






# A<sup>2</sup>-FPN for semantic segmentation of fine-resolution remotely sensed images

Rui Li <sup>a</sup>, Libo Wang <sup>a</sup>, Ce Zhang <sup>b,c</sup>, Chenxi Duan <sup>d</sup> and Shunyi Zheng <sup>a</sup>

<sup>a</sup>School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; <sup>b</sup>Lancaster Environment Centre, Lancaster University, Lancaster, UK; <sup>c</sup>UK Centre for Ecology & Hydrology, Lancaster, UK; <sup>d</sup>Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands

## ABSTRACT

The thriving development of earth observation technology makes more and more high-resolution remote-sensing images easy to obtain. However, caused by fine-resolution, the huge spatial and spectral complexity leads to the automation of semantic segmentation becoming a challenging task. Addressing such an issue represents an exciting research field, which paves the way for scene-level landscape pattern analysis and decision-making. To tackle this problem, we propose an approach for automatic land segmentation based on the Feature Pyramid Network (FPN). As a classic architecture, FPN can build a feature pyramid with high-level semantics throughout. However, intrinsic defects in feature extraction and fusion hinder FPN from further aggregating more discriminative features. Hence, we propose an Attention Aggregation Module (AAM) to enhance multiscale feature learning through attention-guided feature aggregation. Based on FPN and AAM, a novel framework named Attention Aggregation Feature Pyramid Network (A<sup>2</sup>-FPN) is developed for semantic segmentation of fine-resolution remotely sensed images. Extensive experiments conducted on four datasets demonstrate the effectiveness of our A<sup>2</sup>-FPN in segmentation accuracy. Code is available at <https://github.com/lironui/A2-FPN>.

## ARTICLE HISTORY



Received 18 October 2021  
Accepted 9 January 2022

## KEYWORDS

semantic segmentation;  
deep learning; attention  
mechanism

## 1. Introduction

Land-cover information can provide insights from a panoramic perspective to help tackle urgent socioeconomic and environmental challenges, such as food crisis, climate change, and disaster risks. Hence, semantic segmentation, which can assign definite categories to groups of pixels in an image, has become one of the most significant techniques for ground feature interpretation (Li et al. 2021d, 2022). For remotely sensed images, segmentation has played critical roles in several diverse geo-information applications, including urban planning, economic assessment, land resource management, etc. (Zhang et al. 2019; Tong et al. 2020; Zhu et al. 2017). Derived from blooming advances in Earth observation technology, a series of satellite and airborne platforms have been launched (Duan, Pan and Li 2020; Zhang et al. 2020b), thereby making substantial remotely sensed images available. For segmentation,

**CONTACT** Chenxi Duan  [c.duan@utwente.nl](mailto:c.duan@utwente.nl)  Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Hengelosestraat 99 7514 AE Enschede, the Netherlands

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.  
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

traditional methods usually extract vegetation indices of land cover from multispectral/multi-temporal images to manifest the physical properties. However, as the descriptors are hand-crafted, the adaptability and flexibility of these indices are severely limited (Li et al. 2020b; Xiaowei et al. 2020).

Meanwhile, substantial significant leaps of segmentation in remote sensing have been witnessed in recent years (Wang et al. 2021a, 2021b), due to the extensive applications of deep learning and deep convolutional neural networks (CNNs) in particular. Compared with vegetation indices, a wide range of features can be fully extracted by CNNs, such as context information, spectral characteristics, and the mutual effect between different land-cover categories (Wambugu et al. 2021; Bai et al. 2021). Further, benefiting from the powerful ability to capture nonlinear and hierarchical features automatically, CNNs can form the end-to-end framework from the raw image to meaningful information and insights directly (Wang et al. 2021a).

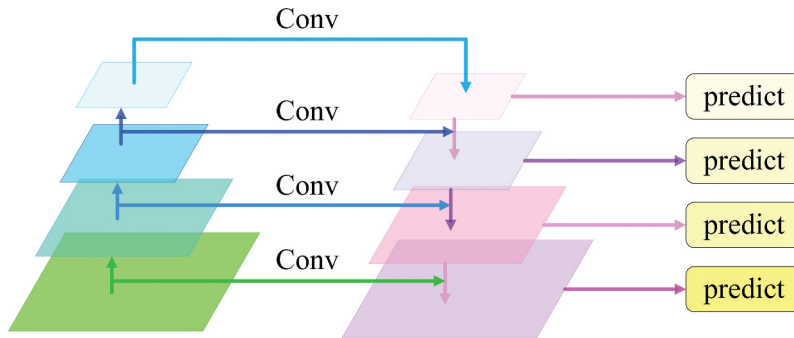
For remote-sensing imagery, the scale variation of geospatial objects is a general phenomenon, which is especially true for those with fine-resolution. Therefore, how to extract the multiscale representation is important for dealing with such an issue. As a widely used framework, Feature Pyramid Network (FPN) (Lin et al. 2017) is a feasible scheme to address the problem of multiscale processing. Specifically, by fusing adjacent features through lateral connections and the top-down pathway, FPN constructs a feature pyramid with abundant semantics at all scales, thereby exploiting the inherent feature hierarchy.

Although effective in multiscale feature representations, the designs of FPN hinder feature pyramids from further aggregating more discriminative features for segmentation. Specifically, in the procedure of feature fusion, feature maps are up-sampled and fused directly, losing the rich context information. Fortunately, dot-product attention mechanisms show strong capabilities to capture long-range dependencies. Different from scaling attention mechanisms which are designed to reinforce informative features and whittle information-lacking features, dot-product attention mechanisms can extract the contextual information by measuring the relationships of every pixel-pair of the input (Li et al. 2021c). However, the memory and computational consumptions of the dot-product attention mechanism increase quadratically with an increase in the spatio-temporal size of the input, which hugely limits their practicability (Li et al. 2021d). Therefore, in this paper, we introduce the linear attention mechanism (Li et al. 2021b), i.e. a simplified dot-product attention mechanism, to the FPN and propose an Attention Aggregation Module (AAM) to enhance multiscale feature learning, thereby designing  $A^2$ -FPN. Compared to mainstream encoder-decoder frameworks,  $A^2$ -FPN is distinctive in two significant aspects: (1) It encodes semantic features from multiscale layers; (2) it extracts discriminative features by extracting global context information.

## 2. Related work

### 2.1. Feature pyramid network

The feature pyramid network is initially designed for object detection, aiming at leveraging the pyramidal feature hierarchy (Lin et al. 2017). The components of the FPN are comprised of a bottom-up pathway, a top-down pathway, and lateral connections, as illustrated in Figure 1. The bottom-up pathway usually takes the ResNet as the backbone (He et al. 2016), where the feature hierarchy is computed with feature maps being generated at multiple scales. The



**Figure 1.** Illustration of the architecture of feature pyramid network for detection.

feature maps at top pyramid levels are spatially coarse but with high-level semantics. The top-down pathway interpolates fine-resolution features by up-sampling from high-level feature maps, which are then merged and refined with features at the same spatial size from the bottom-up pathway via lateral connections. The effectiveness of FPN has been demonstrated in several applications, including object detection (Lin et al. 2017), panoptic segmentation (Kirillov et al. 2019), and super-resolution (Shoebiet al. 2020).

## 2.2. Semantic segmentation

After the first successful Fully Convolutional Network (FCN), deep learning methods have been successfully and extensively introduced and applied to the semantic segmentation, while the remote-sensing area is no exception (Wang et al. 2021a, 2021b). For example, Sherrah (2016) adapted the FCN to semantically label remotely sensed images. Kampffmeyer, Salberg and Jenssen (2016) focused on the segmentation of relatively small objects (e.g. Cars) by quantifying the uncertainty at the pixel level. To investigate the impact of the intermediate features fusion scheme, Maggiori et al. (2017) adopted an auxiliary CNN to learn how to combine features. Audebert, Le Saux and Lefèvre (2018) further leveraged multi-modal data by the V-FuseNet to enhance the segmentation accuracy. However, such a fusion scheme will be invalid if either modality is unavailable in the test phase. Kampffmeyer, Salberg and Jenssen (2018), therefore, proposed a hallucination network aiming to replace missing modalities during testing. Besides, enhancing the segmentation accuracy by optimizing object boundaries is another burgeoning research area (Zheng et al. 2020; Marmanis et al. 2018). Meanwhile, semantic segmentation has shown great potential for practical applications in remote-sensing areas including road detection (Wei, Zhang and Ji 2020; Shamsolmoali et al. 2020), urban resource management (Zhang et al. 2020a; Li et al. 2020a), and land-use mapping (Tu et al. 2020). For example, a novel CNN-based multi-stage framework was introduced by Wei, Zhang and Ji (2020) to extract road surface and center-line tracing simultaneously. Zhang et al. (2020a) characterizes and classifies individual plants based on semantic segmentation methods by continuously increasing patch scale. The recently developed semantic segmentation approaches using deep learning create a new paradigm for land-use mapping (Tu et al. 2020).

### 2.3. The attention mechanism

The accuracy of segmentation relies on inference from sufficient context information. To this end, the dot-product attention mechanism is introduced to capture the global context. However, the memory and computational consumptions which increase quadratically with the input size heavily impede the actual application of the dot-product attention mechanism. Here, we illustrate the principles of the dot-product attention mechanism as well as the attempts to reduce the complexity of the attention mechanism, especially the linear attention mechanism utilized in the proposed A<sup>2</sup>-FPN. By default, vectors in this section refer to column vectors.

