remote sensing

MDPI

Article

Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images

Libo Wang [1,†], Rui Li [1,†], Dongzhi Wang [2,*], Chenxi Duan [3], Teng Wang [2] and Xiaoliang Meng [1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; wanglibo@whu.edu.cn (L.W.); lironui@whu.edu.cn (R.L.); xmeng@whu.edu.cn (X.M.)

[2] Surveying and Mapping Institute, Lands and Resource Department of Guangdong Province, Guangzhou 510500, China; wangteng43@hotmail.com

[3] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; chenxiduan@whu.edu.cn

* Correspondence: harwang1984@foxmail.com; Tel.: +86-38334381

† Equal contribution.

Abstract: Semantic segmentation from very fine resolution (VFR) urban scene images plays a significant role in several application scenarios including autonomous driving, land cover classification, urban planning, etc. However, the tremendous details contained in the VFR image, especially the considerable variations in scale and appearance of objects, severely limit the potential of the existing deep learning approaches. Addressing such issues represents a promising research field in the remote sensing community, which paves the way for scene-level landscape pattern analysis and decision making. In this paper, we propose a Bilateral Awareness Network which contains a dependency path and a texture path to fully capture the long-range relationships and fine-grained details in VFR images. Specifically, the dependency path is conducted based on the ResT, a novel Transformer backbone with memory-efficient multi-head self-attention, while the texture path is built on the stacked convolution operation. In addition, using the linear attention mechanism, a feature aggregation module is designed to effectively fuse the dependency features and texture features. Extensive experiments conducted on the three large-scale urban scene image segmentation datasets, i.e., ISPRS Vaihingen dataset, ISPRS Potsdam dataset, and UAVid dataset, demonstrate the effectiveness of our BANet. Specifically, a 64.6% mIoU is achieved on the UAVid dataset.

Keywords: urban scene segmentation; remote sensing; transformer; attention mechanism

check for updates

Citation: Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. https://doi.org/10.3390/rs13163065

Academic Editors: Sidike Paheding, Zahangir Alom, Maitiniyazi Maimaitijiang and Matthew Maimaitiyiming

Received: 25 June 2021
Accepted: 31 July 2021
Published: 4 August 2021

1. Introduction

Semantic segmentation of very fine resolution (VFR) urban scene images comprises a hot topic in the remote sensing community [1–6]. It plays a crucial role in various urban applications, such as urban planning [7], vehicle monitoring [8], land cover mapping [9], change detection [10], and building and road extraction [11,12], as well as other practical applications [13–15]. The goal of semantic segmentation is to label each pixel with a certain category. Since geo-objects in urban areas are characterized by large within-class and small between-class variance commonly, semantic segmentation of very fine resolution RGB imagery remains a challenging issue [16,17]. For example, urban buildings made of diverse materials show variant spectral signatures, while buildings and roads made of the same material (e.g., cement) exhibit similar textural information in RGB images.

Due to the advantage in local texture extraction, many researchers have investigated the challenging urban scene segmentation task based on deep convolutional neural networks (DCNNs) [18,19]. Especially, the methods based on fully convolutional neural network (FCN) [20], which can be trained end-to-end, have achieved great breakthroughs in urban scene labelling [21]. In comparison with the traditional machine learning methods, such as support vector machine (SVM) [22], random forest [23], and conditional random field (CRF) [24], the FCN-based methods have demonstrated remarkable generalization capability and high efficiency [25,26]. Therefore, numerous specially designed FCN-based networks have been spawned for urban scene segmentation, including UNet and its variants [4,16,27,28], multi-scale context aggregation networks [29,30], and multi-level feature fusion networks [5], attention-based networks [3,31,32], as well as lightweight networks [33]. For example, Sherrah [21] introduced the FCN to semantically label remote sensing images. Kampffmeyer et al. [34] quantified the uncertainty in urban remote sensing images at the pixel level, thereby enhancing the accuracy of relatively small objects (e.g., Cars). Maggiori et al. [35] designed an auxiliary CNN to learn the features fusion schemes. Multi-modal data were further utilized by Audebert et al. [36] in their V-FuseNet to enhance the segmentation performance. However, if either modality is unavailable in the test phase caused by sensors' corruption or thick cloud cover [37], such a multi-modal data fusion scheme will be invalid. Kampffmeyer et al. [38], therefore, proposed a hallucination network aiming to replace missing modalities during testing. In addition, enhancing the segmentation accuracy by optimizing object boundaries is another burgeoning research area [39,40].

The accuracy of FCN-based networks, although encouraging, appears to be incompetent for VFR segmentation. The reason is that almost all FCN-based networks are built on DCNNs, while the latter is designed for extracting local patterns and lacks the ability to model global context in its nature [41]. Hence, extensive investigations have been devoted to addressing the above issue since the long-range dependency is vital for segmenting confusing manmade objects in urban areas. Typical methods include dilated convolutional networks which are designed for enlarging the receptive field [42,43] and attentional networks that are proposed for capturing long-range relational semantic content of feature maps [31,44]. Nevertheless, these two networks have never been able to get rid of the dependence on the convolution operation, impairing the effectiveness of long-range information extraction.

Most recently, with its strong ability in long-range dependency capture and sequence-based image modelling, an entirely novel architecture named Transformer [45] has become prominent in various computer vision tasks, such as image classification [46], object detection [47], and semantic segmentation [48]. The schematic flowchart of the Transformer is illustrated in Figure 1a. First, the Transformer deploys a patch partition to split the 2D input image into non-overlapping image patches. (H, W) and C denotes the resolution and the channel dimension of the input image, respectively. (P, P) is the resolution of each image patch. Then, a flatten operation and a linear projection are employed to produce the 1D sequence. The length of the sequence is N, where $N = (H \times W)/P^2$. M is the output dimension of the linear projection. Finally, the sequence is fed into stacked transformer blocks to extract features with long-range dependencies. As shown in Figure 1b, a standard transformer block is composed of multi-head self-attention (MHSA) [45], layer norm (LN) [49], and multilayer perceptron (MLP) as well as two addition operations. L represents the number of transformer blocks. Benefiting from the non-convolution structure and attention mechanism, Transformer could capture long-range dependencies more effectively [50].

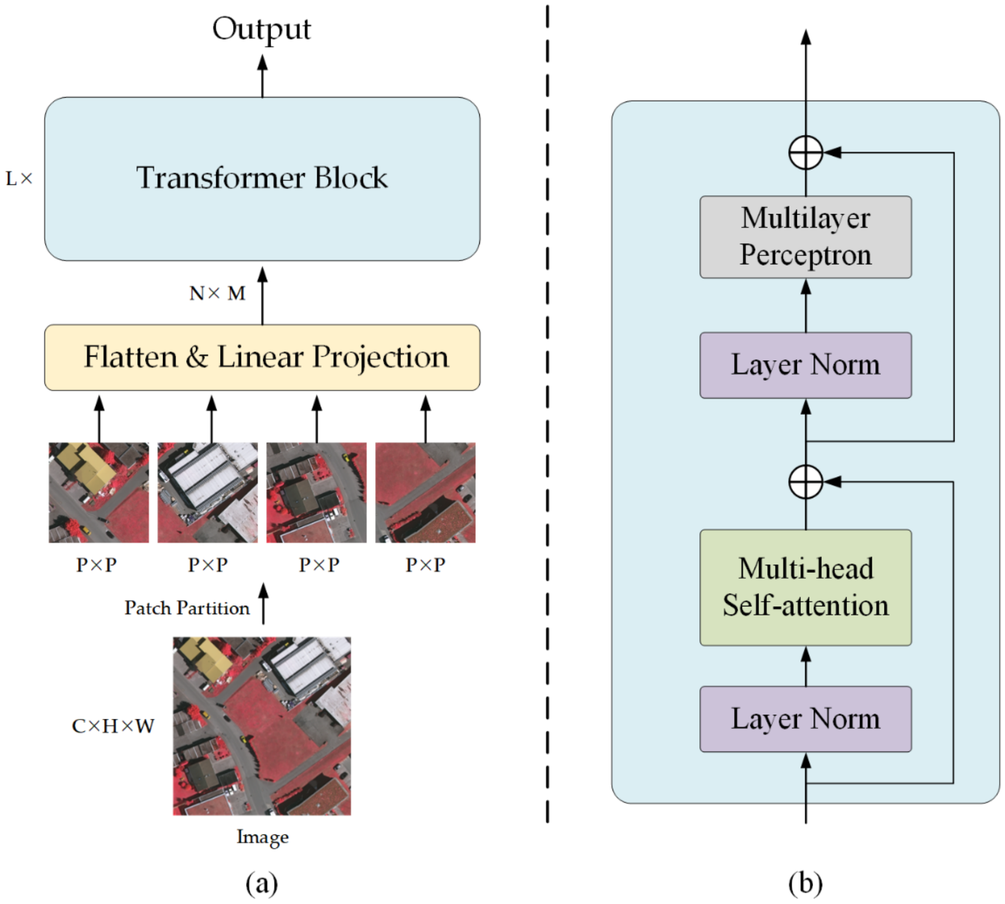

Figure 1. (**a**) Illustration of the schematic flowchart of the Transformer. (**b**) Illustration of a standard transformer block.

Inspired by the advancement of Transformer, in this paper, we propose a Bilateral Awareness Network (BANet) for accurate semantic segmentation of VFR urban scene images. Different from the traditional single-path convolutional neural networks, BANet addresses the challenging urban scene segmentation by constructing two feature extraction paths, as illustrated in Figure 2. Specifically, a texture path using stacked convolution layers is developed to extract the textural feature. Meanwhile, a dependency path using Transformer blocks is established to capture the long-range dependent feature. To leverage the benefits provided by the two features, we design a feature aggregation module (FAM) which introduces the linear attention mechanism to reduce the fitting residual of fused features, thereby strengthening the generalization capability of the network. Experimental results on three large-scale urban scene image segmentation datasets demonstrate the effectiveness of our BANet. In addition, the well-designed bilateral structure could provide a unified solution for semantic segmentation, object detection, and change detection, which undoubtedly boosts deep learning techniques in the remote sensing domain. To sum up, the main contributions of this paper are the following:

(1) A novel bilateral structure composed of convolution layers and transformer blocks is proposed for understanding and labelling very fine resolution urban scene images. It provides a new perspective for capturing textural information and long-range dependencies simultaneously in a single network.

