The network is first supplied with two CBL feature extraction layers, the CBL layer consists of a 3 × 3 convolutional layer, a Batch Norm standard normalization layer, and a LeakRelu activation function layer. Supposing the input feature X, the feature dimension is denoted as $H \times W \times 3$. The formula for the CBL layer is as follows:

$$F_x = LeakRelu(BN(Conv_{3\times3,2}(X))) \tag{1}$$

After the input image undergoes the initial two-layer CBL convolutional feature extraction process, we obtain the feature x_{in} with shape $H \times W \times 3$.

$$x_{in} = CBL(F_x) \tag{2}$$

The basic unit of the ResU-Former is the Swin Residual Transformer Block, as is shown in Figure 1. The task of the network encoder is first to transform the input into a sequential embedding, applying a linear embedding layer to project the feature dimensions into an arbitrary dimensional representation as C. The transformed patch is first passed through a scale adaptive block, a continuous Swin Transformer block to generate the hierarchical feature representation. Among them, the scale adaptive block is responsible for adjusting the input image of an arbitrary scale, which can realize the adjustment of the image scale to be completely adapted to the input scale of the Swin Transformer, solving the problem that the rowing window introduced by the Swin Transformer to enhance the local feature sensory field being insensitive to the input size; the Swin Transformer block is responsible for feature representation learning. Then, the image feature size changed by the scale adaptive block is recovered by the Resize module. The output features are used as residual mappings, which are added to the identity mapping of the starting input to form the output of the residual block, and the above operation is designed to be encapsulated as the Swin Residual Transformer Block. Next, the output features of the Swin Residual Transformer Block are fed into the convolutional layer with the Leakrelu activation layer, which is responsible for downsampling and constitutes the down module as a whole. After the down operation has been performed four times, it enters the up module and starts the task of the network decoder. The decoder is also based on the Swin Transformer composition, and the encoder structure is symmetric; functioning first through upsampling, the neighboring dimensions of the feature mapping reconstructed to a resolution of two times the large feature mapping, the sampled patch through the Swin Residual Transformer Block, and the encoder process are similar.

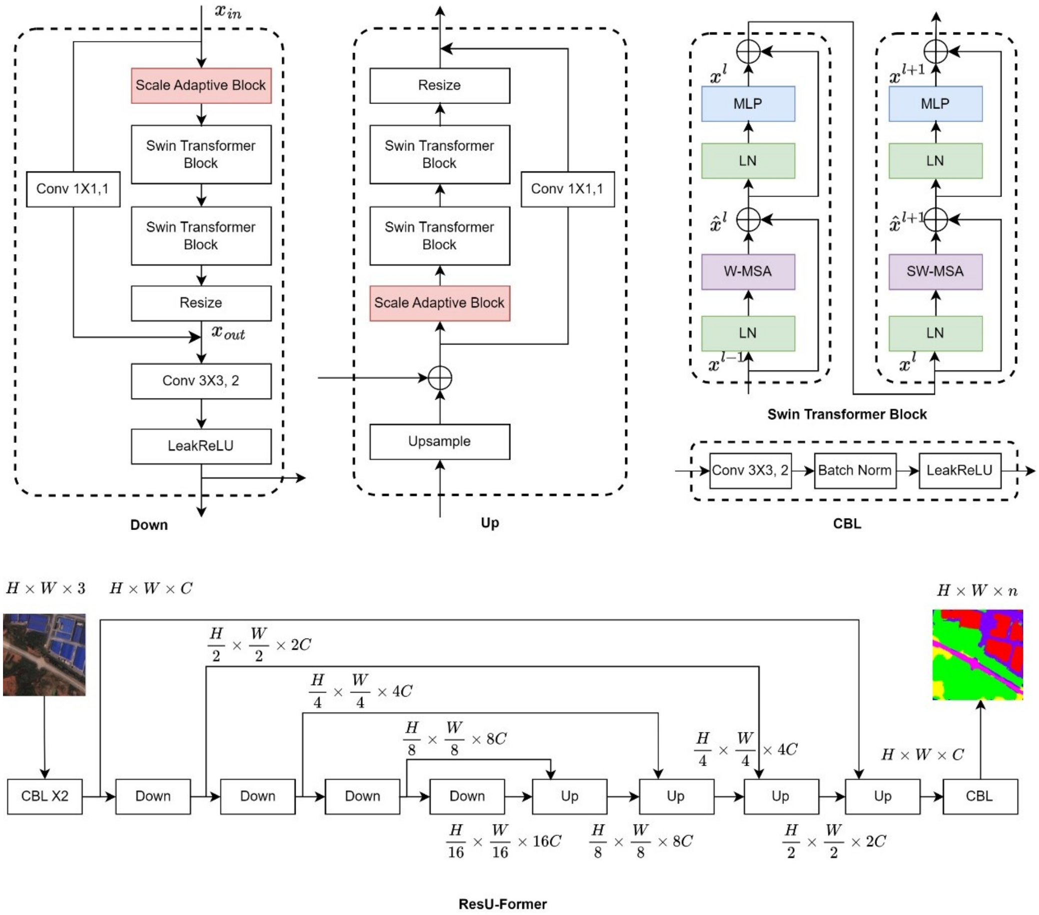

Figure 1. The architecture of ResU-Former Net: in the symmetric structure of a 4-layer encoder and decoder, each layer is based on the unit Swin Residual Transformer Block. With the scale adaptive block and resize module additions, the ResU-Former structure extracts semantic information well.

In the symmetric structure of the decoder and encoder, the extracted contextual features are integrated through skip connections with the decoders to compensate for the loss of spatial information due to downsampling. After four Up module operations, the resolution of the feature mapping is restored to the input resolution $W \times H$, and then a linear projection layer is applied on these final upsampled features to output pixel-level segmentation predictions.

3.2. Details

3.2.1. Scale Adaptive Block

Based on the window partition operation in the Swin Transformer, it commonly sets a fixed 7×7 window size, making one of the limitations of the Swin Transformer that the input image feature size is at least a multiple of 7, resulting in the Swin Transformer being insensitive to the size of the input image. In order to solve this problem, this paper designs a scale adaptive block. The input features are adaptively interpolated and transformed into the input feature, conforming to that in the Swin Transformer based on its structure. The goal is to adjust the size of the input features so that they meet the requirements of the Swin Transformer while not being completely fixed to a particular size. This is an adaptive method that allows the model to handle inputs of different resolutions, thus increasing the range of the applicability of the model.

3.2.2. U-Shaped Architecture

The overall U-shaped structure is supported by Down and Up modules; the encoder contains four Down operations and the decoder contains four Up operations.

For the first four layers of the encoder, the encoder is responsible for gradually reducing the spatial size of the image and adjusting the number of channels to capture information at different scales and consists of multiple Swin Residual Transformer Blocks that compare the attention scores between sequences to capture the contextual relationships; this is used to gradually extract the high-level semantic features of the image as a whole. The number of channels is increased layer by layer by the Swin Residual Transformer Block. The features are subsequently extracted layer by layer by halving the width and height through $Conv_{3\times3/2}$, as the first layer implements the $H \times W \times C \longrightarrow \frac{H}{2} \times \frac{W}{2} \times 2C$. The decoder part then gradually restores the original resolution through an upsampling operation and fuses the features extracted in the encoder with those in the decoder. The decoder part is symmetrical to the encoder part, and the primary body both uses the Swin Transformer, which connects the feature maps in the encoder to the feature maps in the corresponding decoder layer using skip connections after the subsampling, and upsampling operations in order to fuse the low-level and high-level features. This design helps to retain more spatial information and enhance the accuracy of semantic segmentation.

3.2.3. Swin Residual Transformer Block

The most crucial component of the neural network proposed in this paper is the module Swin Residual Transformer Block, which combines the functions of Resnet and the Swin Transformer to significantly increase the neural network's capacity for generalization.