#### 2.3.1. The dot-product attention mechanism

The height, weight, and channels of the input are denoted as  $H$ ,  $W$  and  $C$ , respectively.  $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{N \times C}$  refers to the input feature, where  $N = H \times W$ . First, the dot-product attention mechanism uses three projected matrices  $W_q \in \mathbb{R}^{\mathbb{D}_q \times \mathbb{D}_1}$ ,  $W_k \in \mathbb{R}^{\mathbb{D}_k \times \mathbb{D}_1}$ , and  $W_v \in \mathbb{R}^{\mathbb{D}_v \times \mathbb{D}_c}$  to obtain the *query* matrix  $Q$ , *key* matrix  $K$  and *value* matrix  $V$  as

$$\begin{aligned} Q &= XW_q \in \mathbb{R}^{N \times \mathbb{D}_q}, \\ K &= XW_k \in \mathbb{R}^{N \times \mathbb{D}_k}, \\ V &= XW_v \in \mathbb{R}^{N \times \mathbb{D}_v}. \end{aligned} \quad (1)$$

$Q$  and  $K$  are identical in their shapes. To compute the similarity between the  $i$ -th *query* feature  $q_i^T \in \mathbb{R}^{\mathbb{D}_q}$  and the  $j$ -th *key* feature  $k_j \in \mathbb{R}^{\mathbb{D}_k}$ , a normalization function  $\rho$  is adopted as  $\rho(q_i^T \cdot k_j) \in \mathbb{R}^{\mathbb{K}}$ . Thereafter, similarities between all pairs of pixels are computed and taken as weights. The output is generated by aggregating all positions using weighted summation:

$$D(Q, K, V) = \rho(QK^T)V. \quad (2)$$

For dot-product attention mechanism, the normalization function is set as softmax:

$$\rho(QK^T) = \text{softmax}_{\text{row}}(QK^T). \quad (3)$$

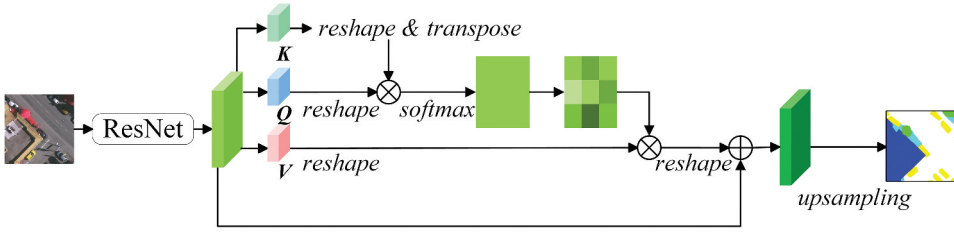
where  $\text{softmax}_{\text{row}}$  denotes that the softmax is operated along the row of matrix  $QK^T$ . The global context information is captured by the  $\rho(QK^T)$  through the modelling of the similarities among all pairs of pixels in the input. However, as  $Q \in \mathbb{R}^{N \times \mathbb{D}_q}$  and  $K^T \in \mathbb{R}^{\mathbb{D}_k \times N}$ , the multiplication between  $Q$  and  $K^T$  belongs to  $\mathbb{R}^{N \times N}$ , leading to the  $O(N^2)$  time and memory complexity (Figure 2).

#### 2.3.2. Generalization and simplification

Given the normalization function is softmax, the  $i$ th row in the output matrix produced by the dot-product attention mechanism can be written as:

$$D(Q, K, V)_i = \frac{\sum_{j=1}^N e^{q_i^T \cdot k_j} v_j}{\sum_{j=1}^N e^{q_i^T \cdot k_j}}. \quad (4)$$

Equation (4) can be generalized into any normalization function as



**Figure 2.** illustration of the architecture of dot-product attention mechanism.

$$D(Q, K, V)_i = \frac{\sum_{j=1}^N \text{sim}(q_i, k_j) v_j}{\sum_{j=1}^N \text{sim}(q_i, k_j)}, \text{sim}(q_i, k_j) \geq 0, \quad (5)$$

$\text{sim}(q_i, k_j)$  depicts the similarity between the  $q_i$  and  $k_j$ , which can be expanded as  $\text{sim}(q_i, k_j) = \phi(q_i)^T \varphi(k_j)$ . We can further rewrite Equation Equation (4) to (6) and then simplify it as Equation (7):

$$D(Q, K, V)_i = \frac{\sum_{j=1}^N \phi(q_i)^T \varphi(k_j) v_j}{\sum_{j=1}^N \phi(q_i)^T \varphi(k_j)}, \quad (6)$$

$$D(Q, K, V)_i = \frac{\phi(q_i)^T \sum_{j=1}^N \varphi(k_j) v_j}{\phi(q_i)^T \sum_{j=1}^N \varphi(k_j)} \quad (7)$$

In particular, Equation (5) is identical to Equation (4), when  $\text{sim}(q_i, k_j) = e^{q_i^T \cdot k_j}$ . Equation (7) can be represented as the vectorized form:

$$D(Q, K, V) = \frac{\phi(Q) \varphi(K)^T V}{\phi(Q) \sum_j \varphi(K)_{ij}^T}, \quad (8)$$

As  $\text{sim}(q_i, k_j) = \phi(q_i)^T \varphi(k_j)$  replaces the softmax function, the order of the commutative operation can be altered, thereby reducing the computationally intensive operations. Specifically, we can compute the multiplication between  $\varphi(K)^T$  and  $V$  first and then multiply the result and  $\phi(Q)$ , resulting in only  $O(dN)$  time and memory complexity. The appropriate  $\phi(\cdot)$  and  $\varphi(\cdot)$  and enable the drastically reduced computation without sacrificing the accuracy (Li et al. 2021c; Katharopoulos et al. 2020).

### 2.3.3. The linear attention mechanism

By replacing the softmax into its first-order approximation of Taylor expansion, we have developed a linear attention mechanism in our previous research (Li et al. 2021b) as

$$e^{\mathbf{q}_i^T \cdot \mathbf{k}_j} \approx 1 + \mathbf{q}_i^T \cdot \mathbf{k}_j, \tag{9}$$

However, the above approximation cannot guarantee the non-negative property of the normalization function. Hence, we normalize  $\mathbf{q}_i$  and  $\mathbf{k}_j$  by  $l_2$  norm to ensure  $\mathbf{q}_i^T \cdot \mathbf{k}_j \geq -1$ :

$$\text{sim}(\mathbf{q}_i, \mathbf{k}_j) = 1 + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right)^T \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right), \tag{10}$$

We then rewrite Equation (5) into Equation (11), and simplify it into Equation (12):

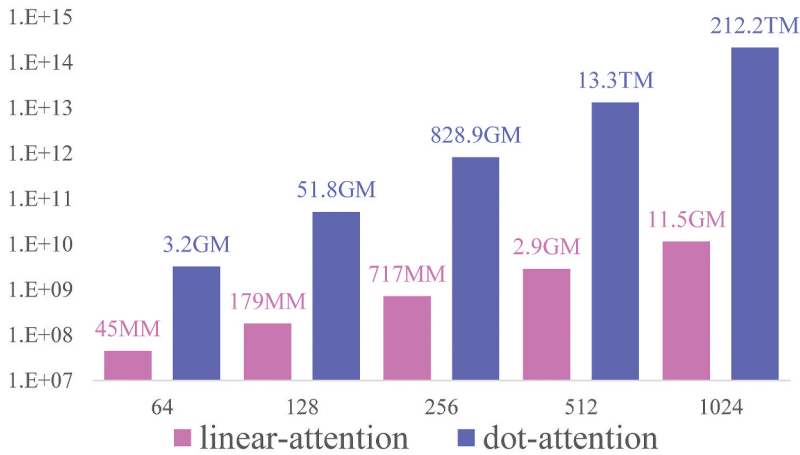
$$D(\mathbf{Q}, \mathbf{K}, \mathbf{V})_i = \frac{\sum_{j=1}^N \left( 1 + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right)^T \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right) \right) \mathbf{v}_j}{\sum_{j=1}^N \left( 1 + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right)^T \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right) \right)}, \tag{11}$$

$$D(\mathbf{Q}, \mathbf{K}, \mathbf{V})_i = \frac{\sum_{j=1}^N \mathbf{v}_j + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right)^T + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right)^T \sum_{j=1}^N \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right) \mathbf{v}_j^T}{N + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right)^T \sum_{j=1}^N \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right)}. \tag{12}$$

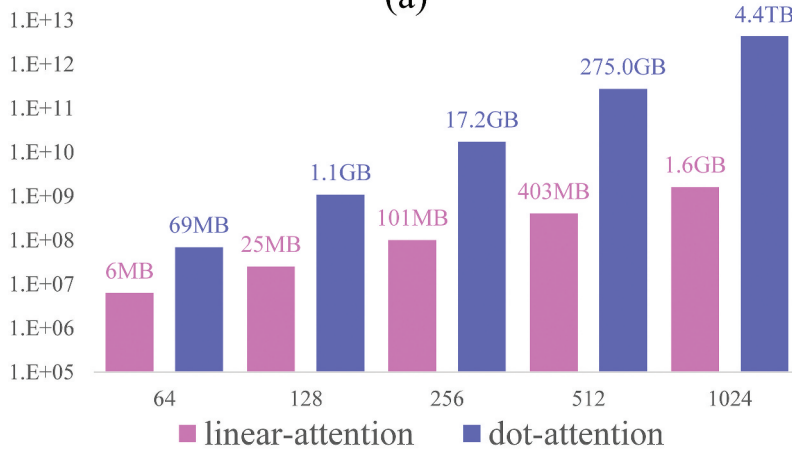
The vectorized form of Equation (12) is

$$D(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\sum_j \mathbf{v}_{ij} + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right) \left( \left( \frac{\mathbf{K}}{\|\mathbf{K}\|_2} \right)^T \mathbf{V} \right)}{N + \left( \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2} \right) \sum_j \left( \frac{\mathbf{K}}{\|\mathbf{K}\|_2} \right)^T_{ij}}. \tag{13}$$

As  $\sum_{j=1}^N \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right) \mathbf{v}_j^T$  and  $\sum_{j=1}^N \left( \frac{\mathbf{k}_j}{\|\mathbf{k}_j\|_2} \right)$  could be computed only once and reused for each query, time and space complexity of the linear attention mechanism based on Equation (13) is  $O(dN)$ . Specifically, given a feature  $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{N \times C}$ , both the dot-attention and linear attention generate the *query* matrix  $Q$ , *key* matrix  $K$  and *value* matrix  $V$ . For the dot-attention, the  $N \times N$  matrix is generated by multiplying the transposed *key* matrix  $K$  and the *value* matrix  $V$ , resulting in  $O(D_k N^2)$  time complexity and  $O(N^2)$  space complexity to compute the similarity using the softmax function. Thus, the dot-attention would occupy at least  $O(N^2)$  memory and require  $O(D_k N^2)$  computation to calculate the similarity between each pair of positions. For linear attention, as the softmax function is substituted for the first-order approximation of Taylor expansion, we can alter the order of the commutative operation and avoid multiplication between the reshaped *key* matrix  $K$  and *query* matrix  $Q$ . Therefore, we can calculate the product between  $K^T$  and  $V$  first and then multiply the result and  $Q$  with only  $O(dN)$  time complexity and  $O(dN)$  space complexity. The concrete comparison can be seen in [Figure 3](#).