(2) A feature aggregation module is developed to fuse the textural feature and long-range dependent feature extracted by the bilateral structure. It employs linear attention to reduce the fitting residual and greatly improves the generalization of fused features.

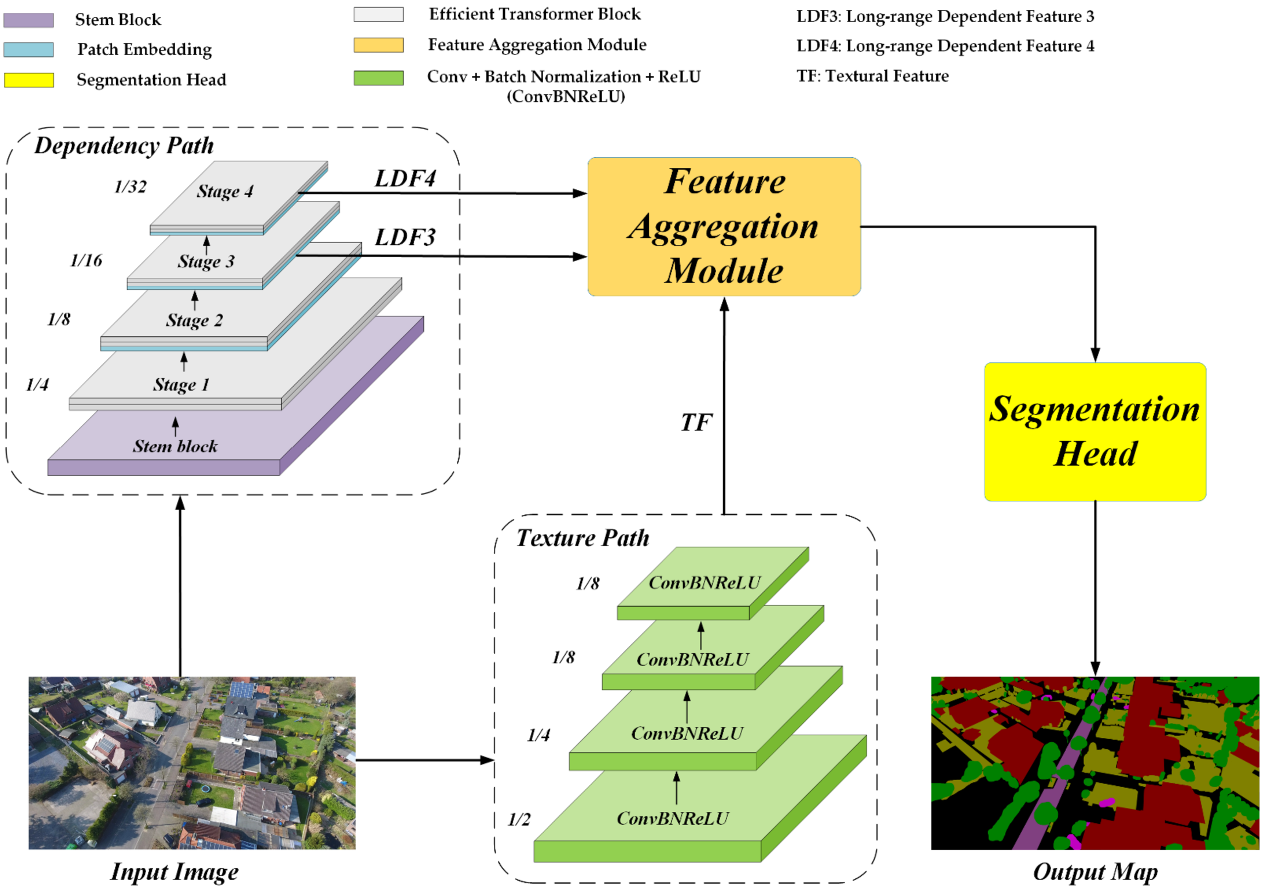

Figure 2. The overall architecture of Bilateral Awareness Network (BANet).

The remainder of this paper is organized as follows. The architecture of BANet and its components are detailed in Section 2. Experimental comparisons on three semantic segmentation datasets (UAVid, ISPRS Vaihingen, and Potsdam) are provided in Section 3. A comprehensive discussion is presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Bilateral Awareness Network

2.1. Overview

The overall architecture of the Bilateral Awareness Network (BANet) is exhibited in Figure 2, where the input image is fed into the dependency path and texture path simultaneously.

The dependency path employs a stem block and four transformer stages (i.e., Stage 1–4) to extract long-range dependent features. Each stage consists of two efficient transformer blocks (ETB). In particular, Stage 2, Stage 3, and Stage 4 involve patch embedding (PE) operations additionally. Proceed by the dependency path, two long-range dependent features (i.e., LDF3 and LDF4) are generated.

The texture path deploys four convolution layers to capture the textural feature (TF), while each convolutional layer is equipped with batch normalization (BN) [51] and ReLU activation function [52]. The downsampling factor is set as 8 for the texture path to preserve spatial details.

Since the outputs of the dependency path and the texture path are in disparate domains, FAM is proposed to merge them effectively. Whereafter, a segmentation head module is attached to convert the fused feature into a segmentation map.

2.2. Dependency Path

The dependency path is constructed by the ResT-Lite [53] pertained on ImageNet. As an efficient vision transformer, ResT-Lite is suitable for urban scene interpretation due to its balanced trade-off between segmentation accuracy and computational complexity. The main basic modules of the ResT-lite include the stem block, patch embedding and efficient transformer block.

Stem block: The stem block aims to shrink the height and width dimension and expand the channel dimension. To capture low-level information effectively, it introduces three 3 × 3 convolution layers with strides of [2, 1, 2]. The first two convolution layers are followed by BN and ReLU. Proceed by the stem block, the spatial resolution is downscaled by a factor of 4, and the channel dimension is extended from 3 to 64.

Patch embedding: The patch embedding aims to downsample the feature map for hierarchical feature representation. The output for each patch embedding can be formalized as

$$\mathrm{PE}(\mathbf{X}') = \mathrm{Sigmoid}(\mathrm{DWConv}(\mathbf{X}'))\cdot\mathbf{X}' \tag{1}$$

$$\mathbf{X}' = \mathrm{BN}(\mathrm{W_s}\cdot\mathbf{X}) \tag{2}$$

where $\mathrm{W_s}$ represents a convolution layer with a kernel size of s+1 and a stride of s. Here, s is set as 2. DWConv denotes a 3 × 3 depth-wise convolution [54] with a stride of 1.

Efficient transformer block: Each efficient transformer is composed of efficient multi-head self-attention (EMSA) [53], MLP and LN. The output for each efficient transformer block can be formalized as

$$\mathrm{ETB}(\mathbf{X}) = \mathrm{G}(\mathbf{X}) + \mathrm{MLP}(\mathrm{LN}(\mathrm{G}(\mathbf{X}))) \tag{3}$$

$$\mathrm{G}(\mathbf{X}) = \mathbf{X} + \mathrm{EMSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \tag{4}$$

The EMSA, a revised self-attention module for computer vision based on MHSA, is the main module of ETB. As illustrated in Figure 3, the detailed steps of EMSA are as follows:

(1) EMSA obtains three vectors $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ from the input vector $\mathbf{X} \in \mathbb{R}^{N\times D}$. Different from the standard multi-head self-attention, EMSA first deploys a depth-wise convolution with a kernel size of r+1 and stride of r to decrease the resolution of $\mathbf{K}$ and $\mathbf{V}$, thereby compressing the computation and memory. For the four transformer stages, r is set as 8, 4, 2, 1, respectively.

(2) To be specific, the input vector $\mathbf{X}$ is reshaped to a new vector with a shape of $D \times H \times W$, where $H \times W = N$. Proceed by the depth-wise convolution, the new vector is reshaped to $D \times h \times w$. Here, $h = H/r$ and $w = W/r$. Then, the new vector is recovered to $n \times D$ as the input of LN, where $h \times w = n$. Thus, the initial shape of $\mathbf{K}$ and $\mathbf{V}$ is $n \times D$. The initial shape of $\mathbf{Q}$ is $N \times D$.

(3) The three vectors $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are fed into three linear projections and reshaped to $k \times N \times m$, $k \times m \times n$ and $k \times n \times m$, respectively. Here, k denotes the number of heads, m denotes the head dimension, $k \times m = D$.

(4) A matrix multiplication operation is applied on $\mathbf{Q}$ and $\mathbf{K}$ to generate an attention map with the shape of $k \times N \times n$.

(5) The attention map is further proceeded by a convolution layer, a Softmax activation function and an Instance Normalization [55] operation.