The following is the formula for the Swin Residual Transformer Block:

Suppose the input feature: x_{in}

The input features with resolution $H \times W \times C$ are subjected to scale adaptive operation, and then are turned into $H_1 \times W_1 \times C$, assuming the resolution of x^{l-1}, where SA denotes the scale adaptive operation.

Subsequently, the scale-adaptive input is put into two subsequent Swin Transformer blocks for representation learning, with constant feature size and resolution. The Swin Transformer's construction is based on a shift window, unlike the conventional Multihead Self-Attention (MSA) module. Figure 1 illustrates two successive Swin Transformer blocks, each of which is made up of a Layer Norm(LN) layer, a multi-head self-attention module, a residual connection, and an MLP. The MLP introduces nonlinear transformations using a nonlinear activation function, allowing the network to extract raw data from more abstract and practical features while also making the network more expressive to increase the model's capacity for fitting and representation. Two subsequent Swin Transformer blocks and a shortcut connection compose the Swin Residual Transformer Block. In the meantime, the codec's Down modules and Up modules are primarily composed of the Swin Residual Transformer Block. The output of the Swin Residual Transformer Block is input into another after up-sampling and a convolutional layer of the kernel, 1×1, doubles the number of channels in the Up module. In the Down module, the output of the Swin Residual Transformer Block is subjected to a one-step convolutional layer and a Leakrelu activation layer operation, which further extracts features and enhances nonlinearity. In order to fully mine both global and local information, the Swin Residual Transformer is constructed with a Swin Residual Transformer for global semantic information mining and the convolutional operation for local feature extraction.

The window-based multi-head self-attention ($W - MSA$)module and the shift window-based multi-head self-attention ($SW - MSA$) module are the succeeding transformer blocks of the multi-head self-attention module, respectively. By using the window-based self-attention mechanism, and only performing a localized region within the window self-attention computation, the computational complexity is reduced while the model's receptive field is maintained. The Swin Transformer process is carried out as follows:

$$\hat{x}^l = W - MSA(LN(x^{l-1})) + x^{l-1} \quad (3)$$

$$x^l = MLP(LN(\hat{x}^l)) + \hat{x}^l \quad (4)$$

$$\hat{x}^{l+1} = SW - MSA(LN(x^l)) + x^l \quad (5)$$

$$x^{l+1} = MLP(LN(\hat{x}^{l+1})) + \hat{x}^{l+1} \quad (6)$$

where $\hat{x}^l$ is assumed to be the output of $W - MSA$ and x^l the output of MLP. The self-attention calculation inside the window is represented as follows:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}} + B)V \quad (7)$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$, the values of B are taken from the bias matrix $\hat{B} \in \mathbb{R}^{(2M+1)(2M-1)}$, M^2 indicates the number of divided windows, and d denotes the dimensions of Q, K.

The upscaling is carried out to increase the number of channels in order to achieve the acquisition of a greater range of information. Supposing the feature Y as the output of the successive Swin Transformer blocks, Y's dimension is $H_1 \times W_1 \times 2C$. This is carried out by fusing the channels in the Swin Residual Transformer block using the convolutional layer of the 1 × 1 kernel. Perform the Resize operation as follows:

$$x_{out} = Resize(Y) \quad (8)$$

After the Resize operation, the x_{out} feature becomes $H \times W \times 2C$. x_{out}, as the designed residual mapping of the Swin Residual Transformer block, added with the identity mapping x_{in}, forms the $G(x)$ through shortcut connections.

$$G(x) = x_{out} + x_{in} \quad (9)$$

Next, the $G(x)$ as input undergoes a convolutional layer with a 3 × 3 kernel and a stride of 2, followed by a LeakyReLU activation layer.

$$f(x) = Conv_{3\times3,2}(G(x)) \quad (10)$$

$$L(x) = LeakRelu(f(x)) \quad (11)$$

The Swin Residual Transformer employs a residual block computation technique, which is deduced above, to facilitate the maintenance of gradient flow and enable deeper layers within the network. Just because of the construction of the deep network involving shortcut connections between multiple layers, the ResU-Former Net network comprises numerous stacked Swin Residual Transformer blocks. Each residual block efficiently incorporates input and output information through shortcut connections, thereby facilitating seamless information flow layer by layer.

3.3. Fusion Loss

In traditional segmentation tasks, a common challenge arises from imbalanced class distributions and irregular object shapes, where standard Cross Entropy Loss often falls short. This limitation hinders the model's performance, especially in delineating precise object boundaries.

To address this issue, we propose an innovative loss function that synergistically fuses Soft Cross Entropy Loss with Lovasz Loss at a 1:1 weight ratio. Our approach is designed to leverage the strengths of both: the Soft Cross Entropy Loss facilitates learning from the probabilistic distribution of classes, while the Lovasz Loss directly targets the optimization of the Jaccard index, which is crucial for achieving high-quality segmentation results.

$$L_{SCE} = -\sum_{i=1}^{N=num\_classes} y_i logP(x_i) \tag{12}$$

where L_{SCE} is the value of the Soft Cross Entropy Loss function, $y_i \in (0,1)$ set to soft label, representing the label value of the i class in the groundtruth labels, and $P(x_i)$ represents the probability of the i class predicted by the model. The cross entropy with a smooth label increases the generalization.

The goal of Lovasz Loss is to gradually improve the prediction results such that they are more similar to ground truth labels. In order to address the issues of category imbalance and the disparity between tough and simple samples, it takes into account boundary samples as well as samples that have been wrongly classified.

$$\Delta J_C : M_C \in \{0,1\}^P \mapsto \frac{|M_C|}{|\{y* = c \cup M_C\}|} \tag{13}$$

where ΔJ_C denotes the loss function to be optimized, y^* denotes the groundtruth, c is the set of prediction error pixels, and M_C is the set of mismatches between network segmentation results and labels. $M_C \in \{0,1\}^P$, p denotes the number of pixels.

The Lovasz extension is utilized for smooth extension and is specifically implemented in multi-class segmentation.

$$m_i(c) = \begin{cases} 1 - f_i(c) & \text{if } c = y_i* \\ f_i(c) \end{cases} \tag{14}$$

where $f_i(c)$ refers to the probability value after the softmax of class c. Use the scoring function $f_i(c)$ to construct a pixel errors $m_i(c)$ vector. Use errors $m(c)$ vector to contruct a loss function replacing the loss function ΔJ_C.

$$loss(f(c)) = \overline{\Delta J_C}(m(c)) \tag{15}$$

During training, Lovasz Loss can generate a smooth gradient signal, which aids in the model's generalization and convergence [29]. It is frequently used for applications like object detection, pixel-level segmentation, and is especially effective at addressing issues with class imbalance and boundary sample problems. The Lovasz softmax loss is defined as follows in order to maximize the evaluation of mloU metrics across all categories by averaging the aforesaid $loss(f(c))$:

$$loss(f) = \frac{1}{|C|} \sum_{c \in C} \overline{\Delta J_C}(m(c)) \tag{16}$$

The resilience and generalization capabilities of the model can be increased, overfitting can be decreased, more accurate gradient signals can be provided, and the model can be made to learn and adjust the parameters more effectively by incorporating the two loss function computations. The LovaszLoss is also simpler to combine with other loss functions because it is mathematically differentiable, thereby enhancing the performance of the model.

The weights between various objectives can be balanced when the two loss computations are combined by changing the weights involved.This allows for flexible modification of the model's degree of optimization on various task indicators to accommodate changing needs.

Combining these two loss functions, with Soft Cross-Entropy Loss providing pixel-level classification accuracy, while Lovász Softmax Loss strengthens the model's ability to predict object boundaries, the model can be motivated to better handle boundary regions while maintaining classification accuracy, especially in cases of category imbalance or ambiguous segmentation boundaries in the semantic segmentation task.