(a)

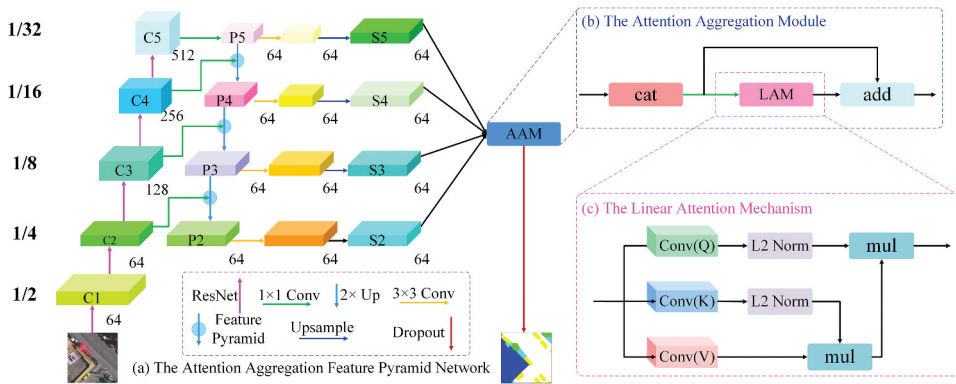


(b)

**Figure 3.** The (a) computation requirement and (b) memory requirement between the linear attention mechanism and dot-product attention mechanism under different input sizes. the calculation assumes  $D = D_v = 2D_k = 64$ . MM denotes 1 Mega multiply-accumulation (MACC), where 1 MACC means 1 multiplication and 1 addition operation. GM means 1 Giga MACC, while TM signifies 1 Tera MACC. Similarly, MB, GB, and TB represent 1 MegaByte, 1 GigaByte, and 1 TeraByte, respectively. Note the figure is shown on the log scale.

### 3. Attention aggregation feature pyramid network

The overall framework of the proposed  $A^2$ -FPN is demonstrated in [Figure 4](#). As a single end-to-end network, the major components of our  $A^2$ -FPN include the bottom-up pathway (i.e. the first column in [Figure 4](#)), the top-down pathway (i.e. the second column in [Figure 4](#)), the lateral connections (i.e. the  $1 \times 1$  convolutional layer between the first and second column in [Figure 4](#)), the feature pyramid (i.e. the second and third columns in [Figure 4](#)), and the Attention Aggregation Module (i.e. [Figure 4\(b\)](#)). We will elaborate on each component below.



**Figure 4.** The structure of (a) the overall framework of our A<sup>2</sup>-FPN, (b) the Attention Aggregation Module, and (c) the Linear Attention Mechanism (taking the attention1 as an example). the figures (e.g. 64, 128, 512) near the features indicate the number of channels.

### 3.1. The bottom-up pathway

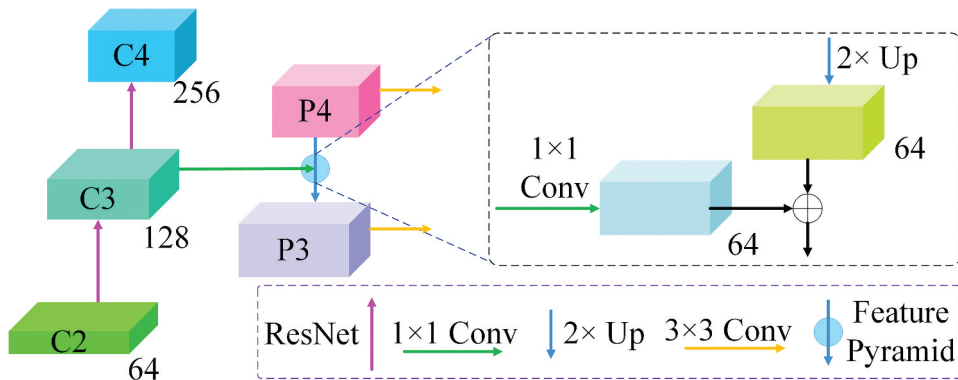
To design a simple and efficient framework, we select the ResNet-18 or ResNet-34 as the backbone of the bottom-up pathway rather than the complicated backbones such as ResNet-101. Based on ResNet backbone, the bottom-up pathway conducts the feed-forward learning and generates the feature hierarchy. The feature maps are generated at different spatial resolutions with a scaling step of 2. The top levels of feature maps have large spatial context with coarse resolution, whereas the bottom levels of feature maps present small context information with fine resolution. We use C2, C3, C4, and C5 to indicate the output feature map of each residual block in ResNets (see above Figure 4), while the spatial size of C2, C3, C4, and C5 are 1/4, 1/8, 1/16, and 1/32 of the input size, respectively. Due to its large memory footprint, C1 is not included in the pyramid.

### 3.2. The top-down pathway and lateral connections

The top-down pathway up-samples semantically rich but spatially coarse feature maps from top pyramid levels to create fine resolution features, which are then merged and refined with corresponding features from the bottom-up pathway via lateral connections. As shown in Figure 5, a top-down layer and a lateral connection constitute a feature pyramid in the proposed A<sup>2</sup>-FPN. The generated feature maps are denoted as P2, P3, P4, and P5 accordingly. With a coarse resolution feature map (e.g. P4 in Figure 5), we up-sample its spatial resolution by a factor of 2, while the up-sampling mode is set as the nearest neighbor for simplicity. By element-wise addition, the up-sampled map is then fused with the corresponding map in the bottom-up pathway, wherein a  $1 \times 1$  convolutional layer is utilized to reduce dimensions of the channel.

The above procedure is iterated until the finest resolution map is generated. To start the iteration, the coarsest resolution map (e.g. P5 in Figure 4) is directly produced by a  $1 \times 1$  convolutional layer on C5. After the merged map generated by the corresponding feature pyramid, a  $3 \times 3$  convolution is attached to produce the final feature map to mitigate the aliasing effect caused by up-sampling operation. The feature pyramid combines low-level contextual information into spatial feature maps, which improves the representation





**Figure 5.** The feature pyramid in the proposed  $A^2$ -FPN.

capability of low-level side networks. Interpreting different scales of land covers requires different levels of context information. Indeed, a large spatial context is contained in the high-level features since the deep convolution layers have larger receptive fields than the shallow ones. Hence, when merged with high-level features, the low-level side networks acquire the multiscale context information to improve its accuracy of segmentation.

### 3.3. The attention aggregation module

The local-aware property severely limits the potential of the CNN to capture the global context information, while the latter is paramount for semantic segmentation. Graphical models and pyramid pooling modules partly remedy the context issue. However, the contextual dependencies for whole input regions are homogeneous and non-adaptive, ignoring the disparity between contextual dependencies and local representation of different categories. Besides, those strategies usually utilized only in one layer do not sufficiently leverage the long-range dependencies of feature maps.

FPN is an effective framework to address the multiscale processing issue. However, the designs of FPN cause the lack of context information in feature maps. Here, to extract the global context information, we design the Attention Aggregation Module to enhance long-range dependencies on multi-level (Figure 4b and Figure 4c). Specifically, the four feature maps (i.e.  $S_2$ ,  $S_3$ ,  $S_4$ , and  $S_5$ ) generated by the corresponding feature pyramid are first concatenated and then fed into the  $1 \times 1$  convolutional layer. Thereafter, the linear attention mechanism is utilized to capture global context information and further refine fused feature maps. Finally, the refined features are added with the original concatenated features.

## 4. Experimental results

### 4.1. Datasets

We test the effectiveness of  $A^2$ -FPN based on the ISPRS Vaihingen and Potsdam datasets (<http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>), the Gaofen Image Dataset (GID) (Tong et al. 2020) as well as the UAvId dataset (Lyu et al. 2020).

#### 4.1.1. *Vaihingen*

There are 33 images as well as normalized digital surface models (nDSMs) in the Vaihingen dataset. The ground sampling distance (GSD) of tiles in Vaihingen is 9 cm and the average size is  $2494 \times 2064$  pixels. The image 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38 are selected for testing, image 30 for validation, and the remaining 15 images for training.

#### 4.1.2. *Potsdam*

The Potsdam dataset contains 38 images and nDSMs. The GSD Potsdam is 5 cm and the size of each tile is  $6000 \times 6000$ . We utilize 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, image 2\_10 for validation, and the remaining 22 images, except 7\_10 with error annotations, for training.

#### 4.1.3. *GID*

The GID contains 150 RGB images (Tong et al. 2020). Each image is in  $7200 \times 6800$  pixels which covers a geographic region of  $506\text{km}^2$  captured by the Gaofen 2 satellite. Following the previous work (Rui et al. 2021a), we select 15 images contained in GID, which cover the whole six categories. We partition each image into non-overlapping patch sets of size  $512 \times 512$  pixels. Thereafter, 50% patches are selected randomly as the training set, 10% patches are chosen as the validation set, and the remaining 40% patches are reserved as the test set.

#### 4.1.4. *UAVid*

UAVid is a fine-resolution Unmanned Aerial Vehicle (UAV) semantic segmentation dataset, which focuses on urban street scenes with a  $4096 \times 2160$  or  $3840 \times 2160$  resolution. UAVid is a very challenging benchmark since the large resolution of images, large-scale variation, and complexities in the scenes. To be specific, there are totally 420 images in the dataset where 200 of them are for training, 70 for validation, and the remaining 150 for testing.

### 4.2. *Evaluation metrics*

For ISPRS and GID datasets, the performance of our  $A^2$ -FPN, as well as comparative methods, is measured by the overall accuracy (OA), the mean Intersection over Union (mIoU), and the F1 score (F1). Based on the accumulated confusion matrix, the OA, mIoU, and F1 are computed as

$$OA = \frac{\sum_{k=1}^N TP_k}{\sum_{k=1}^N TP_k + FP_k + TN_k + FN_k}, \quad (14)$$

$$mIoU = \frac{1}{N} \sum_{k=1}^N \frac{TP_k}{TP_k + FP_k + FN_k}, \quad (15)$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (16)$$

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.

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

$$IoU = \frac{TP_k}{TP_k + FP_k + FN_k}. \quad (17)$$

### 4.3. Experimental setting

We implemented the proposed  $A^2$ -FPN and comparative algorithms using PyTorch under the Python platform and trained them using a single Tesla V100 with Adam optimizer. The learning rate is parametrized as 0.0003. For training, we cropped the original tiles into  $512 \times 512$  patches ( $1024 \times 1024$  for the UAVid dataset) and augmented them by rotating, resizing, horizontal axis flipping, vertical axis flipping, and adding random noise.