(6) A matrix multiplication operation is applied on the proceeded attention map and $\mathbf{V}$. Finally, a linear projection is utilized to generate the output vector. The formalization of EMSA can be referred to the Equation (5).

$$\mathrm{EMSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{LP}\left(\mathrm{IN}\left(\mathrm{Softmax}\left(\mathrm{Conv}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{m}}\right)\right)\right)\cdot\mathbf{V}\right) \tag{5}$$

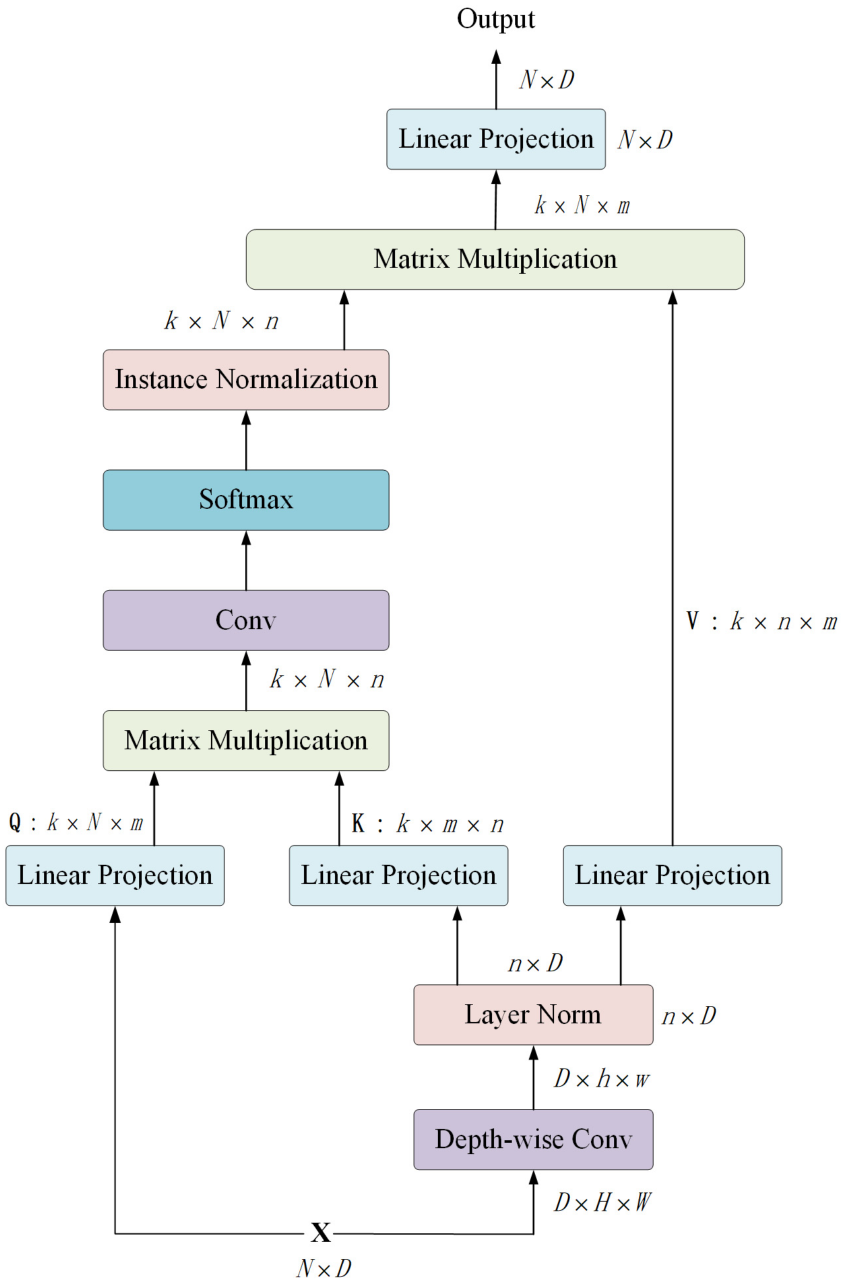

Figure 3. The flowchart of efficient multi-head self-attention.

Here, Conv is a standard 1 × 1 convolution with a stride of 1. IN denotes an instance normalization operation. LP represents a linear projection that keeps a dimension of D.

2.3. Texture Path

The texture path is a lightweight convolutional network, which builds four diverse convolutional layers to capture textural information. The output for the texture path can be formalized as

$$\mathrm{TF}(\mathbf{X}) = \mathrm{T}_4(\mathrm{T}_3(\mathrm{T}_2(\mathrm{T}_1\mathbf{X}))) \qquad (6)$$

Here, T represents a combined function consisting of a convolutional layer, a batch normalization operation, and a ReLU activation. The convolutional layer of T_1 has a kernel size of 7 and a stride of 2, which expands the channel dimension from 3 to 64. For T_2 and T_3, the kernel size and stride are 3 and 2, respectively. The channel dimension is kept as 64. For T_4, the convolutional layer is a standard 1 × 1 convolution with a stride of 1, expanding the channel dimension from 64 to 128. Thus, the output textural feature is downscaled 8 times and has a channel dimension of 128.

2.4. Feature Aggregation Module

The FAM aims to leverage the benefits of the dependent features and texture features comprehensively for powerful feature representation. As shown in Figure 4, the input features for the FAM include the **LDF3**, **LDF4** and **TF**. To fuse those features, we first employ an attentional embedding module (AEM) to merge the **LDF3** and **LDF4**. Thereafter, the merged feature is upsampled to concatenate with the **TF**, obtaining the aggregated feature. Finally, the linear attention module is deployed to reduce the fitting residual of the aggregated feature (**AF**). The pipeline of FAM can be denoted as

$$\mathrm{FAM}(\mathbf{AF}) = \mathbf{AF}\cdot\mathrm{LAM}(\mathbf{AF}) + \mathbf{AF} \tag{7}$$

$$\mathbf{AF}(\mathbf{TF},\ \mathbf{LDF3}, \mathbf{LDF4}) = \mathrm{C}(\mathrm{U}(\mathrm{AEM}(\mathbf{LDF3}, \mathbf{LDF4})), \mathbf{TF}) \tag{8}$$

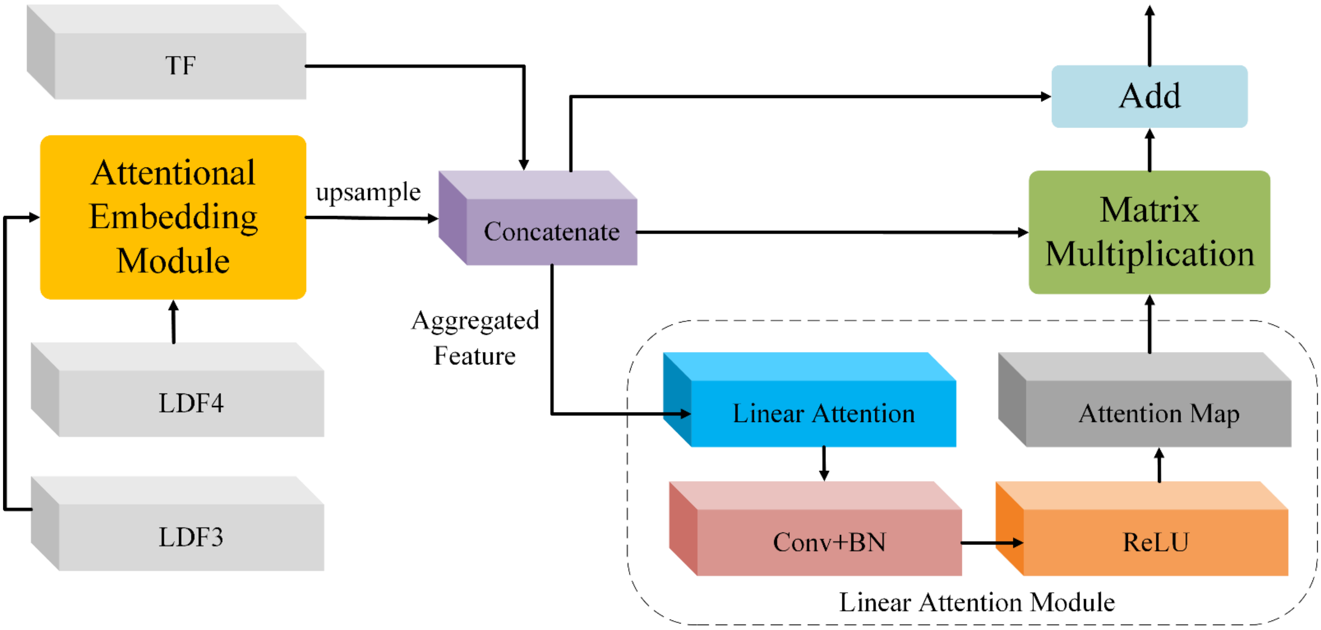

Figure 4. The feature aggregation module.

Here, C represents the concatenate function. U denotes an upsampling operation with a scale factor of 2. The details of LAM and AEM are as follows.

Linear attention module: The conventional dot-product attention mechanism can be defined as

$$D(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = \rho\left(\mathbf{Q}\mathbf{K}^{\mathrm{T}}\right)\mathbf{V}. \tag{9}$$

$$\rho\left(\mathbf{Q}\mathbf{K}^{\mathrm{T}}\right) = softmax_{row}\left(\mathbf{Q}\mathbf{K}^{\mathrm{T}}\right), \tag{10}$$

where query matrix **Q**, the key matrix **K**, and the value matrix **V** are generated by the corresponding standard 1 × 1 convolutional layer with a stride of 1, and $softmax_{row}$ indicates applying the softmax function along each row of matrix $\mathbf{QK}^{\mathrm{T}}$. The $\rho(\mathbf{QK}^{\mathrm{T}})$ models the similarities between each pair of pixels of the input, thoroughly extracting the global contextual information contained in the features. However, as $\mathbf{Q} \in \mathbb{R}^{N\times D_k}$. and $\mathbf{K}^{\mathrm{T}} \in \mathbb{R}^{D_k\times N}$, the product between **Q** and $\mathbf{K}^{\mathrm{T}}$ belongs to $\mathbb{R}^{N\times N}$, which leads to $O(N^2)$ memory and computation complexity. Therefore, the high resource-demanding of dot-product crucially limits the application on large inputs. Under the condition of softmax normalization function, the i-th row of result matrix generated by the dot-product attention module according to Equation (9) can be written as

$$D(\mathbf{Q},\ \mathbf{K},\ \mathbf{V})_i = \frac{\sum_{j=1}^{N} e^{\boldsymbol{q}_i{}^T \boldsymbol{k}_j} \boldsymbol{v}_j}{\sum_{j=1}^{N} e^{\boldsymbol{q}_i{}^T \boldsymbol{k}_j}}, \tag{11}$$

In our previous work on the linear attention (LA) mechanism [3], we design the LA based on first-order approximation of Taylor expansion on Equation (11):

$$e^{\boldsymbol{q}_i{}^T \boldsymbol{k}_j} \approx 1 + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2} \right)^T \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right). \tag{12}$$

where l_2 norm is utilized to ensure $\boldsymbol{q}_i{}^T \boldsymbol{k}_j \geq -1$. Then, Equation (11) can be rewritten as

$$D(\mathbf{Q},\ \mathbf{K},\ \mathbf{V})_i = \frac{\sum_{j=1}^{N} \left(1 + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2} \right)^T \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right) \right) \boldsymbol{v}_j}{\sum_{j=1}^{N} \left(1 + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2} \right)^T \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right) \right)}, \tag{13}$$

and be simplified as

$$D(\mathbf{Q},\ \mathbf{K},\ \mathbf{V})_i = \frac{\sum_{j=1}^{N} \boldsymbol{v}_j + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2} \right)^T \sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right) \boldsymbol{v}_j^T}{N + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2} \right)^T \sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right)}. \tag{14}$$

The above equation can be transformed in a vectorized form as

$$D(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = \frac{\sum_j \mathbf{V}_{i,j} + \left(\frac{\mathbf{Q}}{\|\mathbf{Q}\|_2} \right) \left(\left(\frac{\mathbf{K}}{\|\mathbf{K}\|_2} \right)^T \mathbf{V} \right)}{N + \left(\frac{\mathbf{Q}}{\|\mathbf{Q}\|_2} \right) \sum_j \left(\frac{\mathbf{K}}{\|\mathbf{K}\|_2} \right)_{i,j}^T}. \tag{15}$$

As $\sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right) \boldsymbol{v}_j^T$ and $\sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2} \right)$ can be calculated and reused for every query, time and memory complexity of the proposed LA based on Equation (15) is $O(N)$, while the illustration can be seen in Figure 5.

