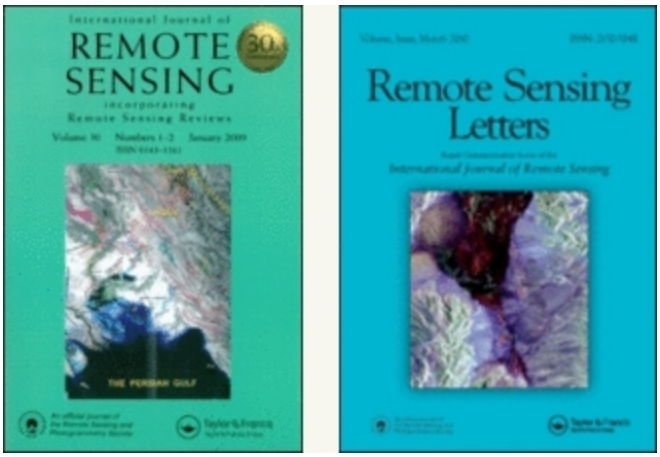


# A²-FPN for Semantic Segmentation of Fine-Resolution Remotely Sensed Images

| | |
|---|---|
| Journal: | *International Journal of Remote Sensing* |
| Manuscript ID | TRES-PAP-2021-1409 |
| Manuscript Type: | IJRS Research Paper |
| Date Submitted by the Author: | 21-Dec-2021 |
| Complete List of Authors: | Li, Rui; Wuhan University, School of Remote Sensing and Information Engineering<br>Wang, Libo; Wuhan University, School of Remote Sensing and Information Engineering<br>Zhang, Ce; Lancaster University, Lancaster Environment Centre; UK Centre for Ecology & Hydrology<br>Duan, Chenxi; University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC)<br>Zheng, Shunyi; Wuhan University, School of Remote Sensing and Information Engineering |
| Keywords: | segmentation, land use |
| Keywords (user defined): | |
| | |

Dear Professor Dongping Ming:
Associate Editor
International Journal of Remote Sensing

On behalf of my co-authors, we thank you very much for allowing us to revise the manuscript, and we are grateful to three reviewers for their constructive comments and suggestions on our manuscript titled "$A^2$-FPN for Semantic Segmentation of Fine-Resolution Remotely Sensed Images" (ID: TRES-PAP-2021-1166.R1).

We have revised the manuscript carefully according to the comments, and have documented our revisions in the part of the "response to reviewers". Manuscripts of the "clean" revision and the "highlight" version of the revision were attached, respectively. In our point-by-point response letter attached below, the comments of each reviewer in plain text followed by our responses in blue text are provided below. The major change we have made in this version is the supplemental experiments on a very large-scale segmentation dataset, i.e., UAVid. To be specific, there are totally 420 images with large resolution in the dataset where 200 of them are for training, 70 for validation, and the remaining 150 for testing. Experimental results demonstrate the effectiveness of the proposed $A^2$-FPN again.

We trust that you will find the revised manuscript acceptable for publication in the *International Journal of Remote Sensing.*

Looking forward to hearing from you.
Best wishes,
Rui Li (lironui@whu.edu.cn) and Chenxi Duan (c.duan@utwente.nl) on behalf of all co-authors.

**Response to Reviewer**

We are grateful to the anonymous reviewers for their constructive comments and suggestions and have revised the manuscript point-by-point carefully in response to their advice. The comments of each reviewer in plain text followed by our responses in blue text are provided below.

## Reviewer #1

------ It seems that this is a manuscript that has been reviewed. This is a work to improve the semantic segmentation model. The method mentioned in the article has already had a lot of similar work without much innovation. It is more like a combination of multiple models. However, the experimental verification process is reasonable and effective. If other reviewers agree to accept, I won't comment much.

**Response:** Many thanks for reviewing our manuscript. Actually, our manuscript is indeed has been reviewed. For your concern about innovation, we do know that there are many FPN-based models for object detection and instance segmentation. But as far as we know, the applications of FPN on semantic segmentation, especially for remote sensing images, are not as much as those on object detection and instance segmentation.