For benchmark comparisons on ISPRS and GID datasets, we considered not only the methods proposed initially for natural images, such as pyramid scene parsing network (PSPNet) (Zhao et al. 2017) and dual attention network (DANet) (Jun et al. 2019), but also the models designed for remote-sensing images, e.g. edge-aware neural network (EaNet) (Zheng et al. 2020). In addition, U-Net (Ronneberger, Fischer and Brox 2015), DABNet (Li et al. 2019), BiSeNetV2 (2021), and CE-Net (Gu et al. 2019) are also taken into account for a comprehensive comparison. The test time augmentation (TTA) in terms of rotating and flipping is applied for all algorithms accordingly.

As the training procedure on the UAVid dataset is extremely time-consuming and there are many publicly available results, we directly utilized models which were tested on the UAVid dataset as the comparative methods. Meanwhile, since most of those models are based on the ResNet-18, the backbone of the proposed  $A^2$ -FPN was also set as ResNet-18 for the UAVid dataset. The comparative models include MSD (Lyu et al. 2020), BiSeNet (Changqian et al. 2018), SwiftNet (Oršičić and Šegvić 2021), ShelfNet (Zhuang et al. 2019), MANet (Li et al. 2021c), BANet (Wang et al. 2021b), and ABCNet (Li et al. 2021d).

### 4.4. Results on the ISPRS Vaihingen dataset

We compare our method with seven existing methods on the Vaihingen test set and quantitative comparisons are shown in Table 1. For a fair comparison, the backbone of ResNet-based algorithms is set as ResNet-34 consistently. Our  $A^2$ -FPN outperforms other encoder-decoder methods (e.g. U-Net and CE-Net), attention-based methods (e.g. DANet), and context aggregation methods (e.g. PSPNet and EaNet) by a significant margin. To be specific, at least 1.6% in mean F1 score, 0.6% in OA, and 2.5% in mIoU higher than the other comparative methods. Especially, the F1 score of Car predicted by our  $A^2$ -FPN is far higher than any other approaches, which increase the second-best CE-Net by a large margin of 5.7%, demonstrating the effectiveness of the Attention Aggregation Module.

**Table 1.** He experimental results on the Vaihingen dataset.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
U-Net	-	84.3	86.5	73.1	83.9	40.8	73.7	82.0	64.0
DABNet	-	87.8	88.8	74.3	84.9	60.2	79.2	84.3	70.2
BiSeNetV2	-	89.9	91.9	82.0	88.3	71.4	84.7	88.0	75.5
PSPNet	ResNet-34	90.3	94.2	82.8	88.6	51.1	81.4	88.8	71.3
DANet	ResNet-34	91.1	94.8	83.5	88.9	63.0	84.3	89.5	74.4
EaNet	ResNet-34	92.8	95.2	82.8	89.3	80.6	88.0	90.0	79.1
CE-Net	ResNet-34	92.7	95.5	83.4	89.5	81.2	88.5	90.4	79.7
<b>A<sup>2</sup>-FPN</b>	<b>ResNet-34</b>	<b>93.0</b>	<b>95.7</b>	<b>84.7</b>	<b>90.0</b>	<b>86.9</b>	<b>90.1</b>	<b>91.0</b>	<b>82.2</b>

To qualitatively illustrate the effectiveness of the proposed A<sup>2</sup>-FPN, we provide qualitative comparisons between different networks via  $512 \times 512$  patches in Figure 6. Particularly, we leverage the red box to mark those intricate regions that are easy to be confused. Designed for real-time segmentation, the speed of BiSeNetV2 is relatively fast. However, the over-simplified structure leads to the deficiency of contextual information. EaNet adopts a large kernel pyramid pooling (LKPP) operation to capture contextual information, but the LKPP is only used for a single-scale feature map. By comparison, the elaborate attention aggregation across multiscale feature maps enables our A<sup>2</sup>-FPN to generate more accurate segmentation maps.

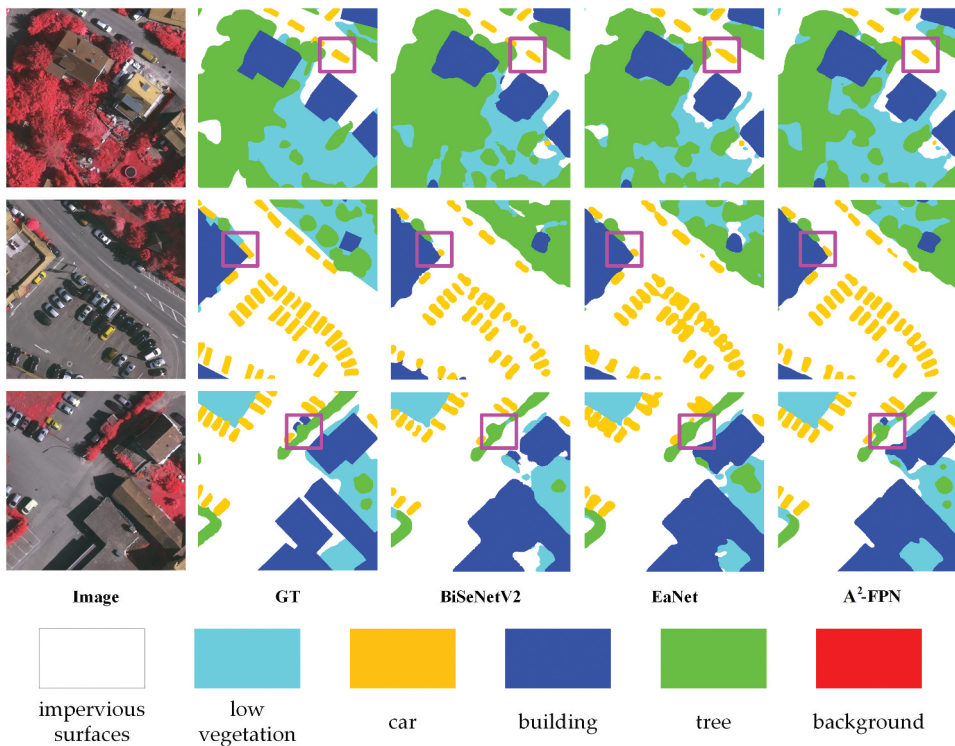
#### 4.5. Results on the ISPRS Potsdam dataset

To further evaluate the effectiveness of A<sup>2</sup>-FPN, we carry out experiments on the ISPRS Potsdam dataset. The training and testing settings on the Potsdam dataset are the same as the Vaihingen dataset. Numerical comparisons with comparative algorithms are listed in Table 2. The A<sup>2</sup>-FPN achieves up to 92.4% in mean F1 score, 91.1% in overall accuracy, and 86.1% in mIoU.

In Figure 7, we further visualize  $512 \times 512$  patches with the intractable regions marked by red rectangles. Our A<sup>2</sup>-FPN produces consistently better segmentation results than other benchmark approaches. Due to the loss of global contextual information, the segmentation maps generated by DABNet are ambiguous, particularly at the contour of objects. For example, in the first row of Figure 7, the edge of the low vegetation is not well recognized by DABNet but precisely captured by the proposed A<sup>2</sup>-FPN. Although CE-Net harnesses the context extractor to exploit contextual information, the utilization is on a single scale which is limited and insufficient. As can be seen in the second row of Figure 7, CE-Net mistakes the building and impervious surfaces. By contrast, the utilization of FPN and AAM enables the proposed A<sup>2</sup>-FPN to exploit the multiscale contextual information, thereby delivering an accurate and robust performance.

#### 4.6. Results on the GID dataset

We conducted experiments on the GID dataset to further test the accuracy of our A<sup>2</sup>-FPN. As listed in Table 3, our A<sup>2</sup>-FPN holds the leading position on the vast majority of the evaluation indexes. Visualized results in Figure 8 also demonstrates the superiority of our method. The built-up category is classified as others wrongly by U-Net on a large scale, while the PSPNet does not recognize the intervals in the meadow. These mistakes are well addressed by our A<sup>2</sup>-FPN, benefiting from the utilization of multiscale contextual information.



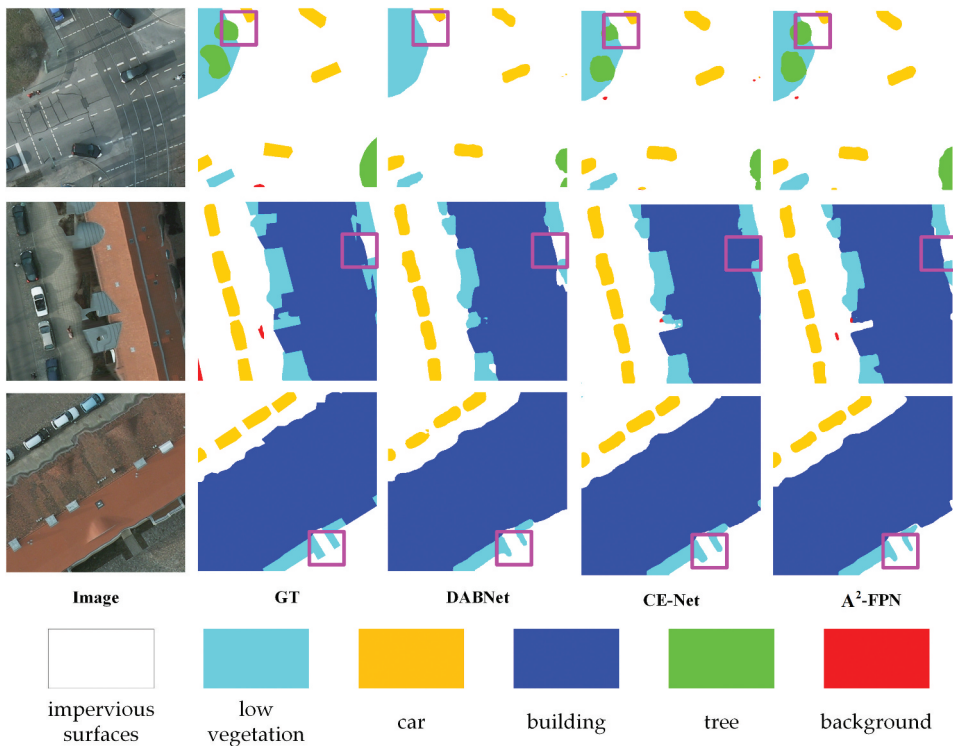
**Figure 6.** Visualization of results on the Vaihingen dataset.