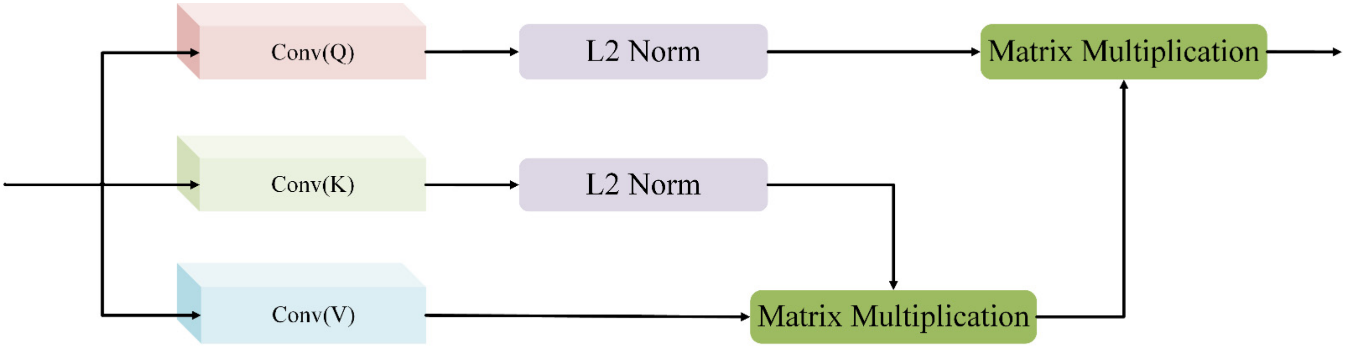

Figure 5. The linear attention.

In the FAM, we first employ LA to enhance the spatial relationships of AF, thereby suppressing the fitting residual. Then, a convolutional layer with BN and ReLU is deployed to obtain the attention map. Finally, we apply a matrix multiplication operation between AF and the attention map to obtain the attentional AF. The pipeline of LAM is defined as

$$\mathrm{LAM}(\mathbf{X}) = \mathrm{Conv}(\mathrm{BN}(\mathrm{ReLU}(\mathrm{LA}(\mathbf{X})))) \tag{16}$$

Here, Conv represents a standard convolution with a stride of 1.

Attentional embedding module: The AEM adopts the LAM to enhance the spatial relationships of **LDF4**. Then, we apply a matrix multiplication operation between the upsampling attention map of **LDF4** and **LDF3** to produce the attentional **LDF3**. Finally, we use an addition operation to fuse the original **LDF3** and the attentional **LDF3**. The pipeline of AEM is illustrated in Figure 6 and can be formalized as

$$\mathrm{AEM}(\mathbf{LDF3}, \mathbf{LDF4}) = \mathbf{LDF3} + \mathbf{LDF3} \cdot \mathrm{U}(\mathrm{LAM}(\mathbf{LDF4})) \tag{17}$$

where U denotes the nearest upsampling operation with a scale factor of 2.

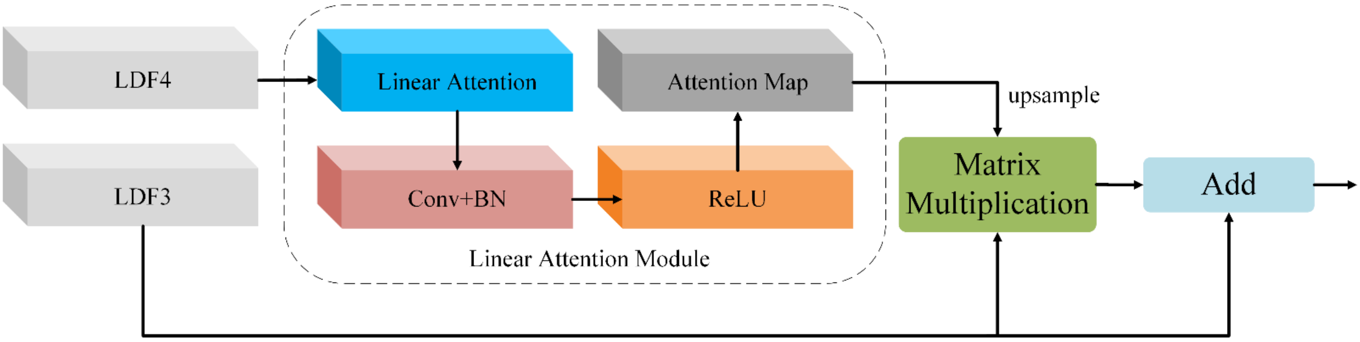

Figure 6. The attentional embedding module.

Capitalizing on the benefits provided by feature fusion, the final segmentation feature is abundant in both long-range dependency and textural information for precise semantic segmentation of urban scene images. In addition, linear attention reduces the fitting residual, strengthening the generalization of the network.

3. Experiments

In this section, experiments are conducted on three publicly available datasets to evaluate the effectiveness of the proposed BANet. We not only compare the performance of our model on the ISPRS Vaihingen and Potsdam datasets (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html, accessed on 20 October 2020) against the state-of-the-art models designed for remote sensing images but also take those proposed for natural images into consideration. Further, the UAVid dataset [56] is utilized to demonstrate the advantages of our method. Please note that as the backbone for the dependency path of our BANet is ResT-Lite with 10.49 M parameters, the backbone for comparative methods is selected as ResNet-18 with 11.7 M parameters correspondingly for a fair comparison.

3.1. Experiments on the ISPRS Vaihingen and Potsdam Datasets

3.1.1. Datasets

Vaihingen: There are 33 VFR images with a 2494 × 2064 average size in the Vaihingen dataset. The ground sampling distance (GSD) of tiles in Vaihingen is 9 cm. We utilize tiles: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38 for testing, tile: 30 for validation, and the remaining 15 images for training. Please note that we use only the near-infrared, red, and green channels in our experiments. The example images and labels can be seen in the top part of Figure 7.

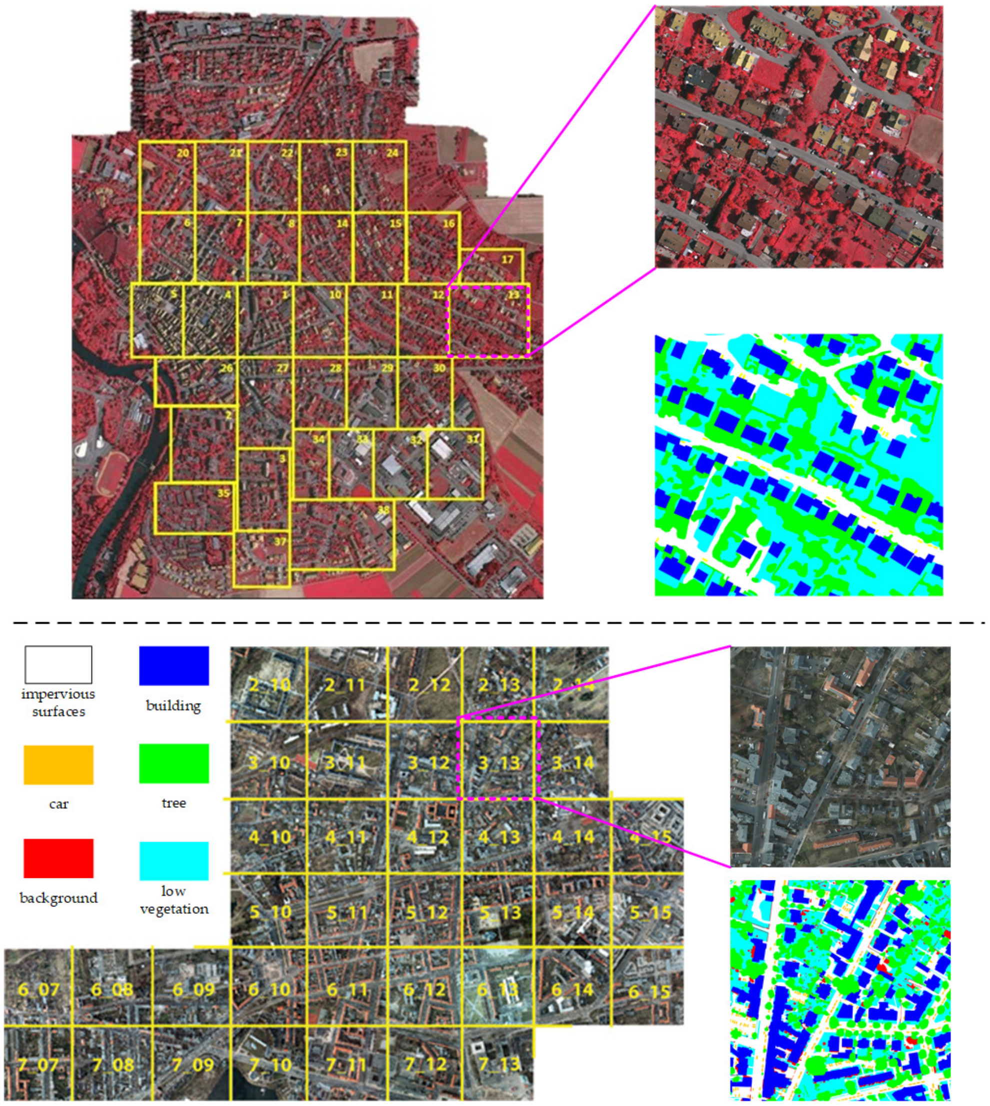

Figure 7. Example images and labels from the ISPRS Vaihingen dataset (**top** part) and Potsdam dataset (**bottom** part).