3.4. Optimizer and FLOPs Params

The network optimizer uses SGD, sets the momentum to 0.9, uses a learning rate of 0.01, and a weight decay of 10^{-4}; the GPU uses a single RTX 3090 (24 GB) and sets the batchsize to 2. The network FLOPs is 15.06 GB and params is 23.89 MB.

4. Experiments

4.1. Datasets

In this paper, WHDLD, Vaihingen, and Postdom datasets [30] are used. The WHDLD dataset contains six types of remote sensing feature types, the training set contains 4446 images, and the validation set contains 494 images, all of which are 256 × 256. The six types include rural territorial characteristics like water, bare soil, and vegetation, and urban territorial characteristics like buildings, pavement, and roads. The Vaihingen and Postdom datasets are released by ISPRS; both contain five types include buildings, trees, low vegetation, roads, and cars, except for the background. Both datasets are used to develop and test algorithms for identifying different scale types of land cover from aerial imagery. Real-world scenarios that benefit from the analysis of this dataset include urban planning and environmental monitoring. The Vaihingen dataset is preprocessed and cut into 5010 training sets and 1003 validation sets of 256 × 256. Similarly, the Postdom dataset is cut into 3581 training sets and 896 validation sets of 256 × 256 for training.

4.2. Metrics

The evaluation metrics use Overall Accuarcy, Frequency Weighted Accuracy, Mean Accuarcy, and Mean IoU. The above metrics are combined to consider the training effect of the neural network.

Used collectively, these metrics provide a comprehensive view of the model's performance. OA provides a snapshot of how much of the model is classifying correctly across all categories. However, if one category has far more samples than the others, OA can be misleading because it may primarily reflect the accuracy of that dominant category. Remote sensing imagery often suffers from an unbalanced distribution of categories. FWA adjusts for this imbalance by giving more weight to high-frequency categories. This helps us understand how the model performs in the most common scenarios. MA, on the other hand, provides a more balanced perspective by telling us how the model performs, on average, in each category, ignoring the imbalance in the distribution of categories. Mean IoU provides information about the spatial accuracy of the segmentation task, in particular the boundary regions, which provides a good measure of the confusion between categories.

Overall Accuracy is a commonly used evaluation metric, denoted as OA, which is used to measure the accuracy of a model in classifying images at the pixel level.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

Frequency Weighted Accuracy is a commonly used evaluation metric denoted as FWA, which represents the frequency-weighted accuracy. FWA is used to measure the accuracy of a model in classifying pixels of different classes, taking into consideration the frequency of each class in the dataset, thus providing a more fair evaluation of the model's performance.

$$FWA = \frac{\sum_{i=1}^{N=num\_classes}(\omega_i \times t_i)}{\sum_{i=1}^{N=num\_classes}(\omega_i \times n_i)} \tag{18}$$

where ω_i represents the proportion of pixels belonging to class i in the entire dataset, t_i represents the number of pixels correctly predicted as class i by the model, and n_i represents the total number of pixels of class i appearing in the model's prediction results.

Mean Accuracy is a commonly used evaluation metric. MA focuses more on the overall classification accuracy of the model, without considering the importance of individual classes. It is used to measure the average accuracy of the model's classification for each class.

$$MA = \frac{\sum_{i=1}^{N=num\_classes} t_i}{\sum_{i=1}^{N=num\_classes} n_i} \tag{19}$$

where t_i represents the number of pixels correctly predicted as class i by the model and n_i represents the total number of pixels of class i appearing in the model's predicted results.

$$IoU = \frac{TP}{TP + FP + FN} \tag{20}$$

$$MeanIoU = \frac{1}{N}\sum_{i=1}^{N}(IoU)_i \tag{21}$$

where N represents the number of semantic segmentation classes. The $MeanIoU$ is a commonly used evaluation metric for assessing the performance of semantic segmentation models. It quantifies the degree of overlap between the predicted segmentation results and the ground truth labels. Its purpose is to evaluate the segmentation capability of the model across different categories.

4.3. Comparative Experiment

Below lies the comparative images selected from the two datasets—WHDLD and Vaihingen. The comparative nets are danet, deepv3+, deepv3, pspnet, pan, fpn, linknet, manet, unet++, and unet [19,31–39].

The ResU-Former emphasizes the improvement of both the global and local semantic abstraction capabilities, which refers to the balance between global semantic space and local visual space, as well as high-level semantic information and low-level feature information. The improvements can be qualitatively observed from the comparative experimental results in the figures and quantitatively obtained from the metrics of the network in the tables. The high OA value in the tables and the high IoU for small feature classification demonstrate the network's excellent ability to perceive local information. The metrics, OA, and Mean IoU are the highest in the datasets, indicating the network's outstanding performance in abstracting semantic information based on global understanding.

As shown in Figures 2–7 and Tables 1 and 2, the ResUformer-net achieves the best performance across all metrics in the WHDLD dataset. In terms of segmenting large-scale features, the ResU-Former effectively captures image features, resulting in clear boundaries and consistent contours. It exhibits minimal false positives and false negatives, demonstrating superior performance in segmenting detailed features. These results highlight the network's ability to integrate and recognize contextual information, effectively balancing low-level features with high-level semantic information, thus enabling precise localization.

As shown in Figures 8–10 and Tables 3 and 4, it can be observed that the ResU-Former achieves improvements in various aspects compared to other mainstream algorithms in the Vaihingen dataset. It achieves the highest values for OA, FWA, and MeanIoU. Although there is still a certain gap in terms of MA compared to the best-performing method, the ResU-Former network excels in its category segmentation ability, particularly in feature recognition and segmentation. By adopting the self-attention mechanism of the Swin Transformer, it is able to capture long-range semantic information and also focus on image feature edges and contours, which leads to an improvement in the IoU metric. Moreover, it demonstrates a clear advantage in segmenting detailed features.

4.4. Multi-Scale Experiments

The network is trained on the Postdom dataset using a multi-scale strategy with six different input-image scales, as shown in Table 5. Multi-scale experiments can provide a more comprehensive and accurate evaluation of the performance of the semantic segmentation model ResU-Former Net.

Table 1. The metrics for comparative networks trained on the WHDLD dataset are presented below.

Net	**Backbone**	**OA**	**MA**	**FWA**
deepvab3+	Resnet50	0.734758	0.551543	0.602162
deepvab3	Resnet50	0.692037	0.525037	0.561441
fpn	Resnet50	0.725924	0.529165	0.588955
linknet	Resnet50	0.723345	0.452256	0.576835
manet	Resnet50	0.723208	0.527974	0.590553
pan	Resnet50	0.724203	0.520272	0.58053
psp	Resnet50	0.722804	0.528253	0.585737
unet	Resnet50	0.736669	0.534071	0.601409
unet++	Resnet50	0.736234	0.542489	0.602753
danet	Resnet50	0.703276	0.516436	0.563059
resuformer	Swin-T	**0.794566**	**0.669945**	**0.683484**

Bolded data are optimal for each indicator.

Table 2. The IoU metrics of comparative networks trained on the WHDLD dataset are provided below.

Net	**MeanIoU**	**Water**	**Building**	**Bare Soil**	**Vegetation**	**Pavement**	**Road**
deepvab3+	0.424909	0.793887	0.408384	0.325546	0.702975	0.279097	0.039563
deepvab3	0.382345	0.793598	0.316567	0.276701	0.649479	0.257514	0.000213
fpn	0.413195	0.789803	0.384012	0.31177	0.687112	0.261684	0.044788
linknet	0.350987	0.782625	0.370225	0.003271	0.701498	0.247631	0.00067
manet	0.403766	0.787113	0.381844	0.290226	0.700507	0.247356	0.015548
pan	0.400813	0.78435	0.354016	0.3106	0.680746	0.268258	0.006905
psp	0.412346	0.797214	0.359893	0.313549	0.678884	0.270671	0.053867
unet	0.414915	0.805472	0.40225	0.303025	0.704142	0.269928	0.004672
unet++	0.417252	0.816493	0.383985	0.309101	0.707612	0.24725	0.039068
danet	0.38925	0.770836	0.33121	0.30178	0.655124	0.272636	0.003915
resuformer	**0.496528**	**0.86254**	**0.434463**	**0.33647**	**0.751058**	**0.338788**	**0.255852**

Bolded data are optimal for each indicator.