**Table 2.** The experimental results on the Potsdam dataset.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
U-Net	-	85.0	88.8	76.7	73.1	90.3	82.8	80.6	74.3
DABNet	-	89.9	93.2	83.6	82.3	92.6	88.3	86.7	79.6
BiSeNetV2	-	91.3	94.3	85.0	85.2	94.1	90.0	88.2	82.3
PSPNet	ResNet-34	91.6	95.8	86.0	87.7	86.5	89.5	89.5	82.6
DANet	ResNet-34	91.9	96.1	85.6	87.6	86.8	89.6	89.6	82.6
EaNet	ResNet-34	92.4	96.3	85.6	87.9	95.1	91.5	89.7	85.2
CE-Net	ResNet-34	92.5	96.4	86.4	87.8	95.3	91.7	90.0	85.4
<b>A<sup>2</sup>-FPN</b>	<b>ResNet-34</b>	<b>93.6</b>	<b>96.9</b>	<b>87.5</b>	<b>88.4</b>	<b>95.7</b>	<b>92.4</b>	<b>91.1</b>	<b>86.1</b>

**Table 3.** The experimental results on the GID dataset.

Method	Backbone	Build-up	Forest	Farmland	Meadow	Water	others	Mean F1	OA (%)	mIoU (%)
U-Net	-	82.3	85.0	89.7	84.1	93.2	69.2	83.9	82.3	73.0
DABNet	-	81.7	86.9	90.6	85.9	94.2	72.7	85.3	83.9	75.0
BiSeNetV2	-	83.0	86.4	90.2	86.4	94.7	72.4	85.5	83.9	75.4
PSPNet	ResNet-34	84.2	89.1	91.5	87.6	95.1	76.4	87.3	86.1	77.9
DANet	ResNet-34	84.8	89.5	91.7	87.8	95.6	77.8	87.9	86.7	78.8
EaNet	ResNet-34	85.2	90.4	91.8	86.4	96.2	78.4	88.1	87.3	79.1
CE-Net	ResNet-34	85.9	90.2	92.2	87.4	96.5	79.4	88.6	87.7	79.9
<b>A<sup>2</sup>-FPN</b>	<b>ResNet-34</b>	<b>86.3</b>	<b>91.0</b>	<b>92.4</b>	<b>87.9</b>	<b>96.8</b>	<b>79.9</b>	<b>89.1</b>	<b>88.3</b>	<b>80.7</b>



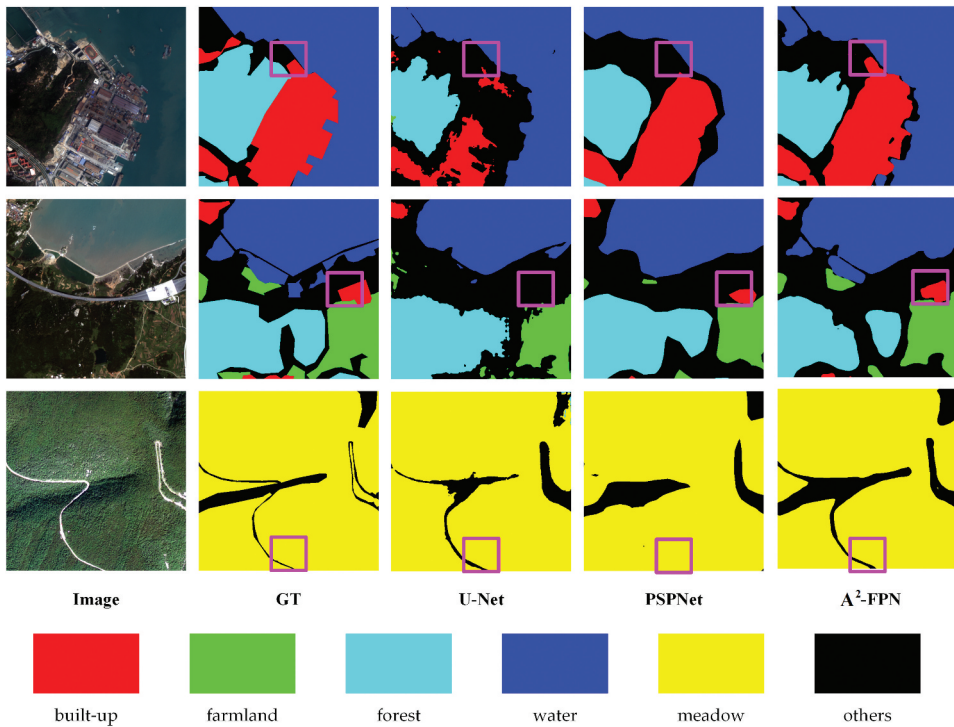
**Figure 7.** Visualization of results on the Potsdam dataset.

**Table 4.** The experimental results on the UAvid dataset.

Method	Backbone	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
BiSeNet	ResNet-18	85.7	78.3	64.7	61.1	<b>77.3</b>	<b>63.4</b>	48.6	17.5	61.5
SwiftNet	ResNet-18	85.3	78.2	64.1	61.5	76.4	62.1	51.1	15.7	61.1
ShelfNet	ResNet-18	76.9	73.2	44.1	61.4	43.4	21.0	52.6	3.6	47.0
MANet	ResNet-18	85.4	77.0	64.5	77.8	60.3	53.6	67.2	14.9	62.6
BANet	ResT-Lite	85.4	78.9	66.6	80.7	62.1	52.8	69.3	21.0	64.6
ABCNet	ResNet-18	86.4	79.9	<b>67.4</b>	<b>81.2</b>	63.1	48.4	69.8	13.9	63.8
A <sup>2</sup> -FPN	ResNet-18	<b>87.2</b>	<b>80.1</b>	<b>67.4</b>	80.2	63.7	53.3	<b>70.1</b>	<b>23.4</b>	<b>65.7</b>

#### 4.7. Results on the UAvid dataset

As illustrated in Table 4, the proposed A<sup>2</sup>-FPN achieves the best IoU score on five out of eight classes and the best mIoU with a 1% gain over the suboptimal BANet. Considering the UAvid is a relatively large-scale dataset, the result strongly demonstrates the effectiveness of the proposed A<sup>2</sup>-FPN. Since the ground truth of the test set is not available now, we visualize and compare the results generated by our A<sup>2</sup>-FPN and the official benchmark, i.e. MSD (Lyu et al. 2020). Compared with the baseline MSD with obvious local and global inconsistencies, the proposed A<sup>2</sup>-FPN can effectively capture the cues to scene semantics. For instance, in the third row of Figure 9, the cars in the pink box are obviously all moving on the road. However, the MSD identifies those cars, which are crossing the street as static cars. In contrast, our A<sup>2</sup>-FPN correctly recognizes all moving cars.



**Figure 8.** Visualization of results on the GID dataset.

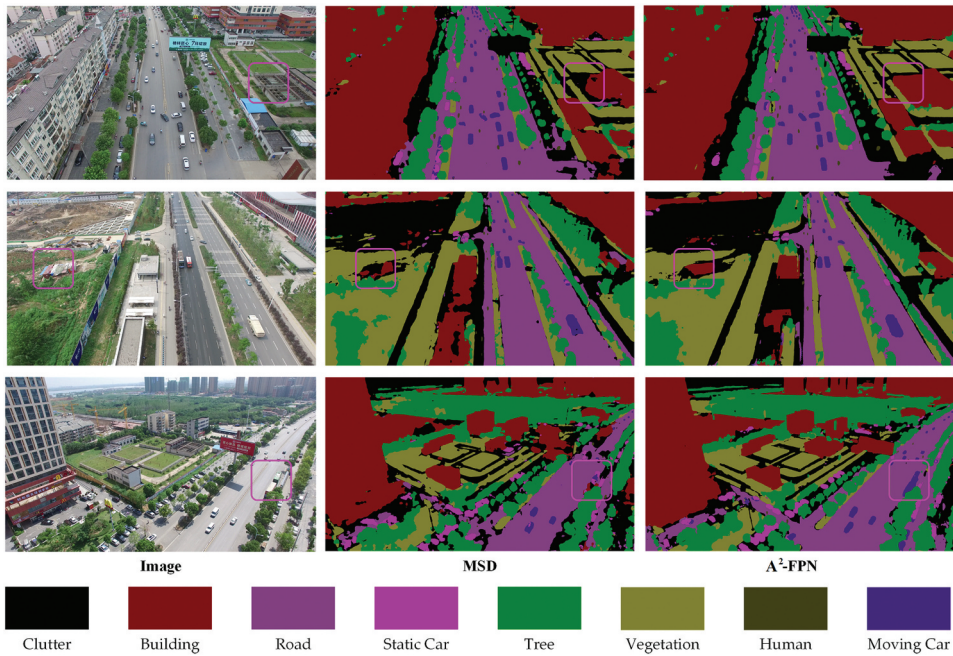
## 5. Discussion

### 5.1. Ablation study about FPN and AAM

Ablation experiments were conducted to test the effectiveness of FPN and AAM in the proposed  $A^2$ -FPN. The encoder-decoder structure based on ResNet-34 is selected as the baseline. As shown in Table 5, the FPN outperforms the encoder-decoder baseline significantly. For the Vaihingen dataset, the introduction of FPN brings more than 3.6% in mean F1 score, 1.1% in OA, and 3.8% in mIoU, while the improvements for the Potsdam dataset is 0.6%, 0.7%, and 2.7%, respectively. The FPN is initially designed for object detection. To tackle the segmentation issue, the feature maps generated by feature pyramids are simply concatenated, lacking the global context information crucial for segmentation. Therefore, the Attention Aggregation Module is developed to address the above limitation. As a specifically designed module for semantic segmentation, the

**Table 5.** Ablation study about FPN and AAM.

Dataset	Method	Backbone	Mean F1	OA	mIoU
Vaihingen	Baseline	ResNet-34	85.9	89.5	77.5
	FPN	ResNet-34	89.5	90.4	81.3
	$A^2$ -FPN	ResNet-34	90.1	91.0	82.2
Potsdam	Baseline	ResNet-34	91.1	89.5	82.7
	FPN	ResNet-34	91.7	90.2	85.4
	$A^2$ -FPN	ResNet-34	92.4	91.1	86.1
GID	Baseline	ResNet-34	87.4	86.1	78.0
	FPN	ResNet-34	88.4	87.5	79.7
	$A^2$ -FPN	ResNet-34	89.1	88.3	80.7



