remote sensing

MDPI

Article

Hybridizing Cross-Level Contextual and Attentive Representations for Remote Sensing Imagery Semantic Segmentation

Xin Li [1,2,3], Feng Xu [1,2,*], Runliang Xia [3], Xin Lyu [1,2], Hongmin Gao [1] and Yao Tong [4]

[1] College of Computer and Information, Hohai University, Nanjing 211100, China; li-xin@hhu.edu.cn (X.L.); lvxin@hhu.edu.cn (X.L.); gaohongmin@hhu.edu.cn (H.G.)

[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; xiarunliang@hky.yrcc.gov.cn

[4] School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China; ytong@ha.edu.cn

* Correspondence: xufeng@hhu.edu.cn

Abstract: Semantic segmentation of remote sensing imagery is a fundamental task in intelligent interpretation. Since deep convolutional neural networks (DCNNs) performed considerable insight in learning implicit representations from data, numerous works in recent years have transferred the DCNN-based model to remote sensing data analysis. However, the wide-range observation areas, complex and diverse objects and illumination and imaging angle influence the pixels easily confused, leading to undesirable results. Therefore, a remote sensing imagery semantic segmentation neural network, named HCANet, is proposed to generate representative and discriminative representations for dense predictions. HCANet hybridizes cross-level contextual and attentive representations to emphasize the distinguishability of learned features. First of all, a cross-level contextual representation module (CCRM) is devised to exploit and harness the superpixel contextual information. Moreover, a hybrid representation enhancement module (HREM) is designed to fuse cross-level contextual and self-attentive representations flexibly. Furthermore, the decoder incorporates DUpsampling operation to boost the efficiency losslessly. The extensive experiments are implemented on the Vaihingen and Potsdam benchmarks. In addition, the results indicate that HCANet achieves excellent performance on overall accuracy and mean intersection over union. In addition, the ablation study further verifies the superiority of CCRM.

Keywords: semantic segmentation; remote sensing imagery; cross-level contextual information; representation enhancement

check for updates

Citation: 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. https://doi.org/10.3390/rs13152986

Academic Editors: Thien Huynh-The and Sun Le

Received: 4 July 2021
Accepted: 26 July 2021
Published: 29 July 2021

1. Introduction

Remote sensing imagery (RSI) semantic segmentation has been a fundamental task in interpreting and parsing the observation areas and objects [1]. It strives to assign pixel-level categorical labels for the image. In recent years, there has been a growing interest in performing semantic segmentation for multi-source remote sensing data since it is of great significance in spreading to various applications, such as urban planning [2,3], water resource management [4,5], precision agriculture [6,7], road extraction [8,9] and so forth.

Conventional methods combined hand-engineered features with machine learning models, such as random forests (RF) [10], support vector machine (SVM) [11], Markov random fields (MRFs) [12] and conditional random fields (CRFs) [13]. Unfortunately, these methods are conditioned on prior knowledge and expertise.

Afterward, DCNNs (deep convolutional neural networks) achieved satisfying performance by learning feature representations and classifier parameters simultaneously, which is deemed to learn more targeted features than conventional methods. Then, FCN (fully convolutional network) [14] has been devised to boost the accuracy by replacing the fully connected layer with the convolutional layer in standard CNN. To mitigate the transformation loss caused by FCN, an encoder-decoder architecture, termed as SegNet, was presented [15]. The core contribution of SegNet lies in mapping low-resolution features to input resolution for pixel-wise prediction. In the decoder stage, the max-pooling indices of the corresponding encoder, which partially retain discerning spatial information, are incorporated to perform non-linear upsampling. As a result, the boundary delineation is boosted without additional parameters. In addition, the satisfactory results both in accuracy and efficiency are demonstrated on several datasets. Specifically, encoder-decoder-based networks have been the dominant solutions in the semantic segmentation task of both natural and remote sensing imagery.

Unlike natural images, which are often more straightforward, current high-resolution remote sensing imagery (HRRSI) covers a wide range of Earth's surface and involves many objects. These objects presented in the same scene exhibit are varying in shape, color, scale, texture and spectrum. Moreover, their features are easily affected by weather, illumination, occlusion and imaging conditions [16]. Therefore, the high intra-class variance and low inter-class variance are widely distributed in HRRSI. However, the baseline encoder-decoder network is inherently limited by local receptive fields and short-range context information due to the fixed geometry structure. Consequently, the primary challenge is to produce refined representations, making the geo-objects more distinguishable and separable.

Striven to strengthen the learned representations with contextual information, numerous variants of encoder-decoder networks were investigated. Ronneberger et al. [17] proposed U-Net, which utilizes multiple skip connections to aggregation context in encoder and decoder. In addition, the results reveal slight improvement. Alternatively, Chen et al. [18] initially exploited Atrous Spatial Pyramid Pooling (ASPP) to capture multiscale context. Furthermore, recently published DeepLab V3+ [19] were enhanced versions embed a functional decoder module to optimize the feature maps. Inspired by DeepLab V3+, Li et al. [20] combined atrous convolution with deep neural forest, leading to an utterly differentiable decision forest, which could generate and leverage superpixel's contextual information. In addition, the experiments demonstrate the desired results on the ISPRS 2D Potsdam dataset.

Although enlarging receptive fields has made remarkable achievements, the expanding of receptive fields caused edge distortions. More specifically, lacking contextual information leads to uncertainty in predictions and missing spatial details provides obscure boundaries.

It is concluded that the context of one position typically refers to a set of positions, e.g., the surrounding pixels. The previous study focuses on the spatial scale of contexts. For instance, DANet (dual attention network for scene segmentation) [21], CFNet (Co-occurrent features in semantic segmentation) [22] and OCNet (object context network for scene parsing) [23] consider the relations between a position and its contextual positions. Incorporated with the produced attention map, the refined representations introducing the contextual positions' representations with higher or lower weights in accordance with the similarity. Inspired by the self-attention mechanism, SCAttNet (semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images) [24] was proposed to learn the attention map to aggregate contextual information for every point adaptively in RSI. Similarly, LANet (local attention embedding to improve the semantic segmentation of remote sensing images) [25] designed a patch attention module to enrich the local context information in addition to the global ones, leading to a competitive performance with fewer computations.

Primarily, the attention mechanisms have successfully verified the capability to capture position-wise dependencies of RSI. In addition, injecting these dependencies into the learnt representations is essential to strengthening the distinguishability of geo-object details to some extent. More importantly, while lots of research has been carried out on extracting and leveraging contextual information, there is still very little scientific understanding of modeling the correlations of the representations that come from different levels. We suppose that this is the main reason that makes the results far from optimal. A superpixel denotes an area generated by grouping pixels in remote sensing data, providing a more natural representation of image data than pixels. Therefore, in addition to independent pixel-level and superpixel-level representations, the cross-level correlations are beneficial for optimizing these two-level features, increasing the probability of correctly classifying the pixels.

Motivated by that the class label assigned to one pixel is the category of the superpixel that the pixel belongs to, we augment the representation of one pixel by exploiting the representation of the superpixel region of the corresponding class. Here, the HCANet (hybridizing cross-level contextual and attentive representations neural network) is proposed for remote sensing imagery semantic segmentation. In addition, the main contributions are as follows:

(1) A cross-level contextual representation module (CCRM) is devised to exploit and harness the superpixel contextual information. After learning superpixels under the supervision of ground truth, the superpixel regions are generated and represented by aggregating the pixels lying in the corresponding regions. Moreover, the cross-level contextual representation is produced by quantifying and formulating the correlations between pixel and superpixels. At last, the correlations are injected to augment the representativeness and distinguishability of features.

(2) A hybrid representation enhancement module (HREM) is designed to fuse cross-level contextual and self-attentive representations flexibly. As discussed above, the self-attention modules can facilitate pixel-wise representations effectively. Thus, a sub-branch that introduces non-local block to refine encoded feature maps is implemented. Afterward, this module adopts a concatenation operation followed by a 1 × 1 convolution layer to realize the injection of both two optimized representations before expansion.

(3) Integrating the above-designed modules, HCANet is constructed based on the encoder-decoder architecture for densely predicting the pixel-level labels. Furthermore, to losslessly boost decoder's efficiency, we implement DUpsampling (Data-dependent upsampling) [26] to recover feature maps with a one-step up-sampling instead of multiple times of up-sampling.