Besides, although there exist some pieces of literature which have explored the combination of attention mechanism and FPN, the attention mechanisms utilized in their models are totally different from ours. Actually, there are two types of technologies with the name of attention mechanism. One is the dot-product attention mechanism, which is designed to model long-range dependencies and enable contextual information extraction at a global scale. The other is scaling attention designed to reinforce informative features and whittle information-lacking features, while the typical examples are the squeeze-and-excitation (SE) model and the convolutional block attention module (CBAM). These two types of attention mechanisms have completely different principles and purposes. A comprehensive comparison between dot-product attention, scaling attention, as well as the simplified dot-product attention mechanism on semantic segmentation, can be seen in our previous work (DOI: https://doi.org/10.1109/TGRS.2021.3093977). The existing researches on the combination of attention mechanism and FPN are based on either dot-product attention which has expensive computing consumptions or scaling attention which is completely different from the attention used in our model. Actually, only in recent two years, the simplification of the dot-product attention mechanism just has gotten more and more focused. Therefore, although there are seemingly many 'similar' researches, the actual similarities are not many, while the main contribution of our work is to combine linear attention and FPN.

Specifically, in the revised version, we supplement the experiment on a very large-scale dataset, i.e., UAVid. There are totally 420 images with large resolution (4096×2160 or 3840×2160) in the dataset where 200 of them are for training, 70 for validation, and the remaining 150 for testing. The quantitative results can be seen in Table 4:

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 | **77.3** | **63.4** | 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 | 67.4 | 81.2 | 63.1 | 48.4 | 69.8 | 13.9 | 63.8 |
| A2-FPN | ResNet-18 | **87.2** | **80.1** | **67.4** | 80.2 | 63.7 | 53.3 | 70.1 | **23.4** | **65.7** |

Considering the UAVid is a relatively large-scale dataset, the result strongly demonstrates the effectiveness of the proposed $A^2$-FPN.

## Reviewer #2

------ Comments to the Author

**Response:** Thank you very much for reviewing our manuscript and for providing us with valuable comments. We studied your comments and responded to them point by point as below.

**Q1.** The paper is fine and the work is competently done but I raised a concern about plagiarism. Is Duplicate Publication a Plagiarism?
- Ref: Hu, Miao, et al. "A2-FPN: Attention Aggregation Based Feature Pyramid Network for Instance Segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

**Response:** Many thanks for your concern. Actually, our manuscript is not published by any Journal in any form before. Meanwhile, the only similarities between the suggested paper and our manuscript are the name of the network and the baseline of FPN. They try to address the **instance segmentation**, while we try to tackle the **semantic segmentation**; they aggregate the context by **scaled cosine-similarity attention mechanism**, while we aggregate the context by the **linear attention mechanism**. Although both are based on FPN, the methods are totally different; although both for segmentation, the instance-oriented and the semantic-oriented tasks are also fully different. Besides, the writing styles are also sheerly different. In summary, based on the same FPN structure, we and Hu Miao et, al utilize different methods to solve different tasks. Most importantly, they submitted the paper to the arXiv on **7 May 2021**, while we submitted our paper to the arXiv on **16 Feb 2021**. Our manuscript does not have any relationships with the suggested one or anyone that has been published. Therefore, we do not think our paper constitutes any academic misconduct, neither plagiarism nor duplicate publication.

**Q2.** Furthermore, the authors do not report standard deviations for their experiments. (Could you report the number of trials and standard deviations for your experiments?)

**Response:** Thanks very much for your question. Actually, we adopt the early stopping strategy when training the models. To be specific, if the accuracy on the validation set does not increase for more than 20 epochs, then we will stop the training procedure. By the above operation, we obtain the optimal model. Therefore, we only report the accuracy of the selected optimal model. In the revised manuscript, to support and demonstrate the effectiveness of the proposed network further, we supplement an experiment on a very large dataset, i.e., UAVid. There are totally 420 images with large resolution (4096×2160 or 3840×2160) in the dataset where 200 of them are for training, 70 for validation, and the remaining 150 for testing. 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 quantitative results can be seen in the official website (https://competitions.codalab.org/competitions/public_submissions/25224) as well as in Table 4:

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 | **77.3** | **63.4** | 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 | 67.4 | 81.2 | 63.1 | 48.4 | 69.8 | 13.9 | 63.8 |
| A2-FPN | ResNet-18 | **87.2** | **80.1** | **67.4** | 80.2 | 63.7 | 53.3 | 70.1 | **23.4** | **65.7** |

Considering the scale of UAVid, the result strongly demonstrates the effectiveness of the proposed A$^2$-FPN.