Table 3. The metrics for comparative networks trained on the Vaihingen dataset are presented below.

Net	**OA**	**MA**	**FWA**
danet	0.801468	0.622492	0.665658
deepv3+	0.793168	0.62533	0.655897
deepv3	0.807184	0.628513	0.676598
pspnet	0.797542	0.637307	0.665285
pan	0.792273	0.619997	0.657551
fpn	0.793692	0.609119	0.657784
linknet	0.809461	0.646067	0.68159
manet	0.796106	0.640289	0.660128
unet++	0.797842	0.630655	0.663521
unet	0.772966	0.589576	0.632654
resuformer	**0.815147**	**0.651314**	**0.689264**

Bolded data are optimal for each indicator.

Table 4. The IoU metrics of comparative networks trained on the Vaihingen dataset are provided below.

Net	MeanIoU	Surface	Building	Low Vegetarian	Tree	Car
danet	0.607947	0.695523	0.749333	0.518418	0.704824	0.371639
deepv3+	0.599365	0.677548	0.717366	0.526712	0.7142	0.360997
deepv3	0.611544	0.704255	**0.759534**	0.543725	0.707851	0.342357
pspnet	0.593678	0.675463	0.731295	0.556682	0.717106	0.287846
pan	0.59517	0.668688	0.696356	0.574505	0.714545	0.321757
fpn	0.593983	0.671545	0.72159	0.543336	0.712726	0.320718
linknet	0.594678	0.674396	0.721338	0.500775	0.704821	0.372051
manet	0.604153	0.678732	0.716838	0.538265	0.718308	0.368624
unet++	0.604827	0.676129	0.709841	0.56384	0.721659	0.352668
unet	0.607947	0.695523	0.749333	0.518418	0.704824	0.371639
resuformer	**0.631888**	**0.712576**	0.736338	**0.594463**	**0.7317345**	**0.3843269**

Bolded data are optimal for each indicator.

The multi-scale experiment uses input-image scales as independent variables, and the scale adaptive module designed by this network can freely adjust the input scale into the Swin Transformer. The performance of the net varies under different input-image scales, as shown in Table 5, and the 200 × 200 scale is the optimum, where OA is 73.99%, FWA is 59.43%, and Mean IoU is 52.22%.

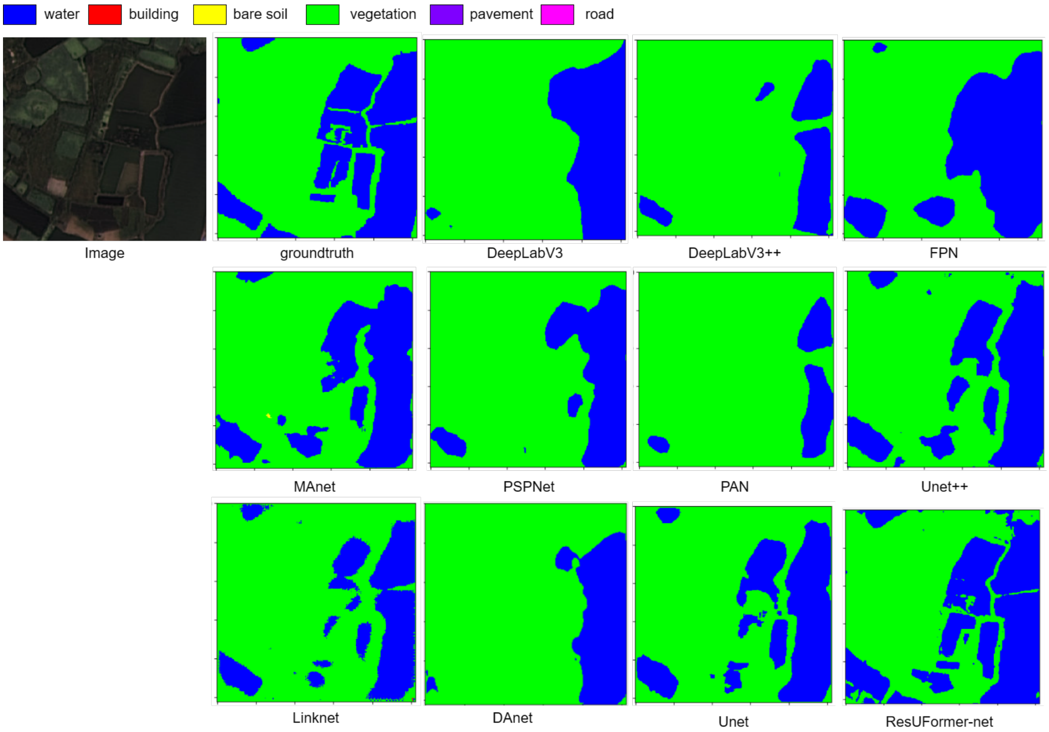

Figure 2. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the dataset-WHDLD. This shows the distribution of zigzag boundaries and fragmented waters, which fully demonstrates the network's grasp of global semantic information, as well as its accurate segmentation of local boundaries and small areas of water, illustrating the network's accurate understanding of local semantic performance.

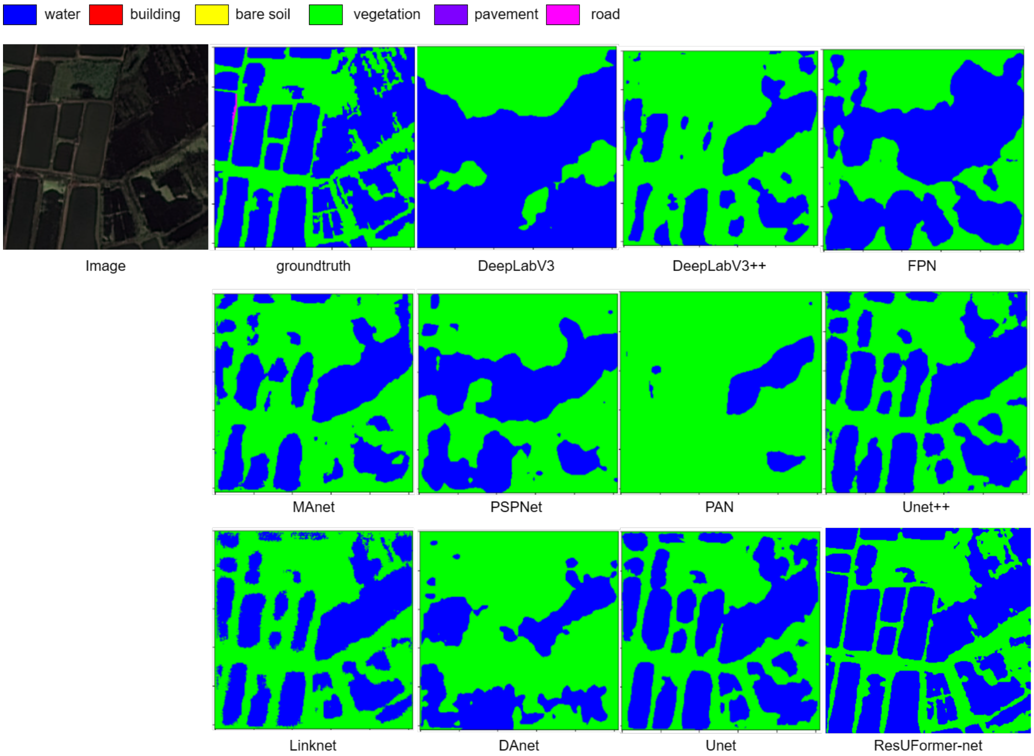

Figure 3. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the dataset-WHDLD. The feature recognition of the water segmentation block demonstrates the network's superiority in local understanding.