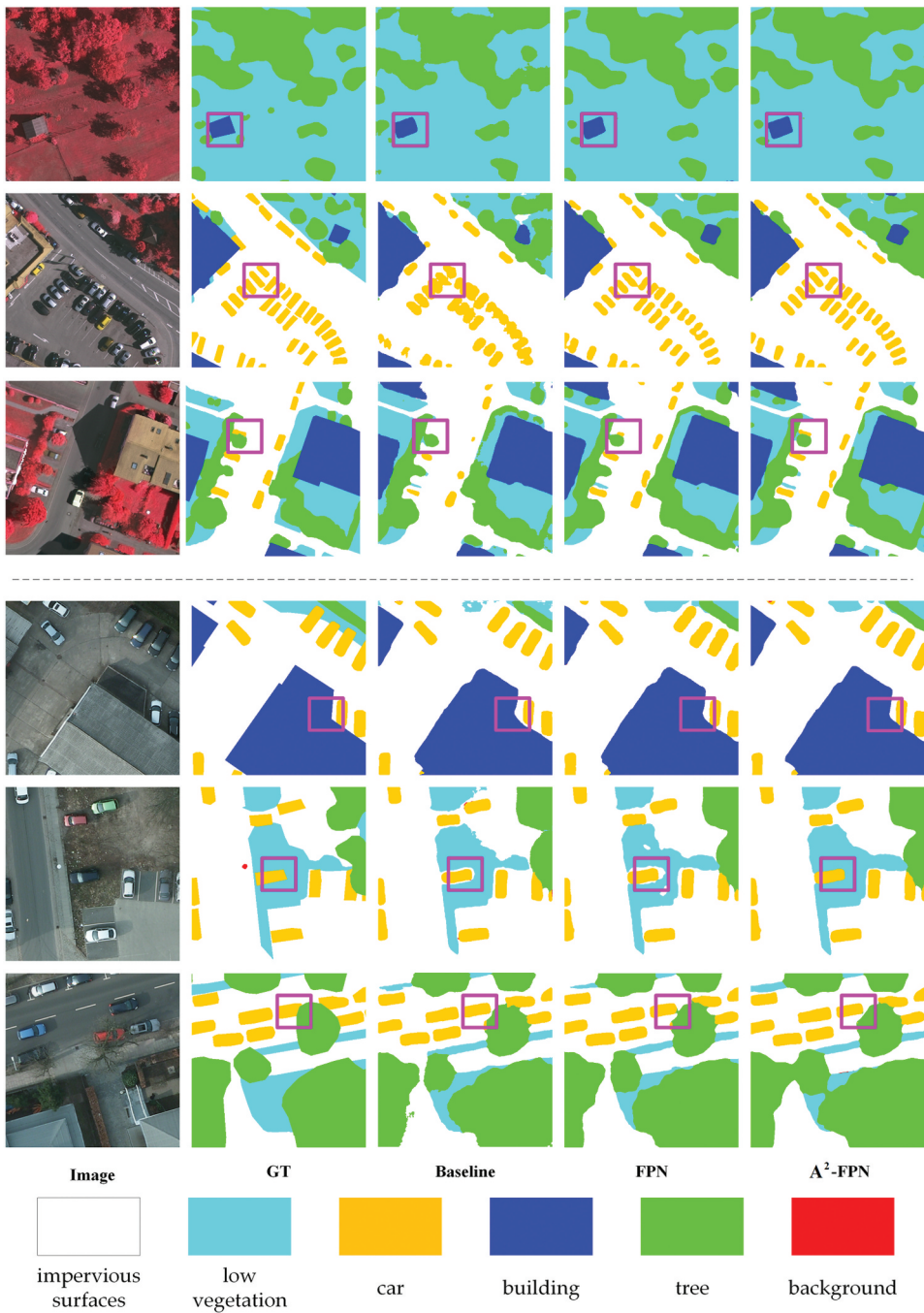
**Figure 9.** Visualization of results on the UAVid dataset.

utilization of AAM contributes to the increase of more than 0.6% in mean F1 score, 0.6% in OA, and 0.9% in mIoU for the Vaihingen dataset, while the figures for the Potsdam dataset are about 0.7%, 0.9%, and 0.7%, respectively. For qualitative comparison, we visualize certain segmentation maps generated by the baseline, FPN, and our  $A^2$ -FPN, which can be seen from Figure 10. Besides, the increases brought by the AAM on the GID dataset are about 0.7% in mean F1 score, 0.8% in OA, and 1.0% in mIoU, and the visualization results are shown in Figure 11.

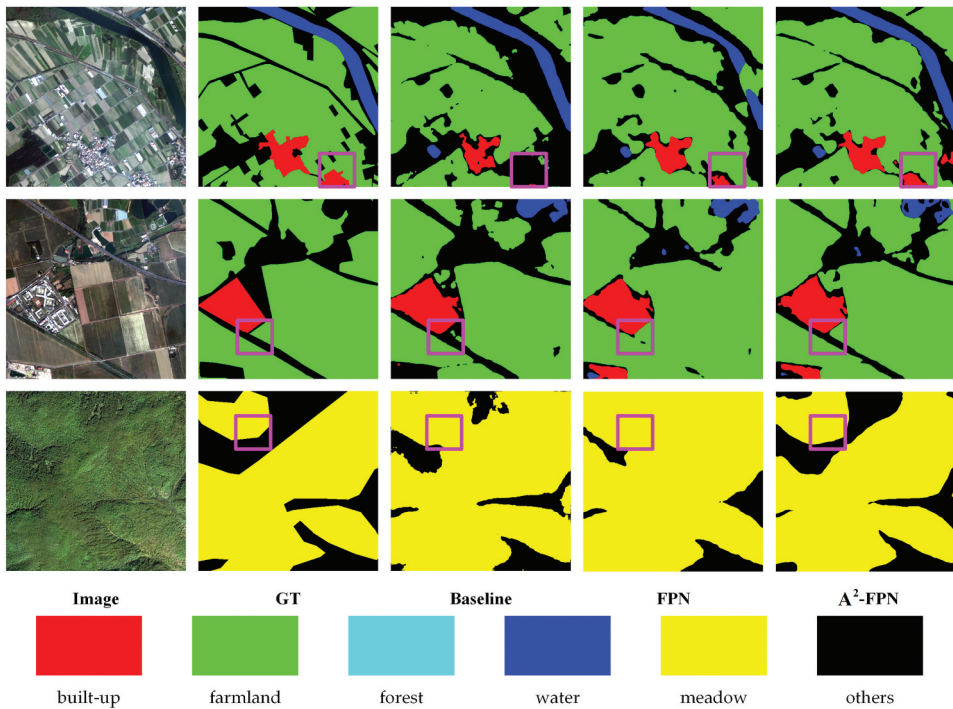
### 5.2. Ablation study about multi-head and dot-product attention

To demonstrate the advancement and efficiency of the proposed AAM, we replace the linear attention mechanism in AAM with the multi-head and dot-product attention mechanism to conduct the ablation study. Meanwhile, the inference speeds measured in frames per second (FPS) on a mid-range notebook graphics card 1660Ti are also reported. As can be seen in Table 6, the multi-head attention, i.e.  $A^2$ -FPN (M), can indeed enhance the performance, but the inference speed (24.98 FPS) will be lowered 2.6 times compared with  $A^2$ -FPN (65.44 FPS), which may be not a cost-effective scheme. After replacing the linear attention mechanism with dot-product attention mechanism, the network, i.e.  $A^2$ -FPN (D), will occupy about 16.4 GB memory under 2 batch sizes for  $512 \times 512$  inputs, while the figure for the raw  $A^2$ -FPN is 15.1 GB under 16 batch sizes. That is, there is more than an 8 times gap between the memory requirements between the  $A^2$ -FPN (D) and the proposed  $A^2$ -FPN. In addition, the inference speed will be lowered to 12.96 FPS due to the high complexity. Therefore, the design of the AAM balances the accuracy and efficiency well.





**Figure 10.** Visualization of ablation study on (top) the Vaihingen dataset and (bottom) the Potsdam dataset.



**Figure 11.** Visualization of ablation study on the GID dataset.

**Table 6.** Ablation study about multi-head attention and dot-product attention mechanism.

Method	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
A2-FPN	93.0	95.7	84.7	90.0	86.9	90.1	91.0	82.2
A2-FPN (M)	93.2	95.7	85.0	89.9	87.7	90.3	91.1	82.6
A2-FPN (D)	92.3	95.1	84.3	89.9	82.8	88.9	90.5	81.5

**Table 7.** The complexity and speed of the proposed A<sup>2</sup>-FPN and other methods.

Method	Complexity (G)	Parameters (M)	256 × 256	512 × 512	1024 × 1024	2048 × 2048
U-Net	247.85	43.42	30.16	10.64	2.75	*
DABNet	5.22	0.75	102.31	87.74	34.88	8.77
BiSeNetV2	13.91	12.30	129.71	111.70	31.23	7.07
PSPNet	22.24	34.14	156.66	83.92	26.08	6.94
DANet	19.58	22.78	111.40	81.54	24.43	7.14
EaNet	28.43	44.34	96.04	54.58	14.90	4.26
CE-Net	39.98	29.00	101.49	45.33	13.71	3.52
A <sup>2</sup> -FPN	22.93	22.27	107.12	65.44	16.87	4.60

The complexity and parameters are measured under the  $512 \times 512$  input, where ‘G’ indicates Gillion (i.e. units for the number of floating point operations) and ‘M’ signifies Million (i.e. units for the number of parameters). For an extensive comparison, we chose  $256 \times 256$ ,  $512 \times 512$ ,  $1024 \times 1024$ , and  $2048 \times 2048$  pixels as the sizes of the input image and report the inference speed measured in frames per second (FPS) on a mid-range notebook graphics card 1660ti. \* means out of memory.

### 5.3. Limitation

Although the proposed A<sup>2</sup>-FPN has bridged the gap between low-level and high-level features and compensated for the weakness of the raw FPN, there are still some potential issues that need to be considered.

First, the total trainable parameters in the  $A^2$ -FPN are 22.27 M, which is less than medium-scale networks such as DANet (22.78 M), PSPNet (34.14 M), and EaNet (44.34 M) while larger than those small-scale networks such as BiSeNetV2 (12.30 M). To extensively compare the efficiency, we report the complexity and the parameters of each method as well as the inference speed. As demonstrated in experimental results, CE-Net and EaNet are significantly superior to other comparative methods except for the proposed  $A^2$ -FPN. In Table 7, we can see that the complexity, parameters, as well as speed of our  $A^2$ -FPN, all have advantages over CE-Net and EaNet, indicating a better structure that balance the accuracy and efficiency well.