## Reviewer #3

------ The method used in this paper is sound.

**Response:** Thank you very much for reviewing our manuscript and for providing us with valuable comments. We studied your comments and responded to them point by point as below.

**Q1.** However there are so many similar research papers related to attention aggregation and feature pyramid network in semantic segmentation. The novelty and new contribution is insufficient.

**Response:** Many thanks for reviewing our manuscript. For your concern about innovation, we do know that there are many FPN-based models for object detection and instance segmentation. But as far as we know, the applications of FPN on semantic segmentation, especially for remote sensing images, are not as much as those on object detection and instance segmentation.

Besides, although there exist some pieces of literature which have explored the combination of attention mechanism and FPN, the attention mechanisms utilized in their models are totally different from ours. Actually, there are two types of technologies with the name of attention mechanism. One is the dot-product attention mechanism, which is designed to model long-range dependencies and enable contextual information extraction at a global scale. The other is scaling attention designed to reinforce informative features and whittle information-lacking features, while the typical examples are the squeeze-and-excitation (SE) model and the convolutional block attention module (CBAM). These two types of attention mechanisms have completely different principles and purposes. A comprehensive comparison between dot-product attention, scaling attention, as well as the simplified dot-product attention mechanism on semantic segmentation, can be seen in our previous work (DOI: https://doi.org/10.1109/TGRS.2021.3093977). The existing researches on the combination of attention mechanism and FPN are based on either dot-product attention which has expensive computing consumptions or scaling attention which is completely different from the attention used in our model. Actually, only in recent two years, the simplification of the dot-product attention mechanism just has gotten more and more focused. Therefore, although there are seemingly many 'similar' researches, the actual similarities are not many, while the main contribution of our work is to combine linear attention and FPN.

**Q2.** Also, the experiments are based on the open datasets with small sizes and well labelled samples, the superiority and fesiblity of the method should be further testified using real remote sensing image with large size. Otherwise the value of this work is limitted.

**Response:** Many thanks for your concern. In the revised version, we supplement the experiment on a very large-scale dataset, i.e., UAVid. There are totally 420 images with large resolution

(4096×2160 or 3840×2160) in the dataset where 200 of them are for training, 70 for validation, and the remaining 150 for testing. 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 quantitative results can be seen in Table 4:

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 | **77.3** | **63.4** | 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 | 67.4 | 81.2 | 63.1 | 48.4 | 69.8 | 13.9 | 63.8 |
| A2-FPN | ResNet-18 | **87.2** | **80.1** | **67.4** | 80.2 | 63.7 | 53.3 | 70.1 | **23.4** | **65.7** |

Considering the UAVid is a relatively large-scale dataset, the result strongly demonstrates the effectiveness of the proposed $A^2$-FPN.

ARTICLE TEMPLATE

# $A^2$-FPN for Semantic Segmentation of Fine-Resolution Remotely Sensed Images

Rui Li [a], Libo Wangt [a], Ce Zhang [b, c], Chenxi Duan [d] and Shunyi Zheng [a]

[a]School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China;
[b]Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK;
[c]UK Centre for Ecology & Hydrology, Library Avenue, Lancaster LA1 4AP, UK;
[d]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands

**ARTICLE HISTORY**
Compiled December 21, 2021

**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 multi-scale feature learning through attention-guided feature aggregation. Based on FPN and AAM, a novel framework named Attention Aggregation Feature Pyramid Network ($A^2$-FPN) is developed for semantic segmentation of fine-resolution remotely sensed images. Extensive experiments conducted on four datasets demonstrate the effectiveness of our $A^2$-FPN in segmentation accuracy. Code is available at https://github.com/lironui/A2-FPN.

**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). For remotely sensed images, segmentation has played critical roles in several diverse geo-information applications, including urban planning, economic assessment, land resource management, etc.

Corresponding author. Email: c.duan@utwente.nl

(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 are available. For segmentation, traditional methods usually extract vegetation indices of land cover from multi-spectral/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; Gu et al. 2020).