Potsdam: There are 38 fine-resolution images that cover urban scenes in the size of 6000 × 6000 pixels with a 5 cm GSD. We utilize ID: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13 for testing, ID: 2_10 for validation, and the remaining 22 images, except image named 7_10 with error annotations, for training. Only the red, green, and blue channels are used in our experiments. The example images and labels can be seen in the bottom part of Figure 7.

3.1.2. Training Setting

For optimizing the network, the Adam is set as the optimizer with the 0.0003 learning rate and 8 batch size. The images, as well as corresponding labels, are cropped into patches with 512 × 512 pixels and augmented by rotating, resizing, and flipping during training. All the experiments are implemented on a single NVIDIA RTX 3090 GPU with 24 GB RAM. The cross-entropy loss function is utilized as the loss function to measure the disparity between the achieved segmentation maps and the ground reference. If OA on the validation set does not increase for more than 10 epochs, the training procedure will be stopped, while the maximum iteration period is 100 epochs.

3.1.3. Evaluation Metrics

The performance of BANet on the ISPRS Potsdam dataset is evaluated using the overall accuracy (OA), the mean Intersection over Union (mIoU), and the F1 score (F1), which are computed on the accumulated confusion matrix:

$$\mathrm{OA} = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k}, \tag{18}$$

$$\mathrm{mIoU} = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{19}$$

$$\mathrm{F1} = 2 \times \frac{precision \times recall}{precision + recall}, \tag{20}$$

where TP_k, FP_k, TN_k, and FN_k indicate the true positive, false positive, true negative, and false negatives, respectively, for object indexed as class k. OA is calculated for all categories including the background.

3.1.4. Experimental Results

A detailed comparison between our BANet and other architectures including BiSeNet [57], FANet [58], MAResU-Net [3], EaNet [40], SwiftNet [59], and ShelfNet [60] can be seen in Tables 1 and 2, based upon the F1-score for each category, mean F1-score, and the OA, and the mIoU on the Vaihingen Potsdam test sets. As it can be observed from the table, the proposed BANet transcends the previous methods designed for segmentation by a large margin, achieving the highest OA of 90.48% and mIoU of 81.35% in the Vaihingen dataset, while the figures for the Potsdam dataset are 91.06% and 86.25%, respectively. Specifically, on the Vaihingen dataset, the proposed BANet brings more than 0.4% improvement in OA and 1.7% improvement in mIoU compared with the suboptimal method, while the improvements for the Potsdam dataset are more than 1.1% and 1.8%. Particularly, as the relatively small objects, the Car is difficult to recognize in the Vaihingen dataset. Even so, the proposed BANet achieves an 86.76% F1-score, preceding the suboptimal method by more than 5.5%.

Table 1. The experimental results on the Vaihingen dataset.

Method	Backbone	Imp. Surf.	Building	Low Veg.	Tree	Car	Mean F1	OA	mIoU
BiSeNet	ResNet-18	89.12	91.30	80.87	86.91	73.12	84.26	87.08	75.82
FANet	ResNet-18	90.65	93.78	82.60	88.56	71.60	85.44	88.87	75.61
MAResU-Net	ResNet-18	91.97	95.04	83.74	89.35	78.28	87.68	90.07	78.58
EaNet	ResNet-18	91.68	94.52	83.10	89.24	79.98	87.70	89.69	78.68
SwiftNet	ResNet-18	92.22	94.84	**84.14**	89.31	81.23	88.35	90.20	79.58
ShelfNet	ResNet-18	91.83	94.56	83.78	89.27	77.91	87.47	89.81	78.94
BANet	ResT-Lite	**92.23**	**95.23**	83.75	**89.92**	**86.76**	**89.58**	**90.48**	**81.35**

Table 2. The experimental results on the Potsdam dataset.

Method	Backbone	Imp. Surf.	Building	Low Veg.	Tree	Car	Mean F1	OA	mIoU
BiSeNet	ResNet-18	90.24	94.55	85.53	86.20	92.68	89.84	88.16	81.72
FANet	ResNet-18	91.99	96.10	86.05	87.83	94.53	91.30	89.82	84.16
MAResU-Net	ResNet-18	91.41	95.57	85.82	86.61	93.31	90.54	89.04	83.87
EaNet	ResNet-18	92.01	95.69	84.31	85.72	95.11	90.57	88.70	83.38
SwiftNet	ResNet-18	91.83	95.94	85.72	86.84	94.46	90.96	89.33	83.84
ShelfNet	ResNet-18	92.53	95.75	86.60	87.07	94.59	91.31	89.92	84.38
BANet	ResT-Lite	**93.34**	**96.66**	**87.37**	**89.12**	**95.99**	**92.50**	**91.06**	**86.25**

To qualitatively validate the effectiveness, we visualize the segmentation maps generated by our BANet and comparative methods in Figure 8. Due to the limited receptive field, the BiSeNet, EaNet, and SwiftNet assign the classification of a specific pixel only by considering a few adjacent areas, leading to fragmented maps and confusion of objects. The direct utilization of the attention mechanism (i.e., MAResU-Net) and the structure of multiple encoder-decoder (i.e., ShelfNet) brings certain improvements. However, the issue of the receptive field is still not entirely resolved. By contrast, we construct the dependency path in our BANet based on an attention-based backbone, i.e., ResT, to capture the long-range global relations, thereby tackling the limitation of the receptive field. Furthermore, a texture path built on convolution operation is equipped in our BANet to utilize the spatial details information in feature maps. Particularly, as shown in Figure 8, the complex circular contour of the Low vegetation is preserved completely by our BANet. In addition, the outlines of the Building generated by our BANet are smoother than those obtained by comparative methods.

3.2. Experiments on the UAVid Dataset

3.2.1. Dataset

As a fine-resolution Unmanned Aerial Vehicle (UAV) semantic segmentation dataset, the UAVid dataset (https://uavid.nl/, accessed on 10 May 2021) is focusing on urban street scenes with a 3840 × 2160 resolution. UAVid is a challenging benchmark since the large resolution of images, large-scale variation, and complexities in the scenes. To be specific, there are 420 images in the dataset where 200 are for training, 70 for validation, and the remaining 150 for testing. The example images and labels can be seen in Figure 9.

We adopt the same hyperparameters and data augmentation as those for experiments on ISPRS datasets, except batch size as 4 and the patch size as 1024 × 1024 during training.

3.2.2. Evaluation Metrics

For the UAVid dataset, the performance is assessed from the official server based on the intersection-over-union metric:

$$\mathrm{IoU} = \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{21}$$

where TP_k, FP_k, TN_k, and FN_k indicate the true positive, false positive, true negative, and false negatives, respectively, for object indexed as class k.

3.2.3. Experimental Results

Quantitative comparison with MSD [56], Fast-SCNN [61], BiSeNet, SwiftNet, and ShelfNet are reported in Table 3. As can be seen, the proposed BANet achieves the best IOU score on five out of eight classes and the best mIoU with a 3% gain over the suboptimal BiSeNet. Qualitative results on the UAVid validation set and test set are demonstrated in Figures 10 and 11, respectively. Compared with the benchmark MSD with obvious local and global inconsistencies, the proposed BANet can effectively capture the cues to scene semantics. For example, in the second row of Figure 11, the cars in the pink box are obviously all moving on the road. However, the MSD identity the left car which is crossing the street as the static car. In contrast, our BANet successfully recognizes all moving cars.