Second, the incorporation of auxiliary information (e.g. DSMs) might further increase the accuracy. However, these require intelligent approaches to handle computationally intensive operations to include more information. Our future work will, therefore, be devoted to realizing real-time semantic segmentation, as well as developing efficient techniques to fuse DSMs or nDSMs, thereby further enhancing the segmentation performance.

## 6. Conclusion

The automatic semantic segmentation from fine-resolution remotely sensed images remains a complicated and challenging task, due to the limited spatial and contextual information utilized. In this research, we employ the Feature Pyramid Network to combine the extracted spatial and contextual features comprehensively. In particular, the pyramidal hierarchy enables FPN to combine low-level detailed spatial information with high-level abundant semantic features thoroughly. Besides, to enhance the segmentation accuracy, we propose an Attention Aggregation Module to not only effectively merge the feature maps but also to fully extract the context information. Although there exist some pieces of literature which have explored the combination of attention mechanisms and FPN, the attention mechanisms utilized in their models are either dot-product attention mechanisms or scaling attention mechanisms. The former has expensive computing consumptions, while the latter is unable to extract contextual information. By contrast, we first introduce the linear attention mechanism, i.e. a simplified version of dot-product attention mechanisms to the FPN. Substantial experiments conducted on the ISPRS Vaihingen, Potsdam, and GID datasets demonstrate the effectiveness of our  $A^2$ -FPN. The extensive ablation studies illustrate the validity of FPN and AAM accordingly.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported in part by the National Natural Science Foundation of China [No.41671452].

## ORCID

Rui Li  <http://orcid.org/0000-0001-7858-3160>  
Libo Wang  <http://orcid.org/0000-0001-8096-6531>  
Ce Zhang  <http://orcid.org/0000-0001-5100-3584>  
Chenxi Duan  <http://orcid.org/0000-0003-0056-3295>

## Author's Contributions

This work was conducted in collaboration with all authors. Shunyi Zheng supervised the research work and provided experimental facilities. Rui Li and Chenxi Duan designed the semantic segmentation model and conducted the experiments. This manuscript was written by Rui Li and Chenxi Duan. Ce Zhang and Libo Wang checked the experimental results. All authors have read and agreed to the published version of the manuscript.

## References

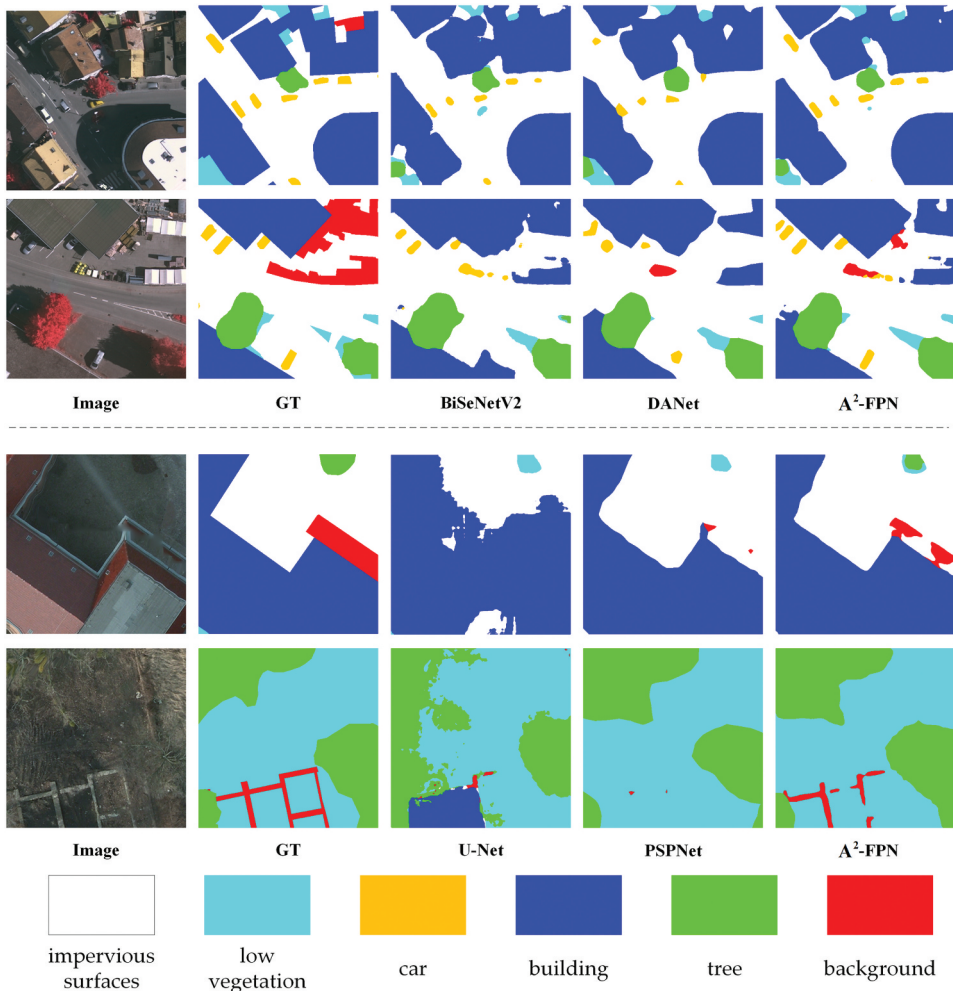
- Audebert, N., B. Le Saux, and S. Lefèvre. 2018. "Beyond RGB: Very High Resolution Urban Remote Sensing with Multimodal Deep Networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 140: 20. doi:10.1016/j.isprsjprs.2017.11.011.
- Bai, Y., G. Sun, Y. Li, P. Ma, G. Li, and Y. Zhang. 2021. "Comprehensively Analyzing Optical and Polarimetric SAR Features for Land-Use/Land-Cover Classification and Urban Vegetation Extraction in Highly-Dense Urban Area." *International Journal of Applied Earth Observation and Geoinformation* 103: 102496. doi:10.1016/j.jag.2021.102496.
- Changqian, Y., C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang. 2021. "Bisenet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation." *International Journal of Computer Vision* 129 3051 doi:10.1007/s11263-021-01515-2 .
- Changqian, Y., J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. 2018. "Bisenet: Bilateral Segmentation Network for Real-Time Semantic Segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, 325.
- Duan, C., J. Pan, and R. Li. 2020. "Thick Cloud Removal of Remote Sensing Images Using Temporal Smoothness and Sparsity Regularized Tensor Optimization." *Remote Sensing* 12 (20): 3446. doi:10.3390/rs12203446.
- Gu, Z., J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. 2019. "Ce-Net: Context Encoder Network for 2d Medical Image Segmentation." *IEEE Transactions on Medical Imaging* 38 (10): 2281. doi:10.1109/TMI.2019.2903562.
- He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770.
- Jun, F., J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. 2019. "Dual Attention Network for Scene Segmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146.
- Kampffmeyer, M., A.-B. Salberg, and R. Jenssen. 2016. "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*, 1.
- Kampffmeyer, M., A.-B. Salberg, and R. Jenssen. 2018. "Urban Land Cover Classification with Missing Data Modalities Using Deep Convolutional Neural Networks." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (6): 1758. doi:10.1109/JSTARS.2018.2834961.
- Katharopoulos, A., A. Vyas, N. Pappas, and F. Fleuret. 2020. "Transformers are Rnns: Fast Autoregressive Transformers with Linear Attention." In *International Conference on Machine Learning*, 5156. PMLR.
- Kirillov, A., R. Girshick, K. He, and P. Dollár. 2019. "Panoptic Feature Pyramid Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6399.
- Li, R., C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson. 2021a. "MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images." *IEEE Geoscience and Remote Sensing Letters* 19 1 doi:10.1109/LGRS.2021.3052886 .
- Li, G., I. Yun, J. Kim, and J. Kim. 2019. "Dabnet: Depth-Wise Asymmetric Bottleneck for Real-Time Semantic Segmentation British Machine Vision Conference." .

- Li, H., C. Zhang, S. Zhang, and P. M. Atkinson. 2020a. "Crop Classification from Full-Year Fully-Polarimetric L-Band UAVSAR Time-Series Using the Random Forest Algorithm." *International Journal of Applied Earth Observation and Geoinformation* 87: 102032. doi:10.1016/j.jag.2019.102032.
- Li, R., S. Zheng, C. Duan, J. Su, and C. Zhang. 2021b. "Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* 19 1 doi:10.1109/LGRS.2021.3063381 .
- Li, R., S. Zheng, C. Duan, L. Wang, and C. Zhang. 2022. "Land Cover Classification from Remote Sensing Images Based on Multi-Scale Fully Convolutional Network." *Geo-Spatial Information Science* 1 doi:10.1080/10095020.2021.2017237.
- Li, R., S. Zheng, C. Duan, Y. Yang, and X. Wang. 2020b. "Classification of Hyperspectral Image Based on Double-Branch Dual-Attention Mechanism Network." *Remote Sensing* 12 (3): 582. doi:10.3390/rs12030582.
- Li, R., S. Zheng, C. Zhang, C. Duan, S. Jianlin, L. Wang, and P. M. Atkinson. 2021c. "Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images 60 ." , 1 doi:10.1109/TGRS.2021.3093977.
- Li, R., S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson. 2021d. "Abcnet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 181: 84. doi:10.1016/j.isprsjprs.2021.09.005.
- Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature Pyramid Networks for Object Detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117.
- Lyu, Y., G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Ying Yang. 2020. "Uavid: A Semantic Segmentation Dataset for UAV Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 165: 108. doi:10.1016/j.isprsjprs.2020.05.009.
- Maggiore, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2017. "High-Resolution Aerial Image Labeling with Convolutional Neural Networks." *IEEE Transactions on Geoscience and Remote Sensing* 55 (12): 7092. doi:10.1109/TGRS.2017.2740362.
- Marmanis, D., K. Schindler, J. Dirk Wegner, S. Galliani, M. Datcu, and U. Stilla. 2018. "Classification with an Edge: Improving Semantic Image Segmentation with Boundary Detection." *ISPRS Journal of Photogrammetry and Remote Sensing* 135: 158. doi:10.1016/j.isprsjprs.2017.11.009.
- Oršičić, M. and S. Šegvić. 2021. "Efficient Semantic Segmentation with Pyramidal Fusion." *Pattern Recognition* 110: 107611. doi:10.1016/j.patcog.2020.107611.
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, 234. Springer.
- Shamsolmoali, P., M. Zareapoor, H. Zhou, R. Wang, and J. Yang. 2020. "Road Segmentation for Remote Sensing Images Using Adversarial Spatial Pyramid Networks." *IEEE Transactions on Geoscience and Remote Sensing* 59 6 4673 doi:10.1109/TGRS.2020.3016086. .
- Sherrah, J. 2016. "Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery." *arXiv preprint arXiv:1606.02585*.
- Shoebly, M., A. Armin, S. Aliakbarian, S. Anwar, and L. Petersson. 2020. "Mosaic Super-Resolution via Sequential Feature Pyramid Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 84.
- Tong, X.-Y., G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang. 2020. "Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models." *Remote Sensing of Environment* 237: 111322.
- Tu, Y., B. Chen, T. Zhang, and B. Xu. 2020. "Regional Mapping of Essential Urban Land Use Categories in China: A Segmentation-Based Approach." *Remote Sensing* 12 (7): 1058.
- Wambugu, N., Y. Chen, Z. Xiao, M. Wei, S. Aminu Bello, J. Marcato Junior, and J. Li. 2021. "A Hybrid Deep Convolutional Neural Network for Accurate Land Cover Classification." *International Journal of Applied Earth Observation and Geoinformation* 103: 102515.
- Wang, L., R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang. 2021a. "A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* doi:10.1109/LGRS.2022.3143368.