Meanwhile, substantial significant leaps of segmentation in remote sensing have been witnessed in recent years (Su et al. 2021; Wang et al. 2021a,b), 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 (Tong et al. 2021; Wang et al. 2021a; Zhang et al. 2021).

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 multi-scale 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 multi-scale 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 multi-scale 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. To remedy the defect of FPN, we propose an Attention Aggregation Module (AAM) based on the linear attention mechanism (Li et al. 2021b) to enhance multi-scale 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 form multi-scale layers; (2) It extracts discriminative features by extracting global context information.

## 2. Related Work

### 2.1. *Feature Pyramid Network*

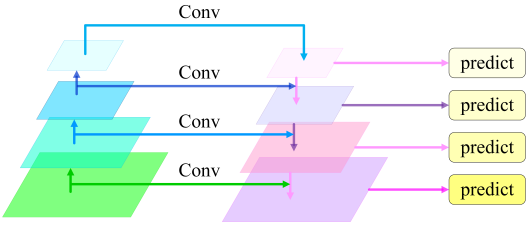


**Figure 1.** Illustration of the architecture of Feature Pyramid Network for detection.

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 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 (Shoeiby et 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. 2021b,a). For example, Sherrah (Sherrah 2016) adapted the FCN to semantically label remotely sensed images. Kampffmeyer et al. (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. (Maggiori et al. 2017) adopted an auxiliary CNN to learn how to combine features. Audebert et al. (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 et al. (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 is 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 impedes 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

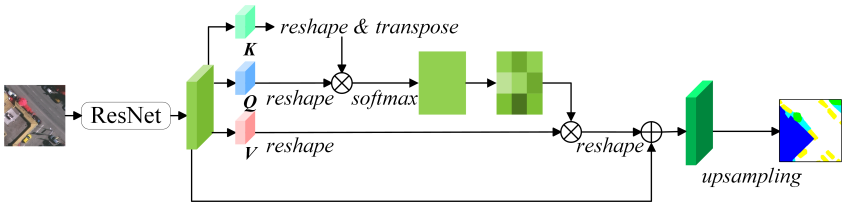


**Figure 2.** Illustration of the architecture of dot-product attention mechanism.

$A^2$-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. $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N] \in \mathbb{R}^{N \times C}$ refers to the input feature, where $N = H \times W$. First, the dot-product attention mechanism uses three projected matrices $\boldsymbol{W}_q \in \mathbb{R}^{D_x \times D_k}$, $\boldsymbol{W}_k \in \mathbb{R}^{D_x \times D_k}$, and $\boldsymbol{W}_v \in \mathbb{R}^{D_x \times D_v}$ to obtain the *query* matrix $\boldsymbol{Q}$, *key* matrix $\boldsymbol{K}$ and *value* matrix $\boldsymbol{V}$ as:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{q}} \in \mathbb{R}^{N \times D_k}, \\ \boldsymbol{K} &= \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{k}} \in \mathbb{R}^{N \times D_k}, \\ \boldsymbol{V} &= \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{v}} \in \mathbb{R}^{N \times D_v}. \end{aligned} \tag{1}$$

$\boldsymbol{Q}$ and $\boldsymbol{K}$ are identical in their shapes. To compute the similarity between the $i$-th *query* feature $\boldsymbol{q}_i^T \in \mathbb{R}^{D_k}$ and the $j$-th *key* feature $\boldsymbol{k}_j \in \mathbb{R}^{D_k}$, a normalization function $\rho$ is adopted as $\rho(\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j) \in \mathbb{R}^1$. Thereafter, similarities between all pairs of pixels are computed and taken as weights. The output is generated by aggregating all positions using weighted summation:

$$D(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \rho(\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V}. \tag{2}$$