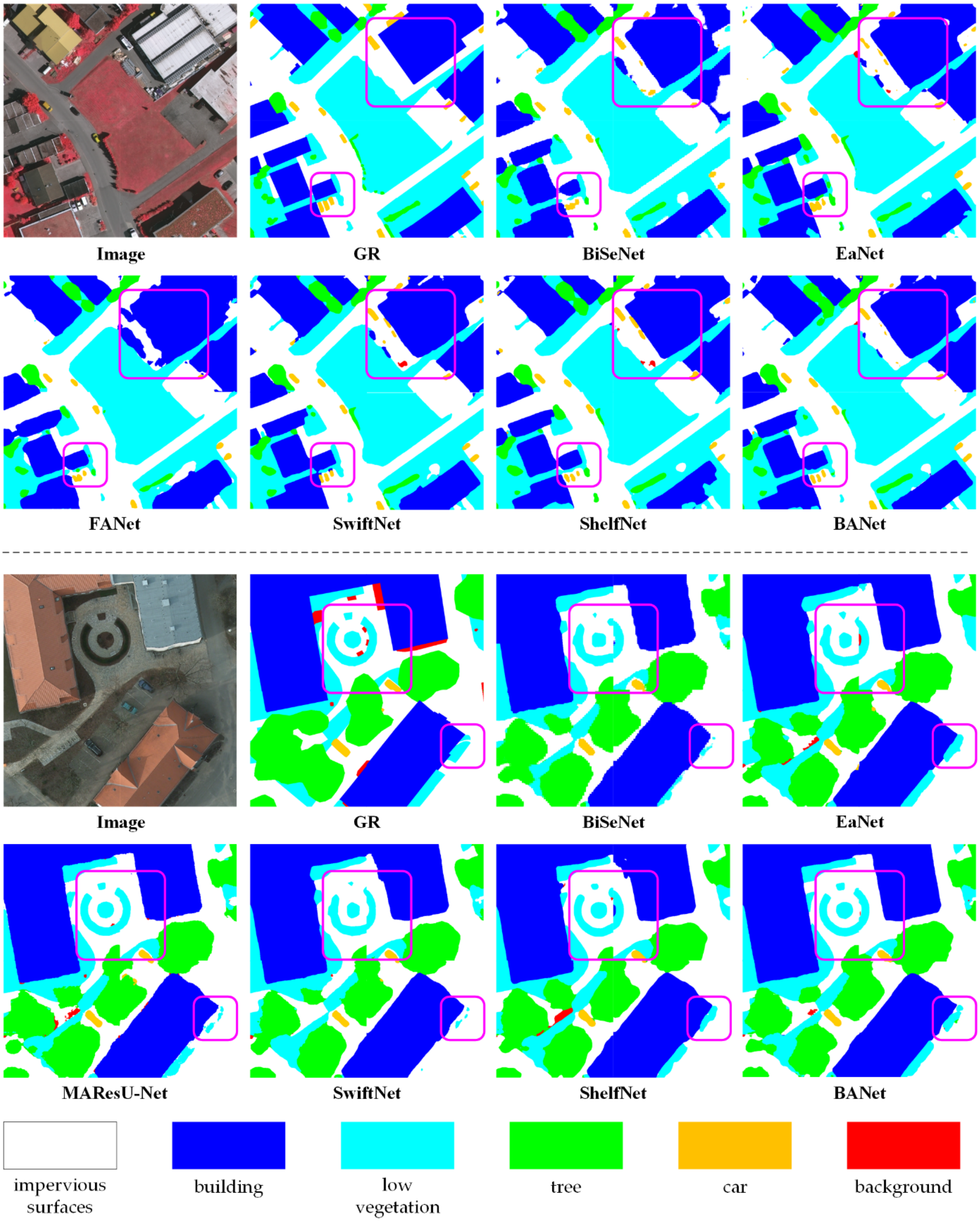

Figure 8. The experimental results on the ISPRS Vaihingen dataset (**top** part) and Potsdam dataset (**bottom** part). GR represents Ground Reference.

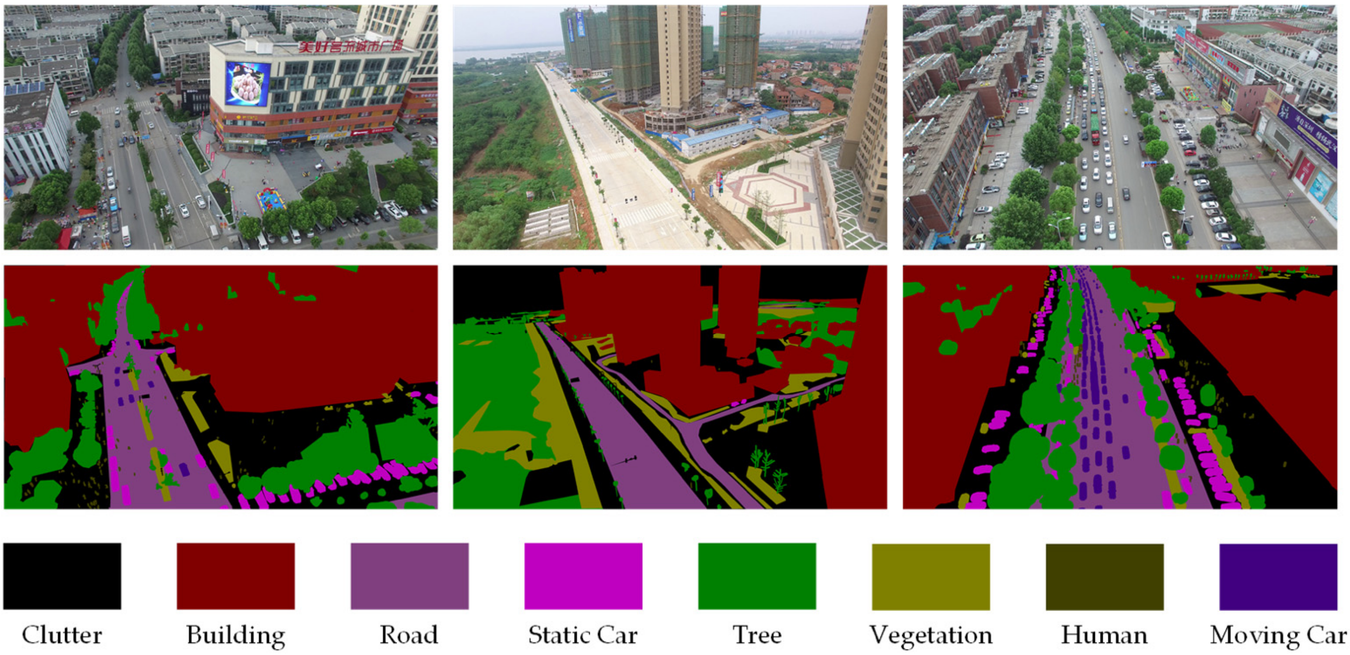

Figure 9. Example images and labels from the UAVid dataset.

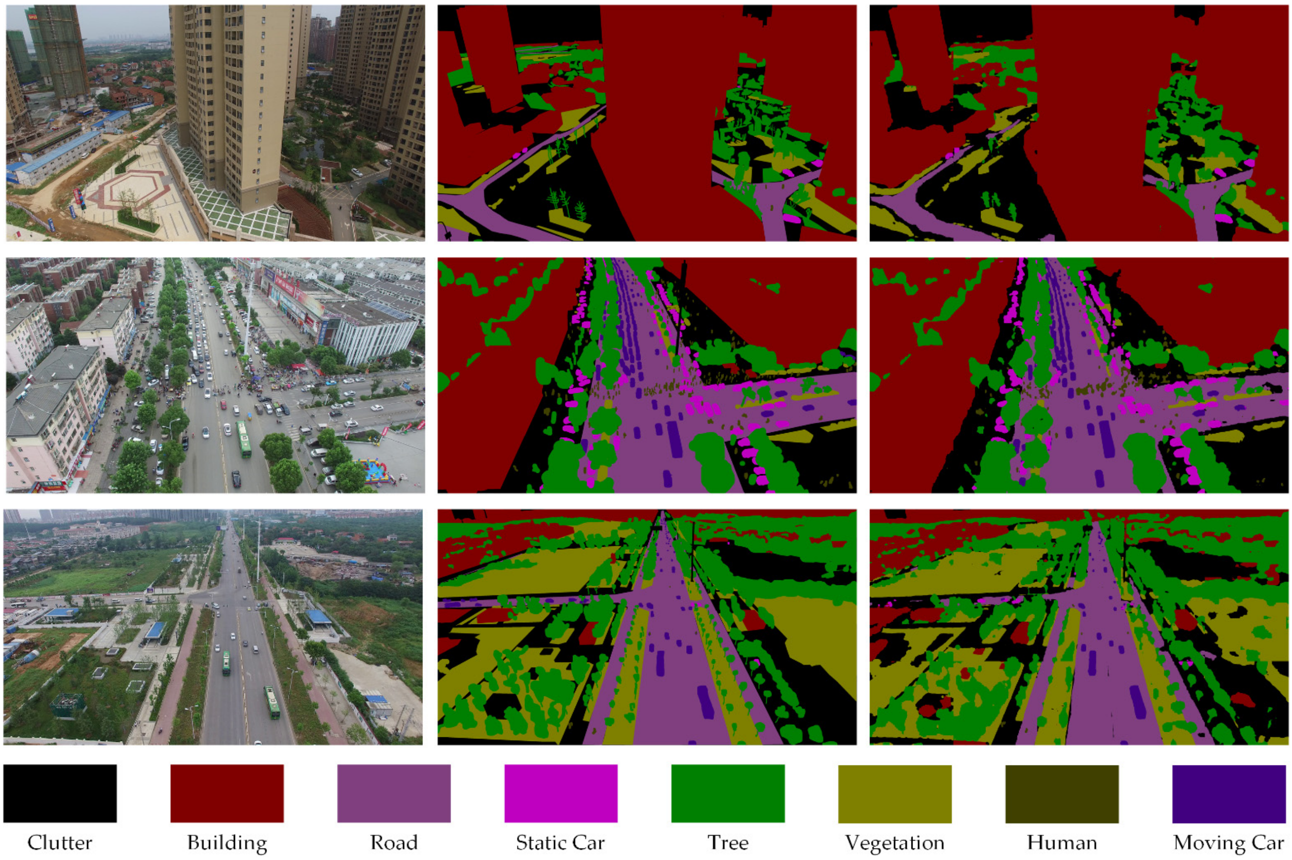

Figure 10. The experimental results on the UAVid validation set. The first column illustrates the input RGB images; the second column depicts the ground reference, and the third column shows the predictions of our BANet.