(4) The extensive experiments are implemented on ISPRS 2D Semantic Labeling Challenging for Vaihingen [27] and Potsdam [28]. In addition, the performance is evaluated by numerical indices and visualization inspections. In addition, the necessary ablation studies are conducted to verify the proposed method further.

This article is organized as follows. Section 2 briefly introduces the related works on the topic of remote sensing imagery semantic segmentation and attention mechanism. Section 3 presents the network architecture and embedding modules. Section 4 designs the experiments and discusses the results. Finally, Section 5 draws the conclusions and presents the future directions.

2. Related Works

2.1. Semantic Segmentation of RSI

Semantic segmentation of RSI profits from the advancements of deep learning methods [29,30]. For example, Mou et al. [31] devised two network modules to discover relationships between any two positions of RS imagery. ScasNet (self-cascaded convolutional neural network) [32] utilized a self-cascade network to improve the labeling coherence with sequential global-to-local contexts aggregation. SDNF (superpixel-enhanced deep neural forest) [20] fused DCNNs with decision forests to form a training network with specific architecture. Marmanis et al. [33] captured the edges' detail for fine-tuning the semantic boundaries of objects. Likely, Edge-FCN [34] fuses prior edge knowledge and learnt representations, guiding the network to obtain better segmentation performance. Additionally, a remarkable neural network named ResUNet-a provided a framework for the task of semantic segmentation of monotemporal very high-resolution aerial images.

Based on U-Net, ResUNet-a innovatively integrates residual connections, optimized Dice loss function to realize accurate inference.

It is essential to refine the representations with sufficient contextual information, including edge, surrounding pixels, homogeneous positions. However, the methods mentioned above inject these clues in handcrafted fashion and not learnable.

2.2. Attention Mechanism

The attention mechanism is a strategy of allocating biased computational resources to highlight the informative parts. This mechanism allows the inputs to interact with each other "self" and determine what they should pay more attention to [35]. In other words, this mechanism layer stacking few parallelizable matrix calculations to form attentive maps, which provides an efficient tool to capture short and long-range dependencies. For example, a SE (squeeze-and-excitation) block is proposed [36], generating channel-wise attention. This block emphasizes the scene-relevant feature maps with spatial-irrelevant information. Convolutional block attention module (CBAM) integrated spatial and channel attention modules to contribute the representations with informative regions [37]. Moreover, the non-local block was devised for several visual tasks. In addition, the results indicate the superiority in accuracy and efficiency [38].

In addition to natural image processing, many attention-based works were advocated for RSI. In [39], the channel attention block was embedded for recalibrating the channel-wise feature maps. Cui et al. [40] exploited the attention mechanism to match the caption nouns with objects in RSI. Then, the global attention upsampling [41] was introduced to provide global guidance from high-level features to low-level ones. Specifically, both positional and channel-wise relations are captured and integrated with serial and parallel manners, producing reasonable cues for inference [42]. Moreover, SCAttNet [24] was proposed to learn the attention map to aggregate contextual information for every point adaptively in RS imagery. Along with the analysis of local context, LANet [25] was proposed to bridge the gap between high- and low-level features. The representations are refined by patch attention modules. As a result, the performance on ISPRS 2D benchmarks reaches the SOTA (state-of-the-art) over several attention-based methods. In the same way, CCANet (class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation) [43] authorized class information constraints to obtain exact long-range correlational context. Furthermore, the results on two RS datasets verify the fine-grained segmentation performance. Homoplastically, a cascaded relation attention module is designed to ascertain the relationship with channels and positions [44,45]. Sun et al. [46] designed a boundary-aware semi-supervised semantic segmentation network, generating satisfactory segments with limited annotated data.

Although numerous methods were studied and proposed incorporating attention mechanisms and achieved competitive performance, the relational context is insufficient. The previous approaches form superpixel regions unsupervisedly, contributing to uncertain and unreliable context, weakening the classification ability.

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 HCANet and embedded CCRM and HREM are illustrated and formally descripted.

As discussed in Section 1, in CCRM, we first model and formulate the correlations between pixel and corresponding superpixels to enhance the distinguishability of learnt representations. Then, HREM concatenates the superpixel contextual representations and self-attentive representations, generating the final representations for decoding.

The rest of this section first presents the relevant preliminaries used in the construction of HCANet. Specifically, the details of CCRM are presented after introducing the framework, including theoretical analysis, structure description and formalization.

3.1. Preliminaries

3.1.1. Non-Local Block

As a typical design of self-attention, the non-local block (NL) [38] calculates the spatial-wise and channel-wise correlations simultaneously by several matrix multiplications. In addition, this block is flexible for embedding into various frameworks. The topological architecture is illustrated in Figure 1 and the formal description of the NL is presented as follows.

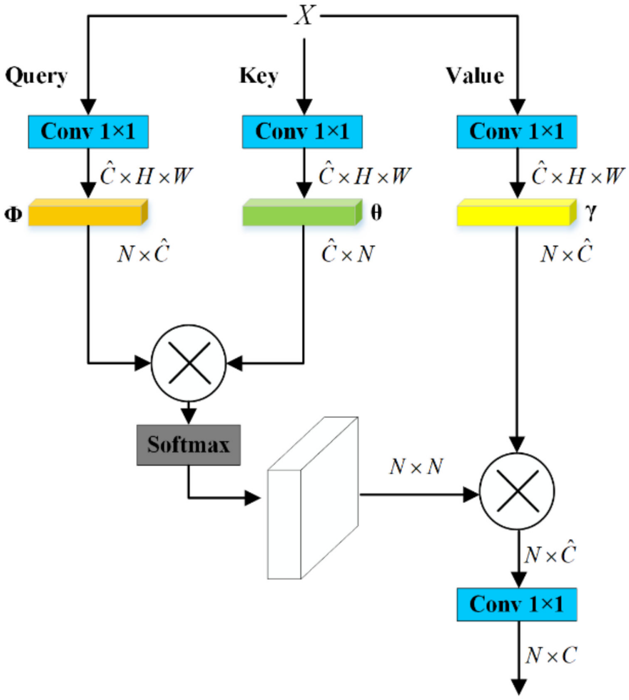

Figure 1. Non-local block.

Let the input feature be $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where C, H and W indicate the channel, height and width, respectively. Three 1×1 convolutions are used as transformation function to produce three diverse embeddings. Formally,

$$\phi = \mathbf{W}_{\phi}(\mathbf{X}) \tag{1}$$

$$\theta = \mathbf{W}_{\theta}(\mathbf{X}) \tag{2}$$

$$\gamma = \mathbf{W}_{\gamma}(\mathbf{X}) \tag{3}$$

where $\phi \in \mathbb{R}^{H\times W\times\hat{C}}$, $\theta \in \mathbb{R}^{H\times W\times\hat{C}}$, $\gamma \in \mathbb{R}^{H\times W\times\hat{C}}$ and $\hat{C}$ is the reshaped feature's channel number. Then, the feature maps are flatten to $\hat{C} \times N$, where $N = H \times W$. In addition, the similar matrix can be produced by,

$$\mathbf{V} = \phi^{\mathrm{T}} \times \theta \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{N\times N}$. After normalization, the similar matrix is transformed to $\vec{\mathbf{V}}$,

$$\vec{\mathbf{V}} = f(\mathbf{V}) \tag{5}$$

where f denotes normalization function. In this paper, the Softmax function is opted as f. As to every position in γ, the output of attention layer is formed as,

$$\mathbf{O} = \vec{\mathbf{V}} \times \gamma^{\mathrm{T}} \tag{6}$$

where $\mathbf{O} \in \mathbb{R}^{N \times \hat{C}}$. In general, the final output is given by

$$f_{\text{NL}} = \mathbf{Y} = \mathbf{W_O}\left(\mathbf{O}^{\text{T}}\right) + \mathbf{X} or \mathbf{Y} = \text{Cat}\left(\mathbf{W_O}\left(\mathbf{O}^{\text{T}}\right), \mathbf{X}\right) \tag{7}$$

where $\mathbf{W_O}$ is weights matrix by 1×1 convolution, $\mathbf{Y}$ is the refined feature maps of $\mathbf{X}$ and $\text{Cat}(\cdot)$ denotes the concatenation process.