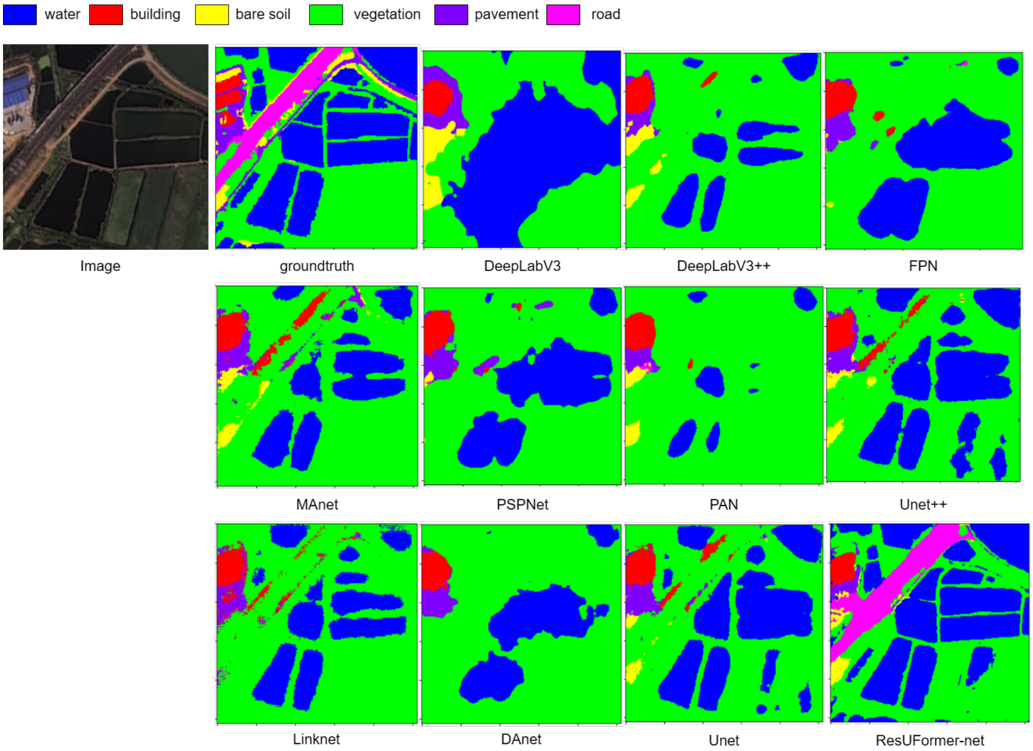

Figure 4. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the dataset-WHDLD. The detection of edge information for the whole image is the clearest compared to other networks, which fully reflects the global understanding ability.

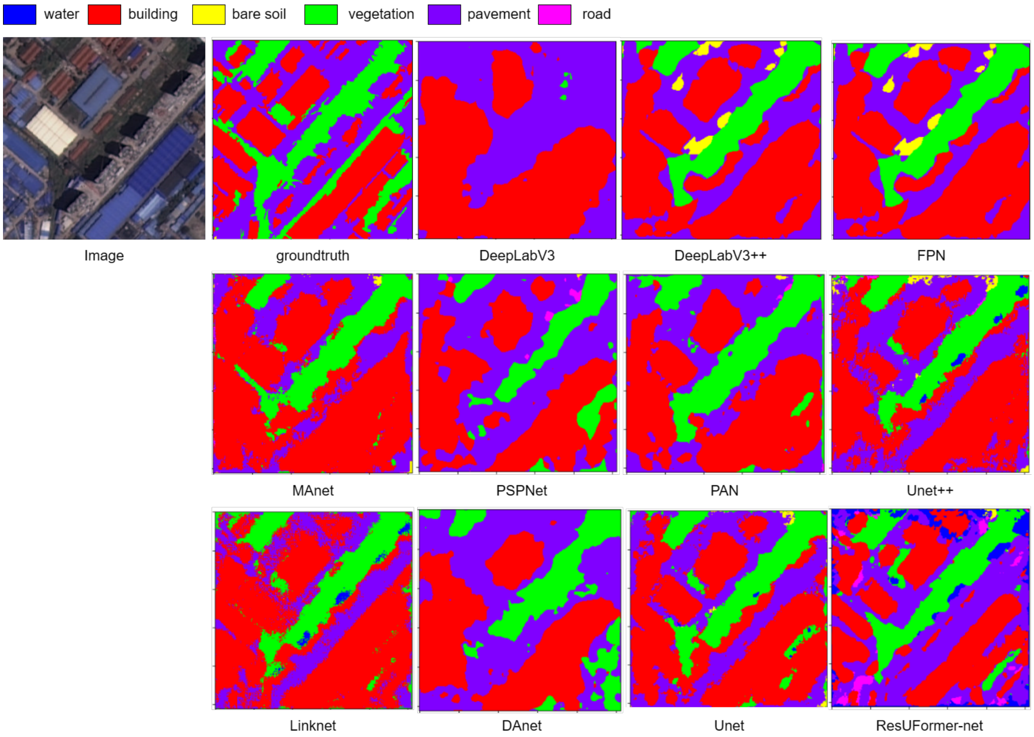

Figure 5. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the dataset-WHDLD. Pavements, buildings, and vegetation are densely distributed in terms of segmentation, and the accurate identification of each category reflects the global semantic understanding of the network, while the identification of dense segmentation is clearer compared to other networks.

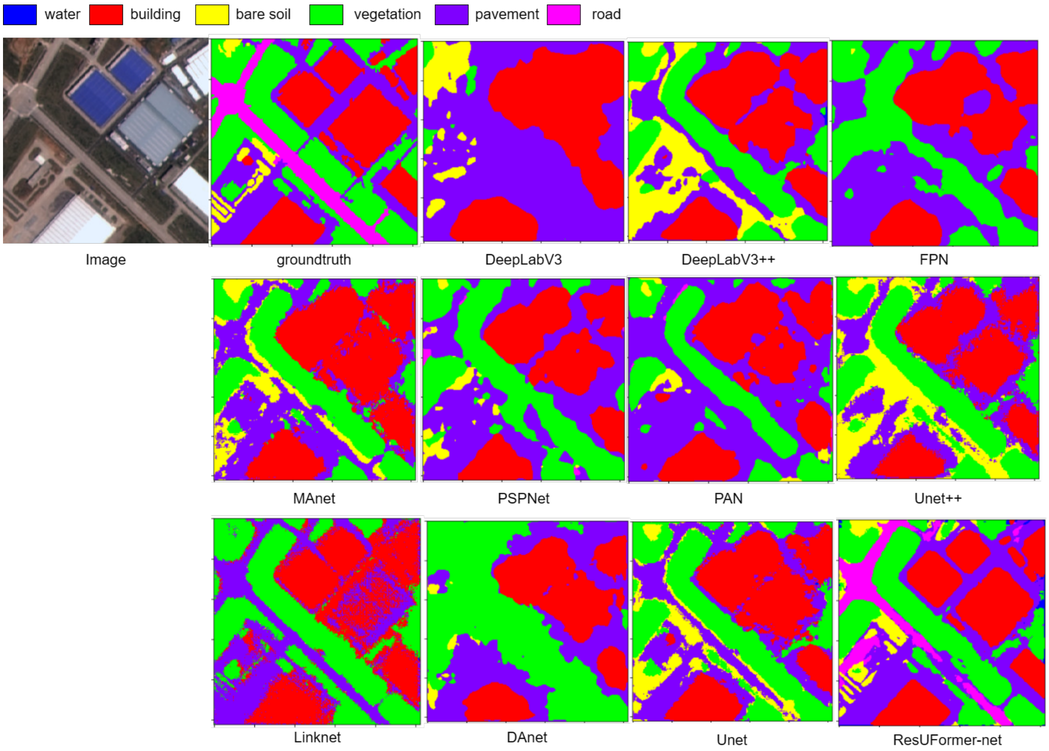

Figure 6. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the dataset-WHDLD. In particular, in the recognition of more classes, the semantic understanding is closest to the ground truth and the discrimination of individual objects is a good indication of the network's excellent understanding of local features.