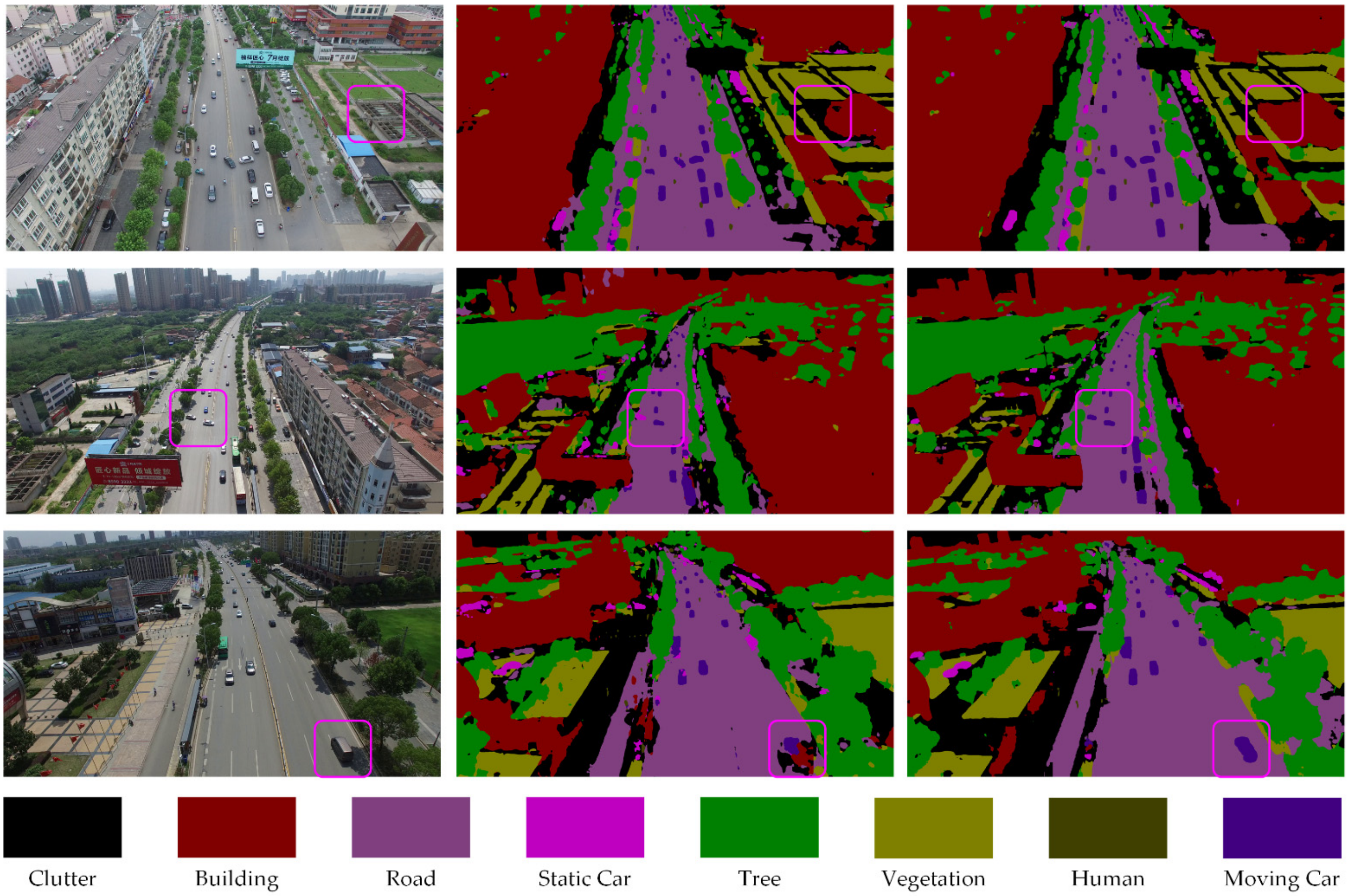

Figure 11. The experimental results on the UAVid test set. The first column illustrates the input RGB images; the second column depicts the outputs of MSD, and the third column shows the predictions of our BANet.

Table 3. The experimental results on the UAVid dataset.

Method	Building	Tree	Clutter	Road	Vegetation	Static Car	Moving Car	Human	mIoU
MSD	79.8	74.5	57.0	74.0	55.9	32.1	62.9	19.7	57.0
Fast-SCNN	75.7	71.5	44.2	61.6	43.4	19.5	51.6	0.0	45.9
BiSeNet	**85.7**	78.3	64.7	61.1	**77.3**	**63.4**	48.6	17.5	61.5
SwiftNet	85.3	78.2	64.1	61.5	76.4	62.1	51.1	15.7	61.1
ShelfNet	76.9	73.2	44.1	61.4	43.4	21.0	52.6	3.6	47.0
BANet	85.4	**78.9**	**66.6**	**80.7**	62.1	52.8	**69.3**	**21.0**	**64.6**

4. Discussion

4.1. Ablation Study

In this part, we conduct extensive ablation experiments on the ISPRS Potsdam dataset to verify the effectiveness of components in the proposed BANet, while the experimental settings and quantitative comparisons are illustrated in Table 4. The results are reported by the average value and corresponding deviation by three-fold experiments. Qualitative comparisons about the ablation study can be seen in Figure 12.

Table 4. The experimental results of the ablation study.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	Mean F1	OA	mIoU
ResNet	90.91 ± 0.45	95.18 ± 0.35	84.86 ± 0.92	86.44 ± 0.37	94.03 ± 0.63	90.28 ± 0.28	88.48 ± 0.50	82.34 ± 0.55
ResT	92.01 ± 0.58	95.73 ± 0.70	85.87 ± 0.58	87.24 ± 0.80	94.13 ± 0.49	91.00 ± 0.60	89.63 ± 0.49	83.80 ± 0.74
Dp+Tp(Sum)	92.11 ± 0.48	95.63 ± 0.43	86.5 ± 0.59	87.09 ± 0.97	94.44 ± 0.30	91.15 ± 0.53	89.87 ± 0.63	84.15 ± 0.72
Dp+Tp(Cat)	92.30 ± 0.55	95.99 ± 0.53	86.18 ± 0.64	87.57 ± 1.04	94.58 ± 0.77	91.32 ± 0.68	90.35 ± 0.23	85.31 ± 0.75
BAResNet	92.46 ± 0.21	95.37 ± 0.43	85.92 ± 0.84	87.24 ± 0.70	94.79 ± 0.28	91.16 ± 0.48	89.60 ± 0.54	84.07 ± 0.64
BANet	93.27 ± 0.11	96.53 ± 0.15	87.19 ± 0.23	88.63 ± 0.45	95.58 ± 0.44	92.24 ± 0.23	90.86 ± 0.18	85.78 ± 0.46

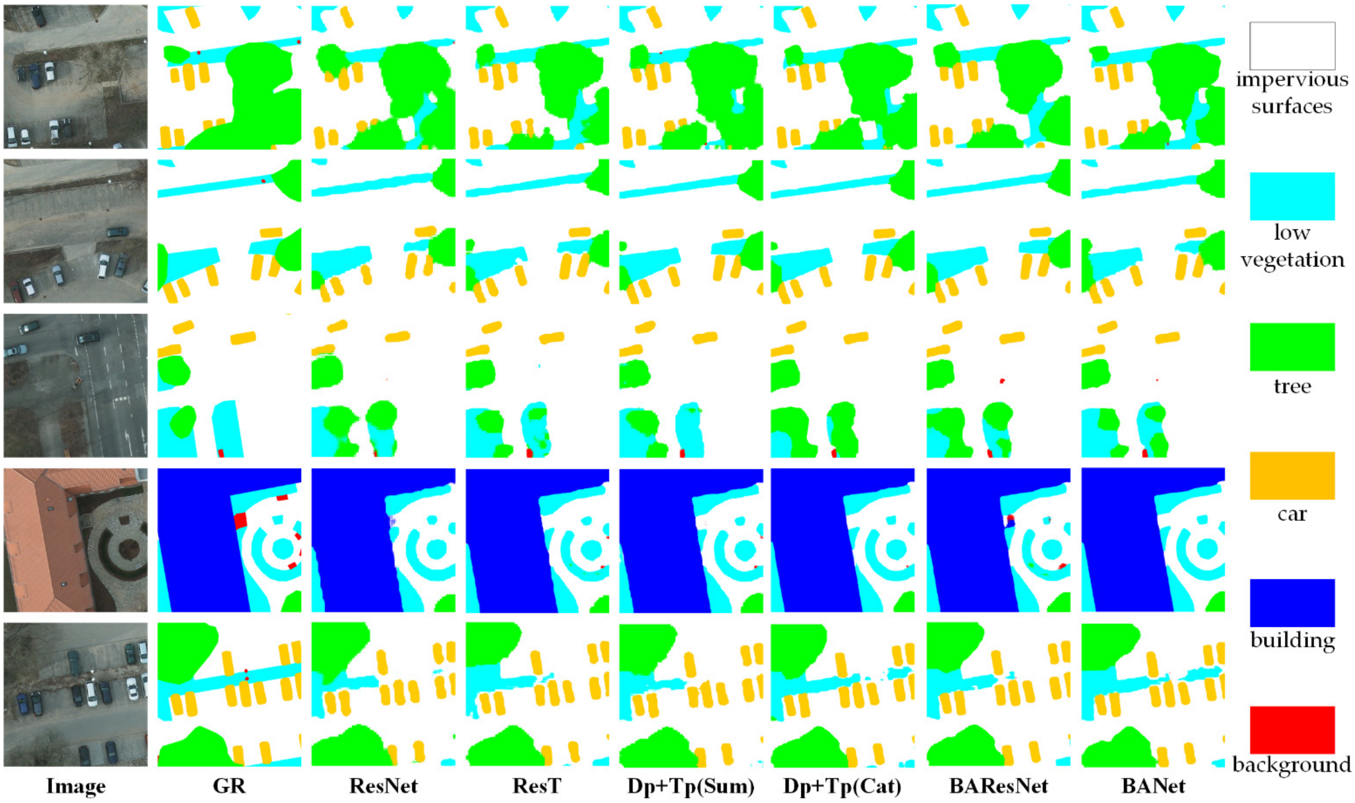

Figure 12. The ablation study on the ISPRS Potsdam dataset. GR represents Ground Reference.

Baseline: We select two baselines in ablation experiments, the dependency path which utilizes the ResNet-18 (denoted as ResNet) as the backbone and the dependency path which adopts the ResT-Lite (denoted as ResT) as the backbone. The feature maps generated by the dependency path are directly upsampled to restore the shape for final segmentation.

Ablation for the texture path: As rich spatial details are important for segmentation, the texture path conducted on the convolution operation is designed in our BANet for preserving the spatial texture. Table 4 illustrates that even the simple fusion schemes such as summation (indicated as Dp+Tp(Sum)) and concatenation (signified as Dp+Tp(Cat)) to merge the texture information can enhance the performance in OA at least 0.2%.

Ablation for feature aggregation module: Given the information obtained by the dependency path and the texture path are in different domains, neither summation nor concatenation is the optimal feature fusion scheme. As shown in Table 4, more than 0.5% improvement in OA brings by our BANet compared with Dp+Tp(Sum) and Dp+Tp(Cat) explains the validity of the proposed feature aggregation module.

Ablation for ResT-Lite: Since a novel transformer-based backbone, i.e., ResT, is introduced in our BANet, it is valuable to compare the accuracy between the ResNet and ResT. As illustrated in Table 4, the replacement of the backbone in the dependency path brings more than the 1% improvement in OA. In addition, we substitute the backbone in our BANet with ResNet-18 (denoted as BAResNet) to further evaluate the performance. As can be seen in Table 4, a 1.2% gap in OA illuminates the effectiveness of the ResT-Lite. Note that the number of parameters for BAResNet is 14.77 million (59.0 MB for weights file), while the figure for BANet is 15.44 million (56.4 MB for weights file). The inference speed of BAResNet is 73.2 FPS on a single mid-range GPU card, i.e., 1660Ti, for 512 × 512 input images, while the speed of BANet is 33.2 FPS, both satisfy the requirement of real-time ($\geq$30 FPS) scenarios. Please notice that the Nvidia GPU has the specialized optimization for CNN, while the optimization for Transformer is not available now. Therefore, the comparison is not completely fair now.