- Wang, L., R. Li, D. Wang, C. Duan, T. Wang, and X. Meng. 2021b. "Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images." *Remote Sensing* 13 (16): 3065.
- Wei, Y., K. Zhang, and S. Ji. 2020. "Simultaneous Road Surface and Centerline Extraction from Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing." *IEEE Transactions on Geoscience and Remote Sensing* 58 (12): 8919.
- Xiaowei, G., P. Angelov, C. Zhang, and P. Atkinson. 2020. "A Semi-Supervised Deep Rule-Based Approach for Complex Satellite Sensor Image Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi:10.1109/TPAMI.2020.3048268.
- Zhang, C., P. M. Atkinson, C. George, Z. Wen, M. Diazgranados, and F. Gerard. 2020a. "Identifying and Mapping Individual Plants in a Highly Diverse High-Elevation Ecosystem Using UAV Imagery and Deep Learning." *ISPRS Journal of Photogrammetry and Remote Sensing* 169: 280.
- Zhang, C., P. A. Harrison, X. Pan, H. Li, I. Sargent, and P. M. Atkinson. 2020b. "Scale Sequence Joint Deep Learning (SS-JDL) for Land Use and Land Cover Classification." *Remote Sensing of Environment* 237: 111593.
- Zhang, C., I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, and P. M. Atkinson. 2019. "Joint Deep Learning for Land Cover and Land Use Classification." *Remote Sensing of Environment* 221: 173.
- Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. 2017. "Pyramid Scene Parsing Network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881.
- Zheng, X., L. Huan, G.-S. Xia, and J. Gong. 2020. "Parsing Very High Resolution Urban Scene Images by Learning Deep ConvNets with Edge-Aware Loss." *ISPRS Journal of Photogrammetry and Remote Sensing* 170: 15.
- Zhu, X. X., D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. 2017. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." *IEEE Geoscience and Remote Sensing Magazine* 5 (4): 8.
- Zhuang, J., J. Yang, L. Gu, and N. Dvornek. 2019. "Shelfnet for Fast Semantic Segmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

## Appendix A More visual results

This section provides more visual results of our method. For ISPRS Vaihingen and Potsdam datasets, the accuracy of the background class which contains complex clatters is relatively low. As can be seen from the top and middle rows in [Figure A1](#), the misclassification of background class (in red color) is more distinct than others. The reason is that the background class is not a well-defined category that may contain several land cover types with different features. For the GID dataset, the fine-grained details cannot be generated successfully. For example, the interval (labeled as others in black color) between the farmland (in green color) in [Figure A2](#) is not distinguished. Similarly, the reason why the accuracy of others is relatively low is that the others class contains all other land cover types except the labeled categories in the GID dataset. Besides, the segmentation maps of the whole image on the ISPRS dataset is provided in [Figure A3](#) and [Figure A4](#).



**Figure A1.** The failure cases in the (top) Vaihingen dataset and (bottom) Potsdam dataset.

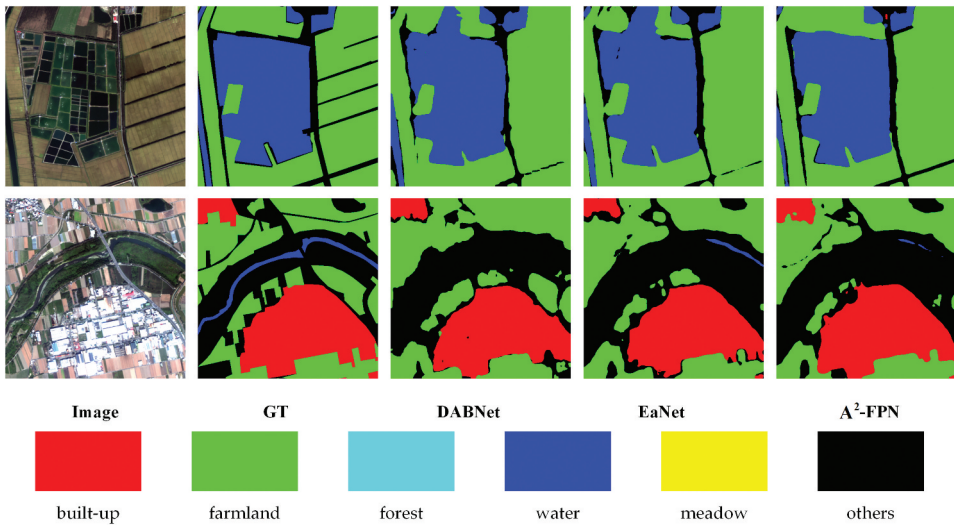


Figure A2. The failure cases in GID dataset.

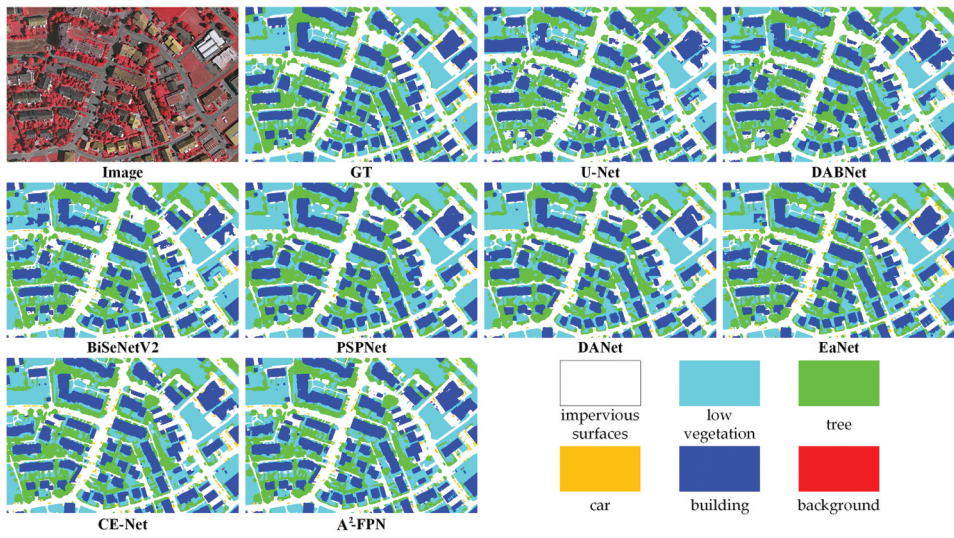
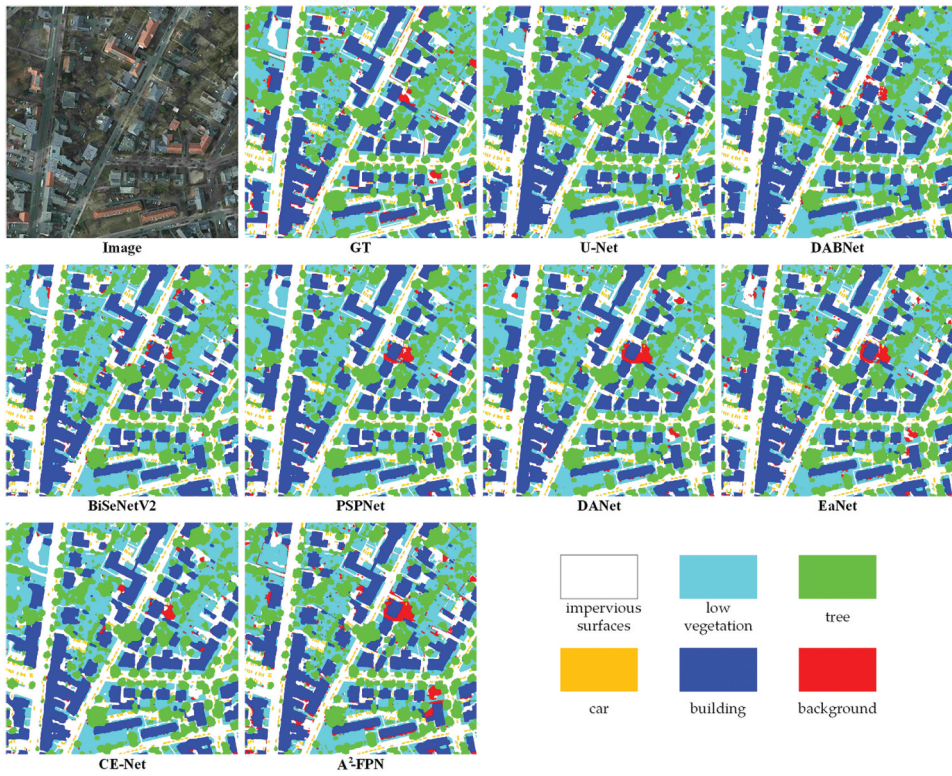


Figure A3. Visualization of tile-38 in the Vaihingen dataset.





**Figure A4.** Visualization of tile-38 in the Potsdam dataset.