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

$$\rho(\boldsymbol{Q}\boldsymbol{K}^T) = \text{softmax}_{row}(\boldsymbol{Q}\boldsymbol{K}^T). \tag{3}$$

where $\text{softmax}_{row}$ denotes that the softmax is operated along the row of matrix $\boldsymbol{Q}\boldsymbol{K}^T$. The global context information is captured by the $\rho(\boldsymbol{Q}\boldsymbol{K}^T)$ through the modeling of the similarities among all pairs of pixels in the input. However, as $\boldsymbol{Q} \in \mathbb{R}^{N \times D_k}$ and $\boldsymbol{K}^T \in \mathbb{R}^{D_k \times N}$, the multiplication between $\boldsymbol{Q}$ and $\boldsymbol{K}^T$ 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:

$$D(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})_i = \frac{\sum_{j=1}^{N} e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j} \boldsymbol{v}_j}{\sum_{j=1}^{N} e^{\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j}}. \tag{4}$$

Equation 4 can be generalized into any normalization function as:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{N}\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)\boldsymbol{v}_j}{\sum_{j=1}^{N}\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)}, \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)\geq 0, \tag{5}$$

$\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)$ depicts the similarity between the $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$, which can be expanded as $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)=\phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$. We can further rewrite equation 4 to equation 6 and then simplify it as equation 7:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{N}\phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)\boldsymbol{v}_j}{\sum_{j=1}^{N}\phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)}, \tag{6}$$

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\phi(\boldsymbol{q}_i)^T\sum_{j=1}^{N}\varphi(\boldsymbol{k}_j)\boldsymbol{v}_j}{\phi(\boldsymbol{q}_i)^T\sum_{j=1}^{N}\varphi(\boldsymbol{k}_j)}. \tag{7}$$

In particular, equation 5 is identical to equation 4, when $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)=e^{\boldsymbol{q}_i^T\cdot\boldsymbol{k}_j}$. The equation 7 can be represented as the vectorized form:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\phi(\boldsymbol{Q})\varphi(\boldsymbol{K})^T\boldsymbol{V}}{\phi(\boldsymbol{Q})\sum_j\varphi(\boldsymbol{K})_{i,j}^T}, \tag{8}$$

As $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)=\phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$ 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 $\varphi(\boldsymbol{K})^T$ and $\boldsymbol{V}$ first and then multiply the result and $\phi(\boldsymbol{Q})$, 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:

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

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

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

We then rewrite equation 5 into equation 11, and simplify it into equation 12:

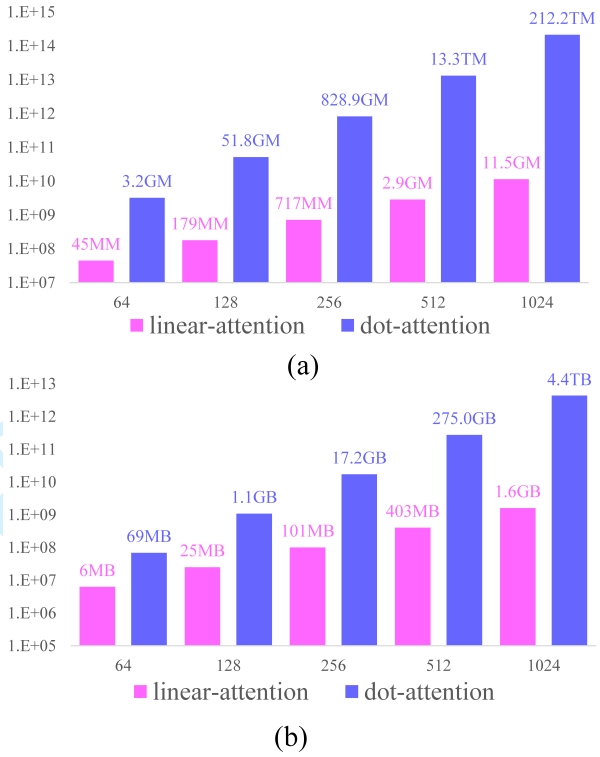


**Figure 3.** The (a) computation requirement and (b) memory requirement between the linear attention mechanism and dot-product attention mechanism under different input sizes. The calculation assumes. The calculation assumes $D = D_v = 2D_k = 64$. MM denotes 1 Mega multiply-accumulation (MACC), where 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.