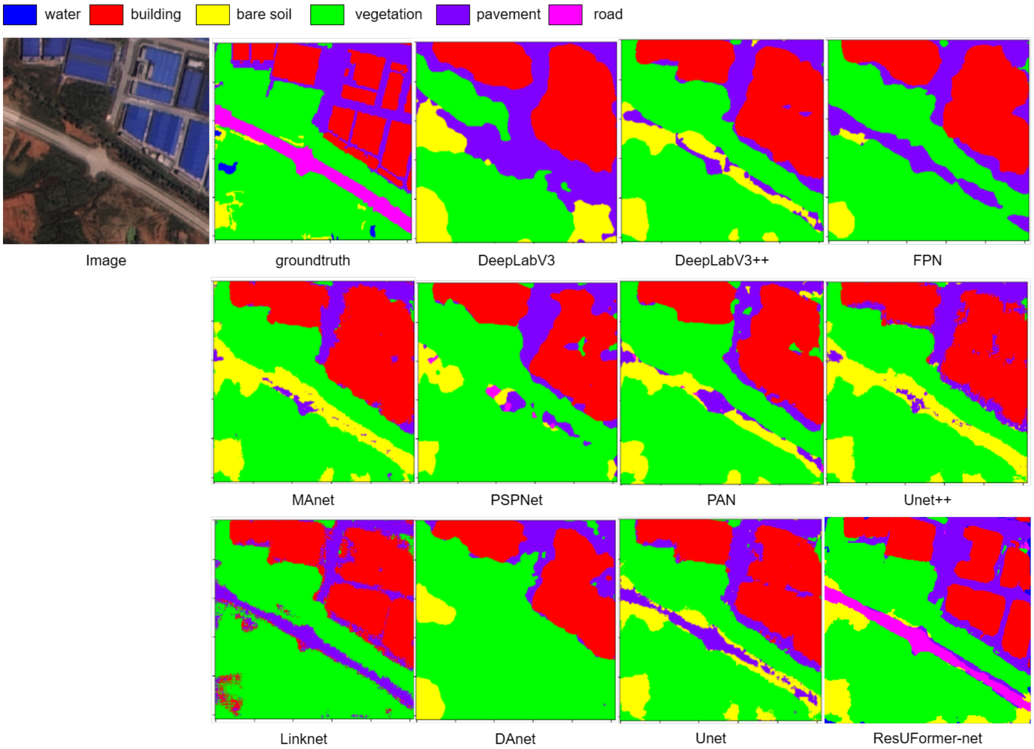

Figure 7. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the dataset-WHDLD. The recognition of roads and the segmentation of building clusters demonstrate the superiority of global semantic comprehension and local detail comprehension.

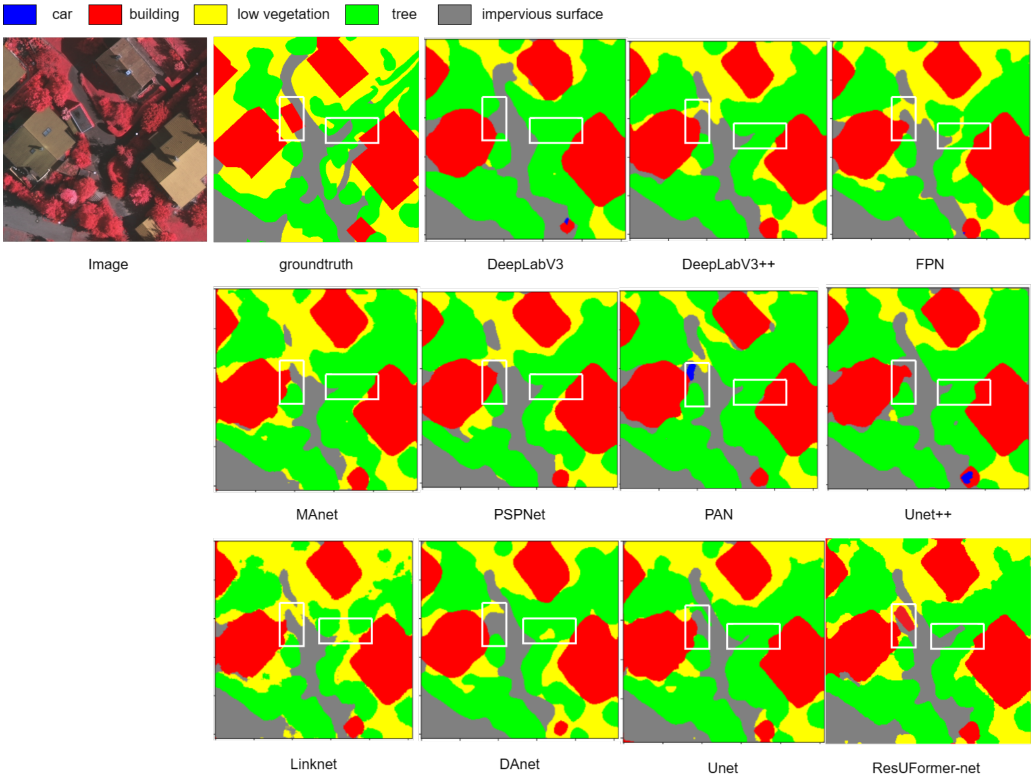

Figure 8. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the Vaihingen dataset. ResU-Former demonstrates a better local feature recognition capability in the identification of small features marked with white bounding boxes, as well as in recognizing the details of road extension.

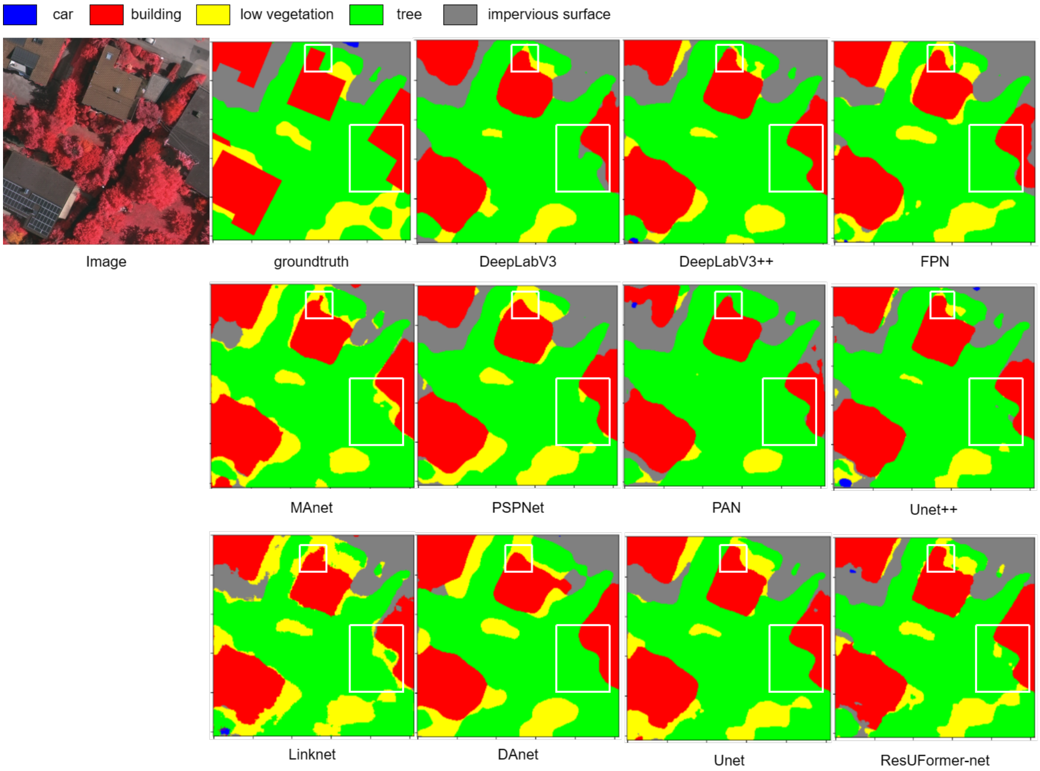

Figure 9. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the Vaihingen dataset.The network exhibits an excellent local feature recognition capability in accurately identifying the edges of objects marked with white bounding boxes.