Non-local block helps the network capture long-range dependencies. This simple yet efficient way is of great significance in semantic segmentation performance. However, only pixel-wise dependencies cross spatial and channel are considered with non-local block, ignoring the pixel-superpixel correlation, which is the vital aspect pivotal for characterizing pixels.

3.1.2. Superpixel Context

Superpixels in RSI are the grouped pixels, which can be explicitly seen as various image regions. These regions help the network learn a more real representation than pixels. Figure 2 shows the illustration of superpixel context. The red line indicates the pixel to be represented. In addition, the partial superpixel region is expected to be marked in the light cyan line. Modeling the relationships between the pixel and superpixel regions positively impacts the distinguishability and separability of encoded representations significantly.

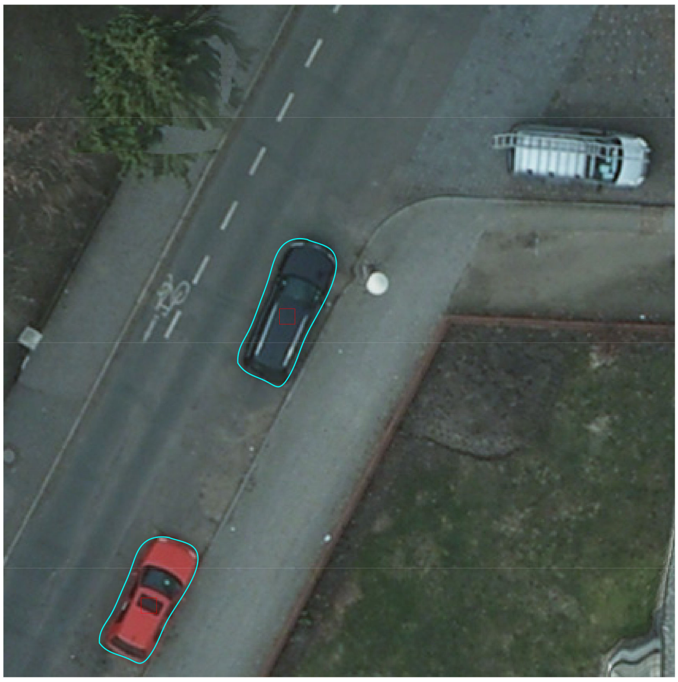

Figure 2. Illustration of superpixels.

For example, we abstractly deem the red box as one pixel lying in a black car object in Figure 2. Baseline networks, such as SegNet, U-Net and DeepLab V3+, extract the features from a regular local area with a size of 3 × 3, 5 × 5 or others (determined by configurations). If the pixel is located in the red car area, the learnt representations are very different. This intra-class inconsistency always leads to misclassification. If we can find a way to represent the similarity between the annotated pixel and the two cars with different colors, the intra-class inconsistency would alleviate automatically by the network.

Resort to the previous studies on similarity matrix analysis, it is achievable to capture the superpixel context and model the pixel-superpixel correlations in a learnable way using an attentive fashion. Hence, motivated by this target, the CCRM is investigated and formulated.

3.1.3. DUpsampling

As illustrated in Figure 3, the pipeline of DUpsampling is presented [26]. The specific pixel in the feature map with a lower resolution is analogously inferred to a 2 × 2 region by a ratio of 2 in DUpsampling.

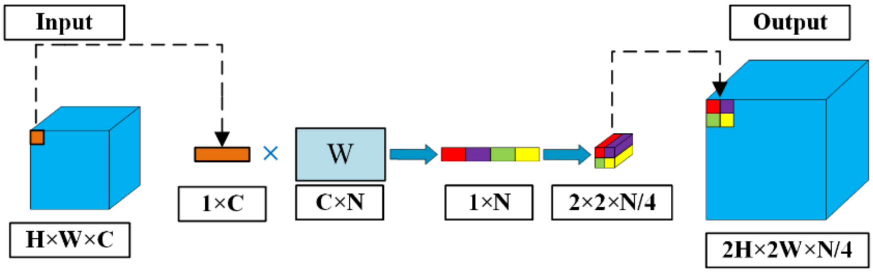

Figure 3. Pipeline of DUpsampling.

Let $F_D \in \mathrm{R}^{\widetilde{H} \times \widetilde{W} \times \widetilde{C}}$ denotes the features to be up-sampled and G indicates the ground truth. As a prerequisite, G can be compressed without loss. Commonly, G is encoded as $G \in \{0,1\}^{H \times W \times C}$, which is a one-hot encoding fashion. The ratio between F_D and G is represented as,

$$\frac{\widetilde{H}}{H} = \frac{\widetilde{W}}{W} = \frac{1}{16} or \frac{1}{32} \tag{8}$$

In the decoder stage, $F_D \in \mathrm{R}^{\widetilde{H} \times \widetilde{W} \times \widetilde{C}}$ is going to be up-sampled to the same spatial size with G. Then, the loss calculation is formed as,

$$L(F_D, G) = Loss(\mathrm{softmax}(B(F_D)),G) \tag{9}$$

where $B(\cdot)$ is the bilinear up-sampling used in former architecture, such as SegNet, FCN, U-Net and so forth. With the incorporation of DUpsampling, the loss function is formed as,

$$L(F_D, G) = Loss(\mathrm{softmax}(D(F_D)),G) \tag{10}$$

where $D(\cdot)$ represents the DUpsampling.

It is supposed that the ground truth label G is not i.i.d. Therefore, G could be compressed without information loss. To determine the transformation matrix W, Tian et al. [26] proposed a learnable way. Firstly, G is compressed to a specific target spatial resolution $\widetilde{G} \in \mathbb{R}^{\widetilde{H} \times \widetilde{W} \times \widetilde{C}}$, which keeps the same size with F_D. In addition, the criteria of W is minimizing the loss between $L(F_D, \widetilde{G})$ and $L(F_D, G)$. Since the spatial correlation of the label is learned, the recovered full size of predictions are more reliable.

As indicated in [26], DUpsampling significantly reduces the computation time and memory footprint of the semantic segmentation method by a factor of 3 or so. In addition, DUpsampling also allows a better feature aggregation to be exploited in the decoder by enlarging the design space of feature aggregation. In addition, Li et al. [47] have proved the efficiency and efficacy for RSI semantic segmentation task with DUpsampling.

3.2. The Framework of HCANet

As previously discussed, contextual information is of great importance in feature optimization. The atrous convolution-based networks, such as DeepLab V3+ and ResUNet-a, cost too much time and space to enrich the contextual information by enlarging the local receptive fields. However, the local information is still limited and insufficient. Alternatively, NLNet [38], CBAM [37], DANet [21], OCNet [23] and SCAttNet [24] employs an attention mechanism to capture long-range dependencies at position-wise. Unfortunately, only focusing on pixel-wise dependencies is not enough to learn the implicit correlations, in which the intra-class inconsistency and inter-class similarity always deteriorate the segmentation performance.

Attempt to optimize the learnt representations by extracting and leveraging richer contextual information, the HCANet is devised. The overall framework is presented in Figure 4. In general, HCANet is based on encoder-decoder architecture. In addition, two modules are designed to enhance the representations. One is CCRM that exploits superpixel context and captures the correlations between pixel and corresponding superpixel. Then, injecting these correlations to pixel-level representations to produce superpixel enhanced representations. Specifically, we first model and formulate the cross-level correlations between pixels and superpixels in RSI. The other is HREM that concatenates the self-attentive and superpixel enhanced representations to generate refined representations for decoding. Furthermore, to alleviate the loss of upsampling in the decoder stage, DUpsampling is incorporated. Finally, the Softmax classifier is employed to predict the pixels densely. The rest of this section will explain CCRM, HREM and DUpsampling in detail.