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

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

The vectorized form of equation 12 is:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\sum_j \boldsymbol{V}_{i,j}+(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2})((\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2})^T\boldsymbol{V})}{N+(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2})\sum_j(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2})_{i,j}^T}. \tag{13}$$

As $\sum_{j=1}^{N}(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2})\boldsymbol{v}_j^T$ and $\sum_{j=1}^{N}(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2})$ 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 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N] \in \mathbb{R}^{N\times C}$, both the dot-attention and linear attention generate the *query* matrix $\boldsymbol{Q}$, *key* matrix $\boldsymbol{K}$ and

*value* matrix $\boldsymbol{V}$. For the dot-attention, the $N \times N$ matrix is generated by multiplying the transposed *key* matrix $\boldsymbol{K}$ and the *value* matrix $\boldsymbol{V}$, 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 $\boldsymbol{K}$ and *query* matrix $\boldsymbol{Q}$. Therefore, we can calculate the product between $\boldsymbol{K}^T$ and $\boldsymbol{V}$ first and then multiply the result and $\boldsymbol{Q}$ with only $O(dN)$ time complexity and $O(dN)$ space complexity. The concrete comparison can be seen in Figure 3.

## 3. Attention Aggregation Feature Pyramid Network

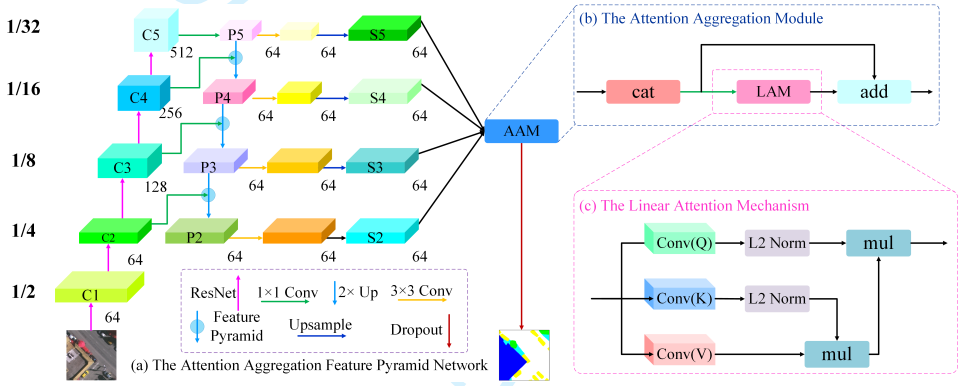


**Figure 4.** The structure of (a) the overall framework of our $A^2$-FPN, (b) the Attention Aggregation Module, and (c) the Linear Attention Mechanism (taking the attention1 as an example). The figures (e.g., 64, 128, 512) near the features indicate the number of channels..

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 4a), the top-down pathway (i.e., the second column in Figure4a), the lateral connections (i.e., the $1 \times 1$ convolutional layer between the first and second column in Figure 4a), the feature pyramid (i.e., the second and third columns in Figure 4a), and the Attention Aggregation Module (i.e., Figure 4b). We will elaborate on each component below.

### 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*

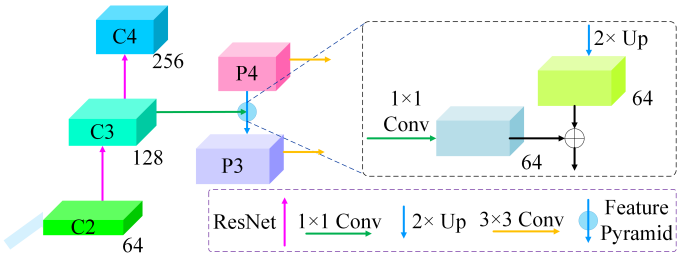


**Figure 5.** The feature pyramid in the proposed $A^2$-FPN.

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 $A^2$-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 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 multi-scale 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 multi-scale 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., S2, S3, S4, and S5) 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).

**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.

**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.

**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 $506 km^2$ captured by the Gaofen 2 satellite. Following the previous work (Li 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 remained 40% patches are reserved as the test set.