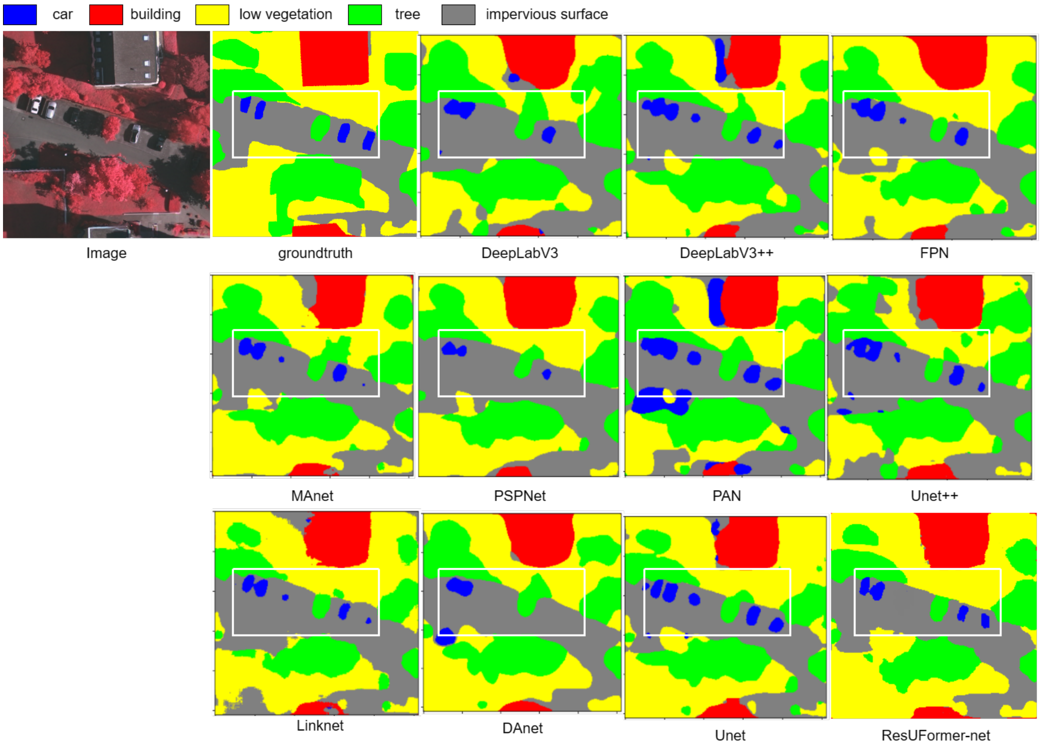

Figure 10. Including the ground truth and the original image, all images from other comparative 10 nets are shown above, from the Vaihingen dataset. The network exhibits excellent local semantic performance by accurately detecting and segmenting small vehicle features and shape of trees marked with white bounding boxes.

Table 5. All metrics of different multi-scale input resolutions from the postdom dataset are listed.

Scale	OA	MA	FWA	MeanIoU
400	0.695949	0.601041	0.537994	0.459527
300	0.721093	0.637878	0.570191	0.489534
256	0.732706	**0.684212**	0.590416	0.520947
200	**0.73987**	0.668063	**0.594294**	**0.522162**
185	0.726978	0.651337	0.577336	0.503144
154	0.644714	0.578341	0.475622	0.411014

Bolded data are optimal for each indicator.

While the metrics of the 154 × 154 scale are low because the input scale features are too small; this leads to the loss of image details and a decrease in resolution, which may result in blurry boundaries and confused categories, thereby reducing the prediction accuracy of the network. The poor performance of the ResU-Former Net on larger-scale images with dimensions of 400 × 400 can potentially be attributed to the limited memory resources for processing large-scale image features, leading to a loss of contextual information. Additionally, the increased number of parameters in the network makes it more prone to overfitting. Moreover, as the feature map size increases, the importance of pixel-level details decreases, which can also contribute to the poor performance of the network on larger-scale inputs.

Based on the analysis of the experimental results, future improvements in semantic segmentation performance will involve the implementation of more intricate multi-scale strategies, such as multi-scale fusion techniques.

4.5. Ablation Study

4.5.1. Ablation Experiments

In order to further analyze the role of each module in the algorithm, we conducted ablation experiments on the WHDLD dataset, which are divided into baseline, baseline and resnet, baseline and fusion loss, and baseline and fusion loss/and resnet. The results of the experiments are shown in Tables 6 and 7.

Table 6. All metrics of the ablation study from the WHDLD dataset are listed.

Net	OA	MA	FWA
baseline	0.746607	0.545161	0.603654
+resnet	0.762989	0.583239	0.637592
+fusion loss	0.748162	0.584657	0.616287
+resnet+fusion loss	**0.76847**	**0.63809**	**0.651536**

Bolded data are optimal for each indicator.

Table 7. The metrics IoU of the ablation study from the WHDLD dataset are listed.

Net	MeanIoU	Water	Building	Bare Soil	Vegetation	Pavement	Road
baseline	0.431722	0.80584	0.402778	0.31591	0.70123	0.257943	0.106629
+resnet	0.465716	0.84091	0.430923	0.312223	0.729559	**0.326153**	0.154525
+fusion loss	0.449872	0.80561	0.424664	0.325412	0.71271	0.284509	0.146325
+resnet/+fusion loss	**0.490832**	**0.845026**	**0.454164**	**0.328907**	**0.745774**	0.304944	**0.266176**

Bolded data are optimal for each indicator.

The introduction of Resnet, which means the shortcut connections embedded within the Swin Residual Transformer Blocks based on the designed successive Swin Transformer blocks, can make the network perform better due to the results shown in Tables 6 and 7. The baseline and resnet has an OA of 76.29%, which is an improvement of 1.63% compared to baseline, an MA of 58.32%, which is an improvement of 3.81% compared to baseline, an FWA of 63.76%, which is an improvement of 3.4% compared to baseline, and a Mean IoU of 46.57%, which is an improvement of 3.4% compared to baseline. This proves that the introduction of Resnet in the Swin Transformer can comprehensively improve the accuracy of network segmentation recognition as well as the performance of the network. This can further utilize the image pixel information, obtaining the evolution of information features on the basis of the Swin Transformer and U-shaped architecture.

The incorporation of fusion loss during training for the ResU-Former means an innovative loss function that synergistically fuses Soft Cross Entropy Loss with Lovasz Loss at a 1:1 weight ratio. Our approach is designed to leverage the strengths of both loss functions. Extensive experiments conducted on the dataset revealed that our fusion loss consistently outperforms traditional loss functions, delivering substantial improvements in key segmentation metrics such as mean IoU and MA compared to the baseline. The empirical results unequivocally substantiate the efficacy of our approach.

By adding both Resnet and fusion loss, the network performance improves significantly: OA is 76.85%, which is 2.19% higher than baseline, MA is 63.8%, which is 9.28% higher than baseline, FWA is 65.15%, which is 4.79% higher than baseline, and Mean IoU is 49.08%, an improvement of 5.91% compared to baseline, which proves that the introduction of fusion loss and Resnet combined further improves the performance of the network, with the metrics MA and Mean IoU particularly improving, verifying that the two modules are beneficial for the network to balance the overall ability to predict all classifications.