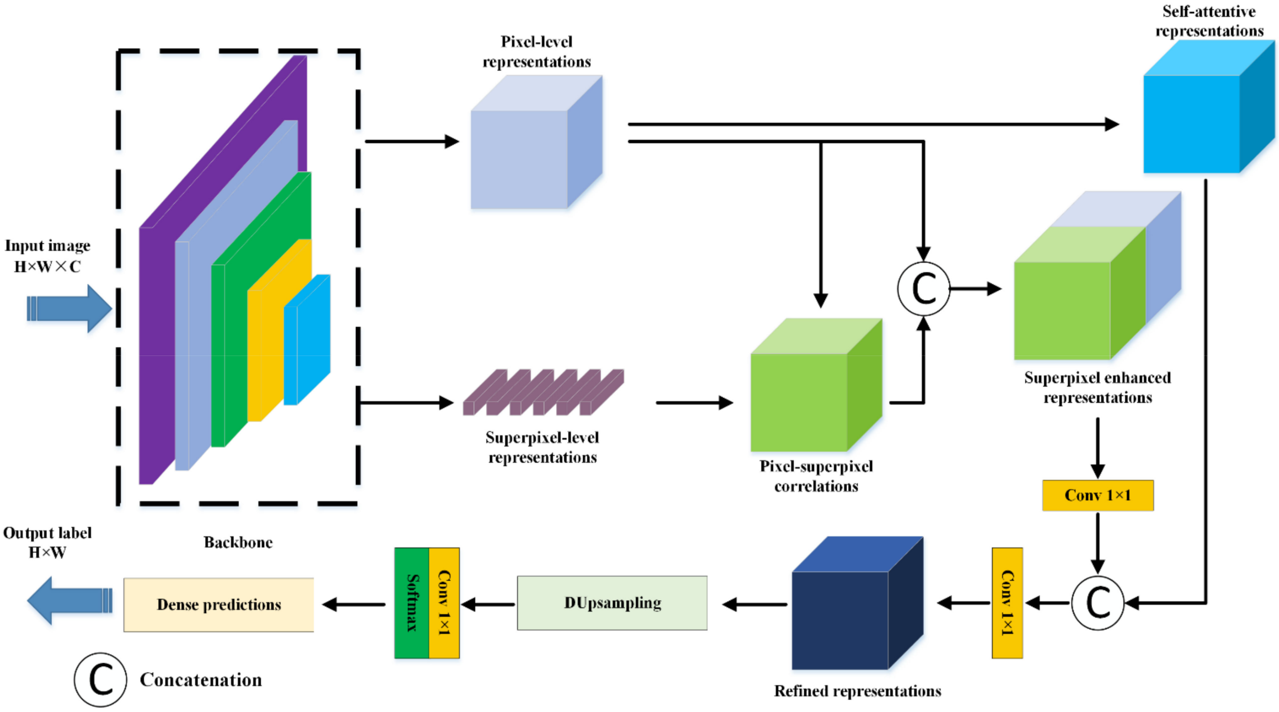
Input image
H×W×C
Backbone
Pixel-level representations
Superpixel-level representations
Self-attentive representations
Pixel-superpixel correlations
C
Superpixel enhanced representations
Conv 1×1
C
Conv 1×1
Refined representations
DUpsampling
Conv 1×1
Softmax
Dense predictions
Output label
H×W
C Concatenation

Figure 4. Overall framework of HCANet.

3.3. Cross-Level Contextual Representation Module

3.3.1. Superpixel Region Generation and Representation

First, of all, the superpixel regions is generated by coarse segmentations, which derives from the intermediate feature maps from backbone. During training, the ground-truth segmentation is investigated to improve the coarse regions by cross-entropy loss supervisiedly.

Let an input image denoted as $\mathbf{F}$, the superpixel regions $[\mathbf{SP}_1, \mathbf{SP}_2, \ldots, \mathbf{SP}_k]$ refer to the number of category k. Then, the representation of superpixel is obtained by aggregating all the pixels' represenitons in the corresponding superpixel region, formally,

$$\mathbf{F}_k = \sum_{i \in F} \alpha_{ki} \mathbf{X}_i \tag{11}$$

where $\mathbf{F}_k$ denotes the kth superpixel's representation, $\mathbf{X}_i$ is the representation of the corresponding pixel p_i, α_{ki} computes the degree for pixel p_i belonging to the kth superpixel. In practical implementation, spatial softmax is applied to normalize superpixel region $\mathbf{SP}_k$.

In the process of experiments, the superpixel regions are produced by coarse segments. In addition, the coarse segments are initially realized using the encoded feature maps of the backbone without any extra computations.

3.3.2. Cross-Level Contextual Representation

As depicted in Figure 5, the pipeline of CCRM, cross-level contextual representation, is concretely introduced to explain the formation of superpixel enhanced representations.

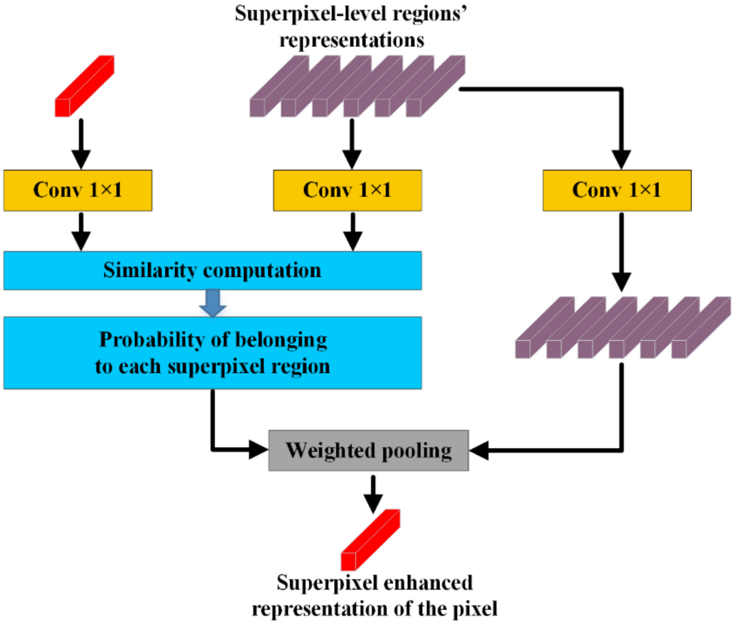
Superpixel-level regions' representations
Conv 1×1
Conv 1×1
Conv 1×1
Similarity computation
Probability of belonging to each superpixel region
Weighted pooling
Superpixel enhanced representation of the pixel

Figure 5. Pipeline of CCRM.

After the computation of superpixels' representations, we calculate the relationship between each pixel and each superpixel region as follows,

$$\beta_{ik} = \frac{e^{g(\mathbf{X}_i,\mathbf{F}_k)}}{\sum_{j=1}^{k} e^{g(\mathbf{X}_i,\mathbf{F}_j)}} \tag{12}$$

where $g(\mathbf{X}_i, \mathbf{F}_k)$ serves as the relation function and formed as,

$$g(\mathbf{X}_i, \mathbf{F}_k) = \theta(\mathbf{X}_i)^{\mathrm{T}} \varphi(\mathbf{F}_k) \tag{13}$$

where $\theta(\cdot)$ and $\varphi(\cdot)$ are transformation functions implemented by 1×1 convolution followed by BN and ReLU layers.

Therefore, the superpixel contextual representation $\mathbf{F}_{sp}(i)$ of pixel p_i is,

$$\mathbf{F}_{sp}(i) = \delta\left(\sum_{k=1}^{K} \beta_{ik}\rho(\mathbf{F}_k)\right) \tag{14}$$

where $\delta(\cdot)$ and $\rho(\cdot)$ are transformation functions implemented by 1×1 convolution followed by BN and ReLU layers, K is the max categories. Overall, the calculation of cross-level correlations and superpixel context are inspired by non-local fashion [38].

Finally, the CCRM produces the superpixel enhanced representations by aggregating two vital parts of context, (1) original pixel-level representations and (2) superpixel contextual representations. Formally,

$$\mathbf{F}_{sr}(i) = \mathrm{Cat}(\mathbf{X}_i, \mathbf{F}_{sp}(i)) \tag{15}$$

where $\mathbf{F}_{sr}(i)$ denotes the superpixel enhanced representations of pixel p_i, $\mathbf{X}_i$ is the original pixel-level representations, $\mathbf{F}_{sp}(i)$ refers to the superpixel contextual representation of p_i and $\mathrm{Cat}(\cdot)$ denotes the concatenation process.

3.4. Hybrid Representation Enhancement Module

The fusion of superpixel enhanced representations and self-attentive pixel-level representations is of great significance to comprehensively leverage contextual information.

Non-local blocks, a kind of self-attention module, have been examined and evaluated in feature optimization [38]. This block helps the network captures and retains positional dependencies, yielding desired enhancements of representations.