**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,

**Table 1.** The 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 |
| $A^2$-FPN | ResNet-34 | **93.0** | **95.7** | **84.7** | **90.0** | **86.9** | **90.1** | **91.0** | **82.2** |

mIoU, and F1 are computed as:

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

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

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \tag{16}$$

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:

$$IoU = \frac{TP_k}{TP_k + FP_k + FN_k}. \tag{17}$$

### 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 512×512 patches (1024×1024 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) (Fu 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 (Yu et al. 2020), and CE-Net (Gu

**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 |
| $A^2$-FPN | ResNet-34 | **93.6** | **96.9** | **87.5** | **88.4** | **95.7** | **92.4** | **91.1** | **86.1** |

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 (Yu et al. 2018), SwiftNet (Oršić 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.

To qualitatively illustrate the effectiveness of the proposed $A^2$-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 which 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 multi-scale feature maps enables our $A^2$-FPN to generate more accurate segmentation maps.

### 4.5. *Results on the ISPRS Potsdam Dataset*

To further evaluate the effectiveness of $A^2$-FPN, we carry out experiments on the ISPRS Potsdam dataset. The training and testing settings on the Potsdam dataset

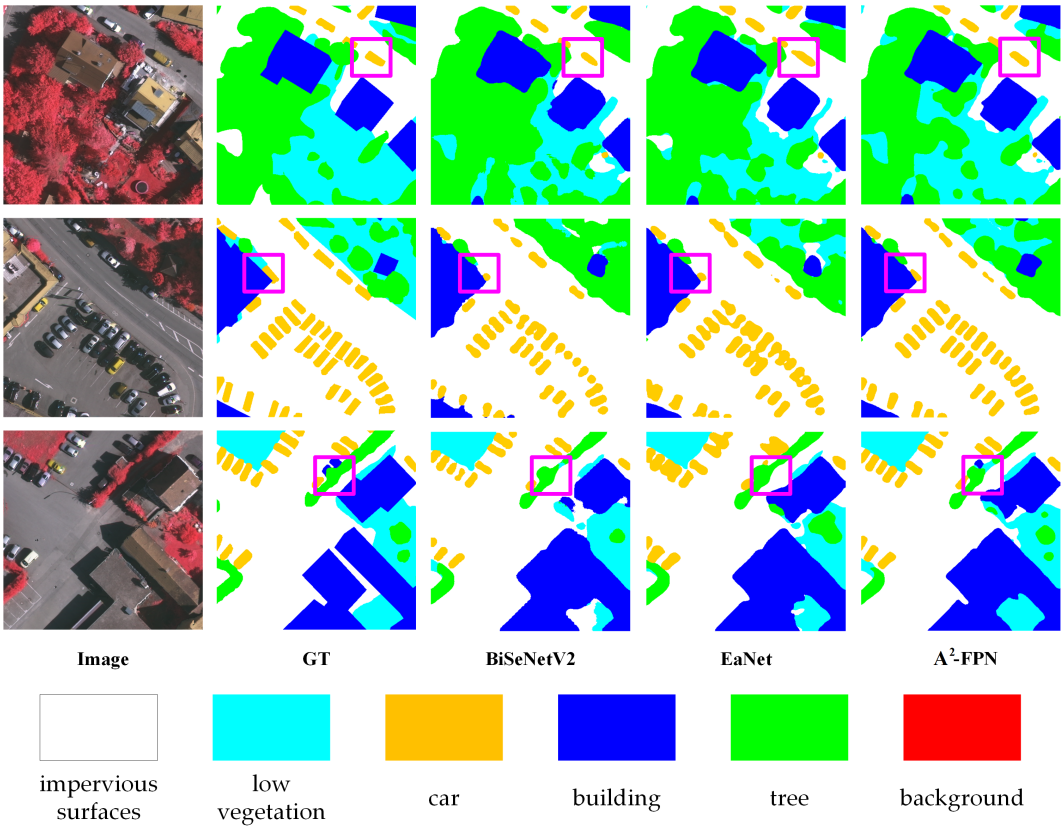


**Figure 6.** Visualization of results on the Vaihingen dataset.

are the same as the Vaihingen dataset. Numerical comparisons with comparative algorithms are listed in Table 2. The $A^2$-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 × 512 patches with the intractable regions marked by red rectangles. Our $A^2$-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 $A^2$-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 $A^2$-FPN to exploit the multi-scale 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 $A^2$-FPN. As listed in Table 3, our $A^2$-FPN holds the leading position on the vast majority of the evaluation indexes. Visualized results in Figure 8 also demonstrates the