4.5.2. Discussion

The introduction of the Resnet, as well as the shortcut connections, improves the metrics like OA, MA, and MIou. Although the design of the Swin Transformer has introduced structures similar to residual connections, the ResU-Former's improvements prove the Resnet is not a redundant connection. The following aspects can be analyzed in terms of network performance improvement.

Shortcut connections allow gradients to flow directly through the network, which helps alleviate the problem of gradient vanishing in deeper networks and makes deeper model training feasible. Also, residual connections allow the network to directly access shallow features at deeper layers, which promotes feature reuse and may help the network learn fine-grained features better. Moreover, by adding shortcut connections, the capacity of the model can be increased without significantly increasing the computational burden, thus improving the expressiveness of the model. More importantly, in deep networks, features at different levels may contribute to the final task to different degrees, while shortcut connections help to combine these different levels of features efficiently.

Throughout the course of the training and the net improvements, the fusion loss with Soft Cross Entropy Loss and Lovasz Loss at a 1:1 weight ratio can apply to these subtasks.

Fusion loss can act as a regularizer to some extent by combining different loss functions, preventing the model from overfitting on a particular loss. Fusing these two loss functions allows the model to make progress in both pixel classification and segmentation quality, like in boundary optimization.

Furthermore, fusion loss can handle category imbalance. In remote sensing images, some categories may be more sparse or smaller in area than others, which can lead to the category imbalance problem. With fusion loss inserting Lovasz Loss, this problem can be alleviated, to some extent, because it focuses directly on IoU, which is a performance metric that does not depend on the category distribution.

The trick of fusion loss can also balance the learning focus for the net. By fusing the two loss functions, the model does not overly focus on one aspect, like optimizing only the pixel-level classification of responsible Soft Cross Entrophy loss, but finds a balance between the pixel-level classification and segmentation quality of responsible Lovasz Loss.

In practice, a 1:1 fusion ratio may not always be optimal, and the ratio needs to be adjusted according to the specific task, dataset, and model performance. For some tasks, more attention may need to be paid to boundary optimization, so the weight of Lovasz Loss can be increased; for other tasks, more attention may need to be paid to classification accuracy. Then, the weight of Soft Cross Entropy can be increased. The changeable weight fusion loss can make the network more generalized and robust.

5. Conclusions

The U-shaped symmetric encoder and decoder structure, using the Swin Residual Transformer for remote sensing images, embedded with scale adaptive block and combined with the fusion loss training trick, realizes the interaction of long-term, long-distance semantic information, supplements contextual information, and effectively balances the underlying and high-level features, which significantly advances its global–local recognition and promotes the net learning visual–semantic space. Moreover, it breaks the restriction of input features into the Swin Transformer and raises the network's generalization and efficiency.

ResU-Former is benchmarked on diverse datasets: WHDLD, Vaihingen, and postdom. Rigorous testing on several remote sensing datasets demonstrated the generalization of ResU-Former. Our method consistently outperformed existing mainstream nets across various landscapes and object classes. The metrics OA and MIoU are extra high, which fully demonstrates that the network's overall segmentation accuracy, full grasp of global semantic information, local detailed feature recognition, and average recognition for all categories are distinguished.

The neural net has evolved along with information, from the earlier eras of the net, which placed greater emphasis on the receptive fields of local features mostly improved by convolution, to the present era, which uses self-attention macro models to address the challenge of balancing the interaction of local and global semantic information.

The objectives set forth at the outset of this paper were successfully met, as evidenced by the enhanced performance in both urban and rural settings. Our contributions not only advance the field of remote sensing image segmentation by introducing an effective architectural innovation but can also fully solve the practical problems of multi-scale features in objects recognition in remote sensing imagery, imbalanced categories in terms of quantity and size, and the insufficient exchange of global semantic information in remote sensing scenes. ResU-Former also lies the groundwork for future exploration into self-attention frameworks that could further exploit the unique advantages of the Swin Transformer.

The wider impact of our work is significant, providing a robust foundation for applications that demand high precision in land cover and land use analysis. This includes environmental monitoring, urban planning, and disaster management. Our approach's adaptability and scalability make it a valuable asset for the remote sensing community as they tackle increasingly complex segmentation tasks.

This network requires some improvements in the future, such as the following:

1. Multi-scale fusion strategies in multi-scale experiments.
2. The incorporation of feature fusion modules or feature enhancement modules.

Author Contributions: Conceptualization, H.L.; methodology, L.L. and H.L.; software, L.L.; validation, H.L.; formal analysis, H.L.; investigation, H.L.; resources, H.L.; data curation, L.L. and H.L.; writing—original draft preparation, H.L.; writing—review and editing, L.Z.; visualization, H.L.; supervision, F.L.; project administration, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset available in a publicly accessible repository. The dataset presented in this study are openly available at https://www.isprs.org/education/benchmarks/UrbanSemLab/Default.aspx (accessed on 15 January 2024), reference number [30], and https://sites.google.com/view/zhouwx/dataset#h.p_hQS2jYeaFpV0 (accessed on 15 January 2024).

Conflicts of Interest: Author Lei Li was employed by the company Aerospace Tianmu (Chongqing) Satellite Science and Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. 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]
2. Li, X.; Liu, B.; Zheng, G.; Ren, Y.; Zhang, S.; Liu, Y.; Gao, L.; Liu, Y.; Zhang, B.; Wang, F. Deep-learning-based information mining from ocean remote-sensing imagery. *Natl. Sci. Rev.* **2020**, *7*, 1584–1605. [CrossRef] [PubMed]
3. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
4. 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]
5. Benediktsson, J.A.; Chanussot, J.; Moon, W.M. Very high-resolution remote sensing: Challenges and opportunities [point of view]. *Proc. IEEE* **2012**, *100*, 1907–1910. [CrossRef]
6. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature split-merge-enhancement network for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
7. Wen, L.; Chen, X.; Guo, P. A Comparative Study on Clustering Algorithms for Multispectral Remote Sensing Image Recognition. In Proceedings of the International Symposium on Neural Networks, Moscow, Russia, 10–12 July 2008; pp. 610–617.
8. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8748–8763.
10. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef] [PubMed]
11. 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.
12. Ghaffarian, S.; Valente, J.; Voort, M.V.D.; Tekinerdogan, B. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. *Remote Sens.* **2021**, *13*, 2965. [CrossRef]
13. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3286–3295.
14. 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 Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
15. 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.
16. 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.* **2023**, *45*, 87–110. [CrossRef] [PubMed]
17. 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 (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
18. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
20. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–9 May 2020; pp. 1055–1059.
21. 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.
22. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring plain vision transformer backbones for object detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 280–296.
23. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.

24. Wang, Q.; Dong, X.; Wang, R.; Sun, H. Swin transformer based pyramid pooling network for food segmentation. In Proceedings of the IEEE 2nd International Conference on Software Engineering and Artificial Intelligence, Xiamen, China, 10–12 June 2022; pp. 64–68.
25. Shi, W.; Xu, J.; Gao, P. SSformer: A lightweight transformer for semantic segmentation. *arXiv* **2022**, arXiv:2208.02034.
26. Yu, L.; Li, Z.; Zhang, J.; Wu, Q. Self-attention on multi-shifted windows for scene segmentation. *arXiv* **2022**, arXiv:2207.04403.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
28. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
29. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-over-Union Measure in Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4413–4421.
30. ISPRS 2D Semantic Labeling Dataset. 2021. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/Default.aspx (accessed on 15 January 2024).
31. 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.
32. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
33. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop (DLMIA 2018), and 8th International Workshop (ML-CDS 2018), Granada, Spain, 16–20 September 2018; pp. 3–11.
34. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
35. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
36. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
37. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
38. 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.
39. Liang, J.; Sun, G.; Zhang, K.; Gool, L.V.; Timofte, R. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4096–4105.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.