With the intention of refining learnt representations, the self-attentive representations are beneficial along with superpixel context. Thereby, another concatenation operation is embedded, which facilitating the representations further. Formally,

$$\mathbf{F}_r(i) = \mathrm{Cat}(\mathbf{F}_{sp}(i), \mathbf{F}_n(i)) \tag{16}$$

where $\mathbf{F}_r(i)$ denotes the refined representation of pixel p_i, $\mathbf{F}_{sp}(i)$ is the superpixel contextual representation of pixel p_i, $\mathbf{F}_n(i)$ refers to the self-attentive representation by non-local block of pixel p_i and $\mathrm{Cat}(\cdot)$ denotes the concatenation operation. Therefore, the output representations of HREM hybridizes original pixel-level representations, superpixel contextual representations and self-attentive representations. For the sake of understanding, 1×1 convolution after concatenation is hidden. Finally, $\mathbf{F}_r$ is fed to be upsampled and classified.

The HREM allows a simple yet effective concatenation operation to fuse the two refined representations. One of them derives from the self-attention module with the contextual information of position-wise dependencies. The other one comes from the CCRM by injecting the pixel-superpixel correlations. Therefore, the comprehensive and hybrid learnt representations could provide reasonable and distinguishable cues for dense prediction.

4. Experiments

4.1. Datasets

To evaluate the performance of the proposed method, extensive experiments are conducted on two ISPRS benchmarks. The data properties are presented in Table 1.

Table 1. Data properties.

Datasets	Vaihingen	Potsdam
Bands used	NIR, R, G	NIR, R, G
GSD	9 cm	5 cm
Sub-patch size	256 × 256	
Data augmentation	Rotate 90, 180 and 270 degrees, Horizontally and vertically flip	

4.1.1. ISPRS Vaihingen Dataset

The public 2D semantic labeling benchmark Vaihingen dataset is released by the International Society for Photogrammetry and Remote Sensing [27]. It contains high-resolution true orthophoto tiles and corresponding digital surface models as well as labeled ground truth. Each tile consists of three spectral bands, red (R), green (G) and near infrared (NIR). The spatial size is around 2500 × 2000 with GSD (ground sample distance) of 9 cm. The available 16 images are partitioned randomly, in which 11 images are for training and 5 for validation and test. The labeled datasets' ground truth is made up of 6 categories: impervious surfaces, building, low vegetation, tree, car and clutter/background.

4.1.2. ISPRS Potsdam Dataset

The 2D semantic labeling benchmark Potsdam dataset [28] is composed of 38 high resolution images of size 6000 × 6000 pixels, with spatial resolution of 5 cm. Similarly, 5 categories are labeled. In addition, the 24 available images are divided into training and validation set. Test set is same to validation set.

4.2. Implement Details

Considering the practical data properties, the matched digital surface models are not involved in experiments. The same sub-patch size and data augmentations are implemented on raw data.

The dataset for training, validation and test are divided randomly and automatically. Subject to the data size, the validation is also the test set. As to the Vaihingen dataset, training set is formed with 11 images (ID: 1, 3, 5, 7, 13, 17, 21, 23, 26, 32 and 37). The validation and test set consists of 5 images (ID: 11, 15, 28, 30 and 34). The training set of Potsdam dataset contains 16 images (ID: 2_10, 2_12, 3_10, 3_11, 4_11, 4_12, 5_10, 5_12, 6_8, 6_9, 6_10, 6_11, 7_7, 7_9, 7_11 and 7_12). In addition, the validation and test set include 7 images (ID: 2_11, 3_12, 4_10, 5_11, 6_7, 6_12, 7_8 and 7_10). Practically, the raw images are cropped to sub-patches with spatial size of 256 × 256 during training.

In the experiments, the settings of hyper-parameters are listed in Table 2. Essentially, the backbone is ResNet 101, which is indicated by black dotted box in Figure 4. Commonly, all the models are implemented on the PyTorch framework version 1.4.1 with NVIDIA Tesla V100-32GB graphic card under Linux OS.

Table 2. Hyper-parameter settings.

Hyper-Parameters	Settings
Backbone	ResNet 101
Batch size	16
Learning strategy	Poly decay
Initial learning rate	0.002
Loss Function	Cross-entropy
Optimizer	Adam
Max epoch	500

4.3. Evaluation Metrics

Two numerical metrics, OA (Overall Accuracy) and mIoU (mean inter-section over union), are chosen to quantitatively evaluate the performance. Formally,

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

$$\mathrm{mIoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{18}$$

where TP denotes the number of true positives, FP denotes the number of true positives, FN denotes the number of false negatives, TN denotes the number of true negatives.

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

In this part, several mainstream methods are compared to analyze the performance. The comparative methods include typical encoder-decoder, attention-based and SOTA RSI semantic segmentation networks.

4.4.1. Results on Vaihingen Test Set

As outlined in the introduction, we reasoned that complementary context is helpful to refine learned representations, which are essential to provide sufficient cues for dense predictions. Table 3 reports the results on the Vaihingen test set. It is evident that the results of HCANet obtained are in exceptionally good agreement with expectation. As can be seen, HCANet performs high consistency with ground truth. Statistically, the OA and mIoU of HCANet, 83.83% and 75.46%, respectively, are the highest.

Table 3. Results on Vaihingen dataset. Accuracy of each category is presented in the OA/IoU form, and the bold number indicates the best.

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	Clutter	OA	mIoU
SegNet [15]	92.37/78.87	90.96/75.71	80.61/62.11	77.22/57.98	60.61/51.11	73.39/57.98	79.19	63.43
U-Net [17]	92.71/78.72	90.84/76.61	79.95/62.25	77.67/58.33	69.87/55.27	74.41/58.33	80.91	65.11
DeepLab V3+ [19]	93.55/80.51	89.96/77.52	81.15/66.61	78.11/63.02	70.49/58.83	77.52/63.02	81.80	69.46
CBAM [37]	93.44/82.77	90.01/78.21	82.27/65.53	77.97/62.09	71.11/66.57	75.59/62.09	81.73	69.75
DANet [21]	93.62/83.54	90.13/77.95	83.31/69.03	78.54/63.05	70.94/65.18	76.62/63.05	82.19	70.03
OCNet [23]	93.31/83.78	91.22/79.01	84.49/68.15	79.61/62.98	72.25/67.79	79.53/62.98	83.40	71.20
NLNet [38]	93.28/83.66	89.95/78.08	83.34/68.85	78.18/63.33	70.99/65.59	76.12/63.33	81.98	70.55
ResUNet-a [45]	93.50/85.81	**97.12**/**80.02**	85.21/69.58	85.83/65.51	79.92/70.06	**81.91**/65.51	87.25	74.42
SCAttNet [24]	89.13/83.18	92.58/79.33	84.97/69.19	82.31/62.57	75.50/67.38	82.23/62.57	84.45	71.63
HCANet	**93.69**/**86.68**	95.11/79.95	**86.67**/**70.04**	**87.71**/**69.38**	**82.21**/**72.55**	81.56/**69.38**	**87.83**	**75.46**

Meanwhile, the category-wise accuracy and IoU also verify the competitive performance of HCANet. Compare to SegNet and U-Net, the OA and mIoU are dramatically increased by about 9% and 12%. It should be noted that SegNet and U-Net are initially devised for natural images and biomedical images, in which the distinguishability of representations is undemanding. Moreover, DeepLab V3+ utilizes the atrous convolution to enlarge the receptive field, which paves an innovative way to capture more contextual information. Likewise, attempt to generate smooth objects' boundary, ResUNet-a adopts complicated multi-tasking inference strategy along with multiple tricks that own a large amount parameters, resulting in a higher OA and mIoU. Specifically, the building and clutter are delineated with the highest accuracy of 97% and 81%. Although satisfactory results are yielded, DeepLab V3+ and ResUNet-a are criticized for the prohibitive computation and GPU memory occupation. Striven to ultimately capture and leverage contextual information, attention-based methods are developed to improve the segmentation performance.

Unfortunately, CBAM, DANet and NLNet are slightly problematic in enhancing separability of representations that extracted from remote sensing images. This is because the disparity between remote sensing and natural imagery. The former one covers a variety of geomorphological details and is acquired with diverse conditions. The demandingly geo-distinguishable of learned representations is challenging. These three models are originally devised for natural images. Facing the remote sensing imagery, the robustness is deficient. NLNet, an advanced self-attention framework, only reaches about 82% in OA and 71% in mIoU. Furthermore, OCNet takes advantages of context to facilitate the performance, which are feasible both on natural and remote sensing images. In addition, the OA and mIoU reach more than 83% and 71%. Inspired by self-attention mechanism, SCAttNet is proposed for the purpose of remote sensing image semantic segmentation. As a result, the results are acceptable with a small amount of extra matrix multiplication.

Figure 6 presents the visualization comparison on random samples from Vaihingen test set. Compared to other methods, HCANet displays progressive performance by learning integrate and complex representations. In addition, the confusable boundary and pixels are considerably classified. For example, HCANet can efficiently segment each car without adhesion and delineate a complete shape. However, the other methods perform the adhesion and incompleteness of cars.

In summary, HCANet comprehensively resorts the pixel-superpixel correlations and pixel-wise attentive maps, revealing the way to refine learned representation. Undeniably, the OA and mIoU are as good as can be expected by hybridizing cross-level contextual and attentive representations of remote sensing images.

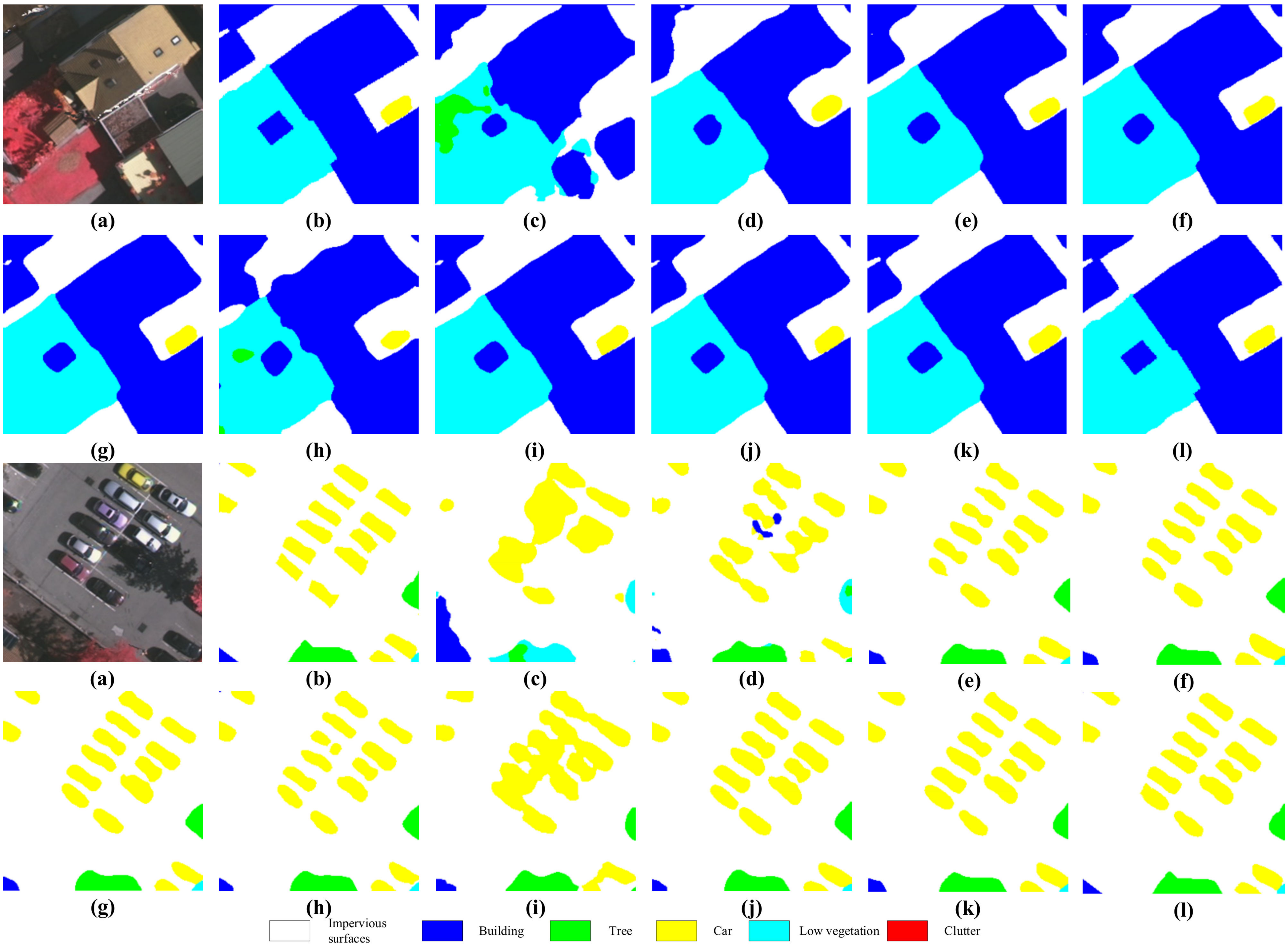

Figure 6. Visualization results on the Vaihingen test set. (**a**) Original image, (**b**) Ground truth, (**c**) SegNet, (**d**) U-Net, (**e**) DeepLab V3+, (**f**) CBAM, (**g**) DANet, (**h**) OCNet, (**i**) NLNet, (**j**) ResUNet-a, (**k**) SCAttNet, (**l**) HCANet.

4.4.2. Results on Potsdam Test Set

As presented in Table 4, the results on the Potsdam test set are collected. It is evident from the results that overall, there is a marked increase compared to other models. Moreover, the performance is almost identical to the Vaihingen test set. Beneficial from the sufficient data size for training, the OA and mIoU exhibit a minor increase with more than 1%. Nevertheless, the accuracy of impervious surfaces, building and clutter is marginally lower than DANet, ResUNet-a and SCAttNet, respectively. It could be concluded that the uncertainties and fluctuations are the dominant reasons. Furthermore, the gaps are narrow and negligible. After all, the OA and mIoU overwhelmingly indicate the superiorities of HCANet. Attention-based methods are of little significance. It is determined by the impractical application to remote sensing image semantic segmentation, which is more than an essential computer vision task. Then, ResUNet-a and SCAttNet incorporate the properties of the remote sensing domain, lending support to accurately labeling pixels. It is therefore desirable to understand the easily-confused pixels as well as possible.

Figure 7 compares the segmentation performance on random samples of test set. Although attention-based methods, SCAttNet and ResUNet-a have a certain degree of competitiveness, HCANet is of great significance due to the refined representations with high distinguishability, separating the heterogeneous pixels readily. For example, low vegetation and tree is relatively rough. In addition, HCANet exhibits acceptable performance.

Table 4. Results on Potsdam dataset. Accuracy of each category is presented in the OA/IoU form, and the bold number indicates the best.

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	Clutter	OA	mIoU
SegNet [15]	94.08/79.66	92.64/76.46	82.10/62.73	78.65/58.56	61.73/51.62	74.75/55.35	80.66	64.06
U-Net [17]	94.43/79.50	92.52/77.37	81.43/62.87	79.11/58.91	71.16/55.82	75.79/60.07	82.41	65.76
DeepLab V3+ [19]	95.28/81.31	91.62/78.29	82.65/67.27	79.56/63.65	71.79/59.42	78.95/70.97	83.31	70.15
CBAM [37]	95.17/83.59	91.68/78.99	83.79/66.18	79.41/62.71	72.43/67.23	76.99/63.96	83.24	70.44
DANet [21]	**95.35**/84.37	91.80/78.73	84.85/69.72	79.99/63.68	72.25/65.83	78.04/62.04	83.71	70.73
OCNet [23]	95.04/84.61	92.91/79.80	86.05/68.83	81.08/63.61	73.59/68.46	81.01/66.14	84.94	71.91
NLNet [38]	95.01/84.49	91.61/78.86	84.88/69.54	79.63/63.96	72.30/66.24	77.53/64.43	83.49	71.25
ResUNet-a [45]	95.22/**86.66**	**98.92**/80.82	86.79/70.27	87.42/66.16	81.40/70.76	83.43/**76.29**	88.86	75.16
SCAttNet [24]	90.78/84.01	94.29/80.12	86.54/69.88	83.83/63.19	76.90/68.05	**83.75**/68.81	86.02	72.34
HCANet	95.23/86.41	96.88/**81.55**	**88.29**/**71.44**	**89.35**/**70.77**	**83.74**/**74.00**	83.08/75.64	**89.44**	**76.64**

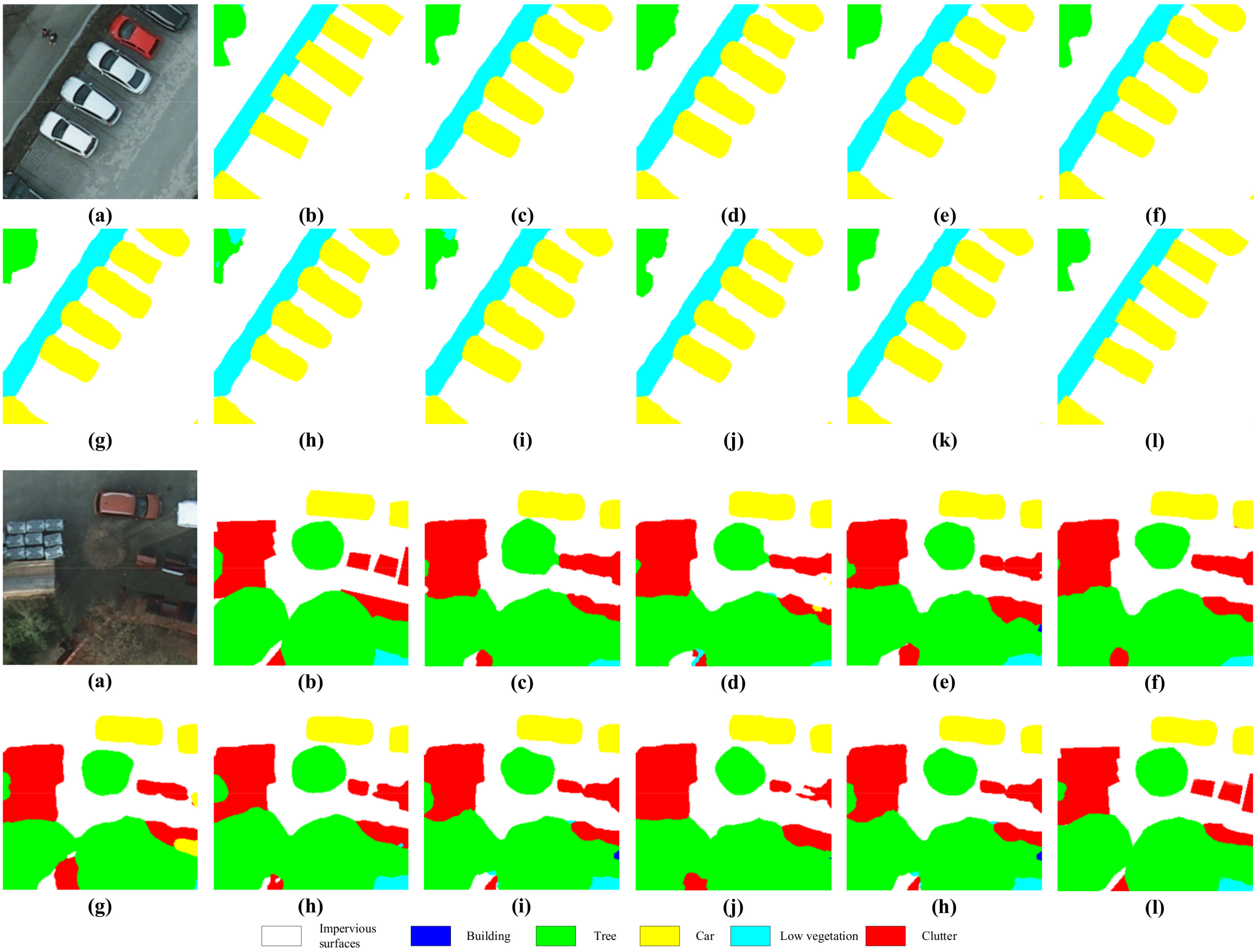

Figure 7. Visualization results on the Potsdam test set. (**a**) Original image, (**b**) Ground truth, (**c**) SegNet, (**d**) U-Net, (**e**) DeepLab V3+, (**f**) CBAM, (**g**) DANet, (**h**) OCNet, (**i**) NLNet, (**j**) ResUNet-a, (**k**) SCAttNet, (**l**) HCANet.

Overall, HCANet captures and fuses superpixel context, pixel-superpixel relationships and self-attentive maps, providing compelling cues for pixel-wise predictions. Hence, it appears to be apparent from the numerical results that the excellent prediction is agreed with the ground truth.

4.5. Ablation Study on CCRM

As reported in Section 1, refining learned representations with contextual information is profitable for producing sufficient cues for pixel-level predictions. Among the context, pixel-superpixel correlation is an essential element. HCANet captures and incorporates this kind of context in a self-attentive way. To comprehensively verify the efficiency and superiority, the ablation study on CCRM is designed and implemented. In addition, the non-CCRM version of HCANet is denoted as nc-HCANet.

Table 5 lists the OA/mIoU of two models tested on the three datasets. Compared to nc-HCANet, the OA increases by about 5% for three datasets and mIoU increases by about 4% correspondingly. The incorporation of superpixel enhanced representations lends support to boost the performance.

Table 5. Quantitative evaluations on two datasets. Accuracy in the OA/mIoU form, and the bold number indicates the best.

Models	Vaihingen	Potsdam
nc-HCANet	82.56/71.03	84.51/71.55
HCANet	**87.83/75.46**	**89.44/76.64**

In addition to test results, the training loss (cross-entropy) per epoch of two models is collected and illustrated in Figures 8 and 9. Accordingly, the training loss is much lower than nc-HCANet both for the Vaihingen (Figure 8) and Potsdam (Figure 9) training sets. After 500 epoches of training, CCRM module drops the loss to 0.0349 while nc-HCANet still has 0.185 for the Vaihingen dataset. As to Potsdam dataset, CCRM helps the network decrease training loss from 0.1973 to 0.0271. As a result, the reduction of loss means the inconsistency of probability distribution tends to be more acceptable.

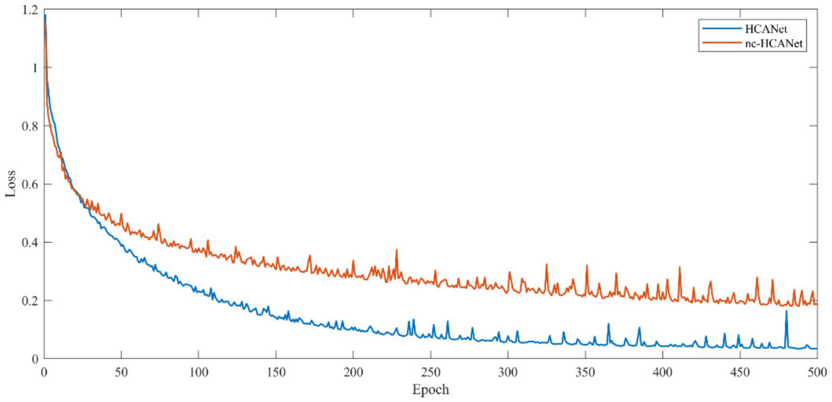

Figure 8. Training loss of Vaihingen dataset.

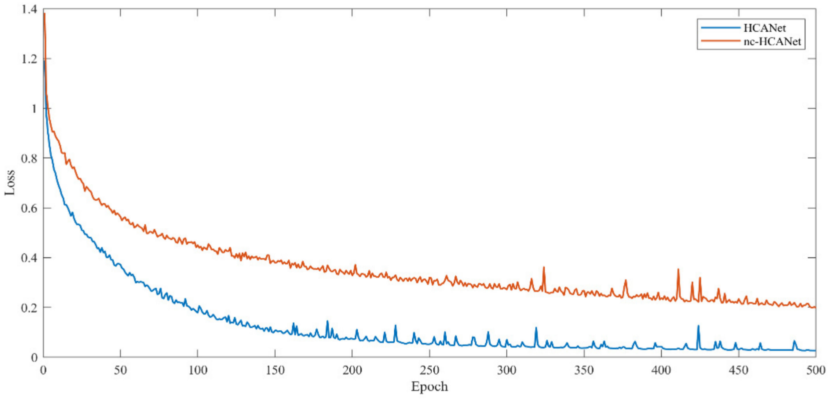

Figure 9. Training loss of Vaihingen dataset.

Apart from the accuracy comparison, the training time per epoch and inference time for a test image with spatial size of 256 × 256 are collected to emphasize the efficiency. As evident from Table 6, the training time per epoch is an average of 500 epochs. Therefore, concerning the significant increase in accuracy, the slight rise in time costs is acceptable. The inference time is generated by equally dividing the time costs for predicting the whole test set. As to single 256 × 256 image, the inference time grows up at 0.4 ms with the incorporation of CCRM.

Table 6. Efficiency comparisons.

Models	Training Time Per Epoch (Seconds)	Inference Time (Milliseconds)
nc-HCANet	376 ± 23	28.1
HCANet	391 ± 17	28.5

In summary, the improvements are intuitive and remarkable, owing to the utilization of more comprehensive contextual information, especially the pixel-superpixel correlation. Meanwhile, both training time and inference time are entirely negligible.

5. Conclusions

Striven to enhance the distinguishability of learned representations by semantic segmentation neural networks, HCANet is devised and implemented. First, of all, inspired by self-attention mechanism, the cross-level contextual representation module (CCRM) is designed to model the pixel-superpixel dependencies, which are injected to produce superpixel enhanced representations. Moreover, hybrid representation enhancement module (HREM) concatenates self-attentive representations implemented by non-local block with superpixel enhance representations generated by CCRM. Furthermore, the DUpsampling is embedded into decoder stage to recover feature maps to original spatial resolution losslessly.

The extensive experimental results provide a straightforward evidence that the complementary and complete contextual information enables high accuracy in pixel-wise semantic labeling. In addition, both short-range and long-range dependencies should be emphasized. Future work will focus on cross spatial resolution feature fusion in-depth at inexpensive time and space cost. In addition to capture the pixel-superpixel correlation of encoded feature maps, the shallow encoders' output feature maps should be further exploited.

Author Contributions: The individual contributions are listed below. Conceptualization, X.L. (Xin Li) and F.X.; methodology, X.L. (Xin Li), R.X., H.G. and X.L. (Xin Lyu); software, X.L. (Xin Li) and Y.T.; validation, X.L. (Xin Li) and X.L. (Xin Lyu); formal analysis, X.L. (Xin Li), X.L. (Xin Lyu) and Y.T.; investigation, X.L. (Xin Li), F.X. and H.G.; resources, X.L. (Xin Li) and R.X.; data curation, X.L. (Xin Li), H.G. and Y.T.; writing—original draft preparation, X.L. (Xin Li), F.X., H.G. and X.L. (Xin Lyu); writing—review and editing, X.L. (Xin Li), F.X. and H.G.; supervision, F.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Key Research and Development Program of China (Grant No. 2018YFC0407905, 2018YFC0407105 and 2017YFC0405505), the Technology Project of China Huaneng Group (Grant No. MW 2017/P28), the National Natural Science Fund of China (Grant No. 51779100 and 51679103), Central Public-interest Scientific Institution Basal Research Fund (Grant No. HKY-JBYW-2020-21 and HKY-JBYW-2020-07).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: [https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/].

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

References

1. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
2. Kouziokas, G.; Perakis, K. Decision support system based on artificial intelligence, GIS and remote sensing for sustainable public and judicial management. *Eur. J. Sustain. Dev.* **2017**, *6*, 397–404. [CrossRef]
3. Azimi, S.; Fisher, P.; Korner, M.; Reinartz, P. Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2920–2938. [CrossRef]
4. Duan, W.; Maskey, S.; Chaffe, P.; Luo, P.; He, B.; Wu, Y.; Hou, J. Recent advancement in remote sensing technology for hydrology analysis and water resources management. *Remote Sens.* **2021**, *13*, 1097. [CrossRef]
5. Zhang, X.; Jin, J.; Lan, Z.; Li, C.; Fan, M.; Wang, Y.; Yu, X.; Zhang, Y. ICENET: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. *Remote Sens.* **2020**, *12*, 221. [CrossRef]
6. Anand, T.; Sinha, S.; Mandal, M.; Chamola, V.; Yu, R.F. AgriSegNet: Deep aerial semantic segmentation framework for iot-assisted precision agriculture. *IEEE Sens. J.* **2021**. [CrossRef]
7. Du, Z.; Yang, J.; Ou, C.; Zhang, T. Smallholder crop area mapped with a semantic segmentation deep learning method. *Remote Sens.* **2019**, *11*, 888. [CrossRef]
8. Chen, Z.; Wang, C.; Li, J.; Xie, N.; Han, Y.; Du, J. Reconstruction bias U-Net for road extraction from optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 2284–2294. [CrossRef]
9. Wei, Y.; Zhang, K.; Ji, S. Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-Based segmentation and tracing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8919–8931. [CrossRef]
10. Yoo, C.; Han, D.; Im, J.; Bechtel, B. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. *ISPRS J. Photogramm. Remote Sens.* **2019**, *157*, 155–170. [CrossRef]
11. Wang, Y.; Yu, W.; Fang, Z. Multiple kernel-based SVM classification of hyperspectral images by combining spectral, spatial, and semantic information. *Remote Sens.* **2020**, *12*, 120. [CrossRef]
12. Zheng, C.; Zhang, Y.; Wang, L. Multigranularity multiclass-layer markov random field model for semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**. [CrossRef]
13. Kong, Y.; Zhang, B.; Yan, B.; Liu, Y.; Leung, H.; Peng, X. Affiliated fusion conditional random field for urban UAV image semantic segmentation. *Sensors* **2020**, *20*, 993. [CrossRef] [PubMed]
14. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
15. 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]
16. Liu, Y.; Xie, Y.; Yang, W.; Zuo, X.; Zhou, B. Target classification and recognition for high-resolution remote sensing images: Using the parallel cross-model neural cognitive computing algorithm. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 50–62. [CrossRef]
17. 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; Volume 9351, pp. 234–241.
18. Chen, L.; Papandreou, G.; Schroff, F.; Hartwig, A. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
19. Chen, L.; Papandreou, G.; Kokkinos, I.; 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] [PubMed]
20. Li, M.; Chen, Z. Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 140–152.
21. 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, CA, USA, 16–20 June 2019; pp. 3141–3149.
22. Zhang, H.; Zhang, H.; Wang, C.; Xie, J. Co-Occurrent Features in Semantic Segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 548–557.
23. Yuan, Y.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
24. Li, H.; Qiu, K.; Li, C.; 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.* **2020**. [CrossRef]
25. 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]
26. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3121–3130.
27. ISPRS. Vaihingen 2D Semantic Labeling Dataset. 2017. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html (accessed on 10 December 2017).

28. ISPRS. Potsdam 2D Semantic Labeling Dataset. 2017. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html (accessed on 10 December 2017).
29. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
30. Muhammad, A.; Wang, J.; Cong, G. Convolutional neural network for the semantic segmentation of remote sensing images. *Mob. Netw. Appl.* **2021**, *26*, 200–215.
31. Mou, L.; Hua, Y.; Zhu, X. A Relation-Augmented Fully Convolutional Network for Semantic Segmentation in Aerial Scenes. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12416–12425.
32. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 78–95. [CrossRef]
33. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
34. He, C.; Li, S.; Xiong, D.; Fang, P.; Liao, M. Remote sensing image semantic segmentation based on edge information guidance. *Remote Sens.* **2020**, *12*, 1501. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
36. Hu, J.; Li, S.; Samuel, A.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
37. Woo, S.; Park, J.; Lee, J.; Kweon, I. 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.
38. 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, CA, USA, 18–22 June 2018; pp. 7794–7803.
39. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasethiern, 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]
40. 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]
41. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. In Proceedings of the 29th British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.
42. Su, Y.; Wu, Y.; Wang, M.; Chen, J.; Lu, G. Semantic Segmentation of High Resolution Remote Sensing Image Based on Batch-Attention mechanism. In Proceedings of the 39th IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 3856–3859.
43. Deng, G.; Wu, Z.; Wang, C.; Xu, M.; Zhong, Y. CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**. [CrossRef]
44. Chen, J.; Wang, H.; Guo, Y.; Zhang, Y.; Deng, M. Strengthen the feature distinguishability of geo-object details in the semantic segmentation of high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 2327–2340. [CrossRef]
45. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
46. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS^{4}Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 5398–5413. [CrossRef]
47. 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]