4.2. Application Scenarios

The main application scenario of our method is urban scene segmentation using remotely sensed images captured by satellite, aerial sensors, and UAV drones. The proposed Bilateral Awareness Network, which consists of a texture path, a dependency path, and a feature aggregation module, provides a unified framework for semantic segmentation, object detection, and change detection. Moreover, our model considers both accuracy and complexity, revealing enormous potential in illegal land use detection, real-time traffic monitoring, and urban environmental assessment.

In the future, we will continue to study the hybrid structure of convolution and Transformer and apply it to a wider range of urban applications

5. Conclusions

This paper proposes a Bilateral Awareness Network for semantic segmentation of very fine resolution urban scene images. Specifically, there are two branches in our BANet, a dependency path built on the Transformer backbone to capture the long-range relationships and a texture path constructed on the convolution operation to exploit the fine-grained details in VHR images. In particular, we further design an attentional feature aggregation module to fuse the global relationship information captured by the dependency path and the spatial texture information generated by the texture path. Extensive experiments on the ISPRS Vaihingen dataset, ISPRS Potsdam dataset, and UAVid dataset demonstrate the effectiveness of the proposed BANet. As a novel exploration to combine the Transformer and convolution in a bilateral structure, we envisage this pioneering paper could inspire practitioners and researchers engaged in this area to explore more possibilities of the Transformer in the remote sensing domain.

Author Contributions: This work was conducted in collaboration with all authors. D.W. and T.W. defined the research theme. X.M. supervised the research work and provided experimental facilities. L.W. and R.L. designed the semantic segmentation model and conducted the experiments. C.D. checked the experimental results. This manuscript was written by L.W. and R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) under grant number 41971352, National Key Research and Development Program of China under grant number 2018YFB0505003.

Data Availability Statement: We are grateful to ISPRS for providing the open benchmarks for 2D remote sensing image semantic segmentation. The data in the paper can be obtained through the following link. Potsdam: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/, accessed on 20 October 2020 Vaihingen: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/, accessed on 20 October 2020 and UAVid: https://uavid.nl/, accessed on 10 May 2021 Code is available at https://github.com/lironui/BANet, accessed on 25 June 2021.

Acknowledgments: The authors are very grateful to the many people who helped to comment on the article, and the Large Scale Environment Remote Sensing Platform (Facility No. 16000009, 16000011, 16000012) provided by Wuhan University, and the supports provided by Surveying and Mapping Institute Lands and Resource Department of Guangdong Province, Guangzhou. Special thanks to editors and reviewers for providing valuable insight into this article.

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

Abbreviations

The following abbreviations are used in this manuscript:

VFR	Very Fine Resolution
DCNNs	Deep Convolutional Neural Networks
FCN	Fully Convolutional Neural Network
SVM	Support Vector Machine
RF	Random Forest
CRF	Conditional Random Field
MHSA	Multi-Head Self-Attention
MLP	Multilayer Perceptron
FAM	Feature Aggregation Module
BANet	Bilateral Awareness Network
TF	Textural Features
AF	Aggregated Feature
LDF	Long-range Dependent Features
BN	Batch Normalization
GSD	Ground Sampling Distance
UAV	Unmanned Aerial Vehicle
LA	Linear Attention
AEM	Attentional Embedding Module
EMSA	Efficient Multi-head Self-attention

References

1. Zhang, C.; Atkinson, P.M.; George, C.; Wen, Z.; Diazgranados, M.; Gerard, F. Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 280–291. [CrossRef]
2. Zhang, C.; Harrison, P.A.; Pan, X.; Li, H.; Sargent, I.; Atkinson, P.M. Scale sequence joint deep learning (SS-JDL) for land use and land cover classification. *Remote Sens. Environ.* **2020**, *237*, 111593. [CrossRef]
3. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for Semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**. [CrossRef]
4. Li, R.; Duan, C.; Zheng, S.; Zhang, C.; Atkinson, P.M. MACU-Net for semantic segmentation of fine-resolution remotely sensed images. *IEEE Geosci. Remote Sens. Lett.* **2021**. [CrossRef]
5. Wang, L.; Fang, S.; Zhang, C.; Li, R.; Duan, C.; Meng, X.; Atkinson, P.M. SaNet: Scale-aware neural network for semantic labelling of multiple spatial resolution aerial images. *arXiv* **2021**, arXiv:2103.07935.
6. Huang, Z.; Wei, Y.; Wang, X.; Shi, H.; Liu, W.; Huang, T.S. AlignSeg: Feature-Aligned segmentation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef] [PubMed]
7. Yao, H.; Qin, R.; Chen, X. Unmanned aerial vehicle for remote sensing applications—A review. *Remote Sens.* **2019**, *11*, 1443. [CrossRef]
8. Audebert, N.; Le Saux, B.; Lefèvre, S. Segment-before-Detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sens.* **2017**, *9*, 368. [CrossRef]
9. Matikainen, L.; Karila, K. Segment-based land cover mapping of a suburban area—Comparison of high-resolution remotely sensed datasets using classification trees and test field points. *Remote Sens.* **2011**, *3*, 1777–1804. [CrossRef]
10. Zhang, Q.; Seto, K.C. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sens. Environ.* **2011**, *115*, 2320–2329. [CrossRef]
11. Wei, Y.; Wang, Z.; Xu, M. Road structure refined CNN for road extraction in aerial image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [CrossRef]
12. Li, E.; Femiani, J.; Xu, S.; Zhang, X.; Wonka, P. Robust rooftop extraction from visible band images using higher order CRF. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4483–4495. [CrossRef]
13. Zhang, Y.; Wang, C.; Ji, Y.; Chen, J.; Deng, Y.; Chen, J.; Jie, Y. Combining segmentation network and nonsubsampled contourlet transform for automatic marine raft aquaculture area extraction from sentinel-1 images. *Remote Sens.* **2020**, *12*, 4182. [CrossRef]
14. Maxwell, A.E.; Bester, M.S.; Guillen, L.A.; Ramezan, C.A.; Carpinello, D.J.; Fan, Y.; Hartley, F.M.; Maynard, S.M.; Pyron, J.L. Semantic segmentation deep learning for extracting surface mine extents from historic topographic maps. *Remote Sens.* **2020**, *12*, 4145. [CrossRef]
15. Kalajdjieski, J.; Zdravevski, E.; Corizzo, R.; Lameski, P.; Kalajdziski, S.; Pires, I.M.; Garcia, N.M.; Trajkovik, V. Air pollution prediction with multi-modal data and deep neural networks. *Remote Sens.* **2020**, *12*, 4142. [CrossRef]
16. 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]

17. Li, R.; Duan, C. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remote sensing images. *arXiv* **2021**, arXiv:2102.02531.
18. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. Joint deep learning for land cover and land use classification. *Remote Sens. Environ.* **2019**, *221*, 173–187. [CrossRef]
19. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. [CrossRef]
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585.
22. Guo, Y.; Jia, X.; Paull, D. Effective sequential classifier training for SVM-based multitemporal remote sensing image classification. *IEEE Trans. Image Process.* **2018**, *27*, 3036–3048. [CrossRef]
23. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
24. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 109–117.
25. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
26. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 96–107. [CrossRef]
27. Yue, K.; Yang, L.; Li, R.; Hu, W.; Zhang, F.; Li, W. TreeUNet: Adaptive Tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS J. Photogramm. Remote Sens.* **2019**, *156*, 1–13. [CrossRef]
28. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241.
29. 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]
30. Yang, M.Y.; Kumaar, S.; Lyu, Y.; Nex, F. Real-time semantic segmentation with context aggregation network. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 124–134. [CrossRef]
31. 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, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
32. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**. [CrossRef]
33. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 405–420.
34. Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
35. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. High-resolution aerial image labeling with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7092–7103. [CrossRef]
36. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
37. Duan, C.; Pan, J.; Li, R. Thick cloud removal of remote sensing images using temporal smoothness and sparsity regularized tensor optimization. *Remote Sens.* **2020**, *12*, 3446. [CrossRef]
38. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1758–1768. [CrossRef]
39. 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]
40. Zheng, X.; Huan, L.; Xia, G.-S.; Gong, J. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 15–28. [CrossRef]
41. 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, 18–22 June 2018; pp. 7794–7803.
42. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
43. Liu, Q.; Kampffmeyer, M.; Jenssen, R.; Salberg, A.B. Dense Dilated Convolutions' Merging Network for Land Cover Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6309–6320. [CrossRef]
44. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T.S. CCNet: Criss-cross attention for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.

46. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, preprint. arXiv:2010.11929.
47. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
48. Wang, L.; Li, R.; Duan, C.; Fang, S. Transformer meets DCFAM: A novel semantic segmentation scheme for fine-resolution remote sensing images. *arXiv* **2021**, arXiv:2104.12137.
49. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
50. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030.
51. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
52. Nair, V.; Hinton, G.E. Rectified linear units improve Restricted Boltzmann machines. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 807–814.
53. Zhang, Q.; Yang, Y. ResT: An efficient transformer for visual recognition. *arXiv* **2021**, arXiv:2105.13677.
54. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
55. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022v3.
56. Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
57. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
58. Hu, P.; Perazzi, F.; Heilbron, F.C.; Wang, O.; Lin, Z.; Saenko, K.; Sclaroff, S. Real-time semantic segmentation with fast attention. *IEEE Robot. Autom. Lett.* **2021**, *6*, 263–270. [CrossRef]
59. Oršić, M.; Šegvić, S. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognit.* **2021**, *110*, 107611. [CrossRef]
60. Zhuang, J.; Yang, J.; Gu, L.; Dvornek, N. Shelfnet for fast semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 847–856.
61. Poudel, R.P.K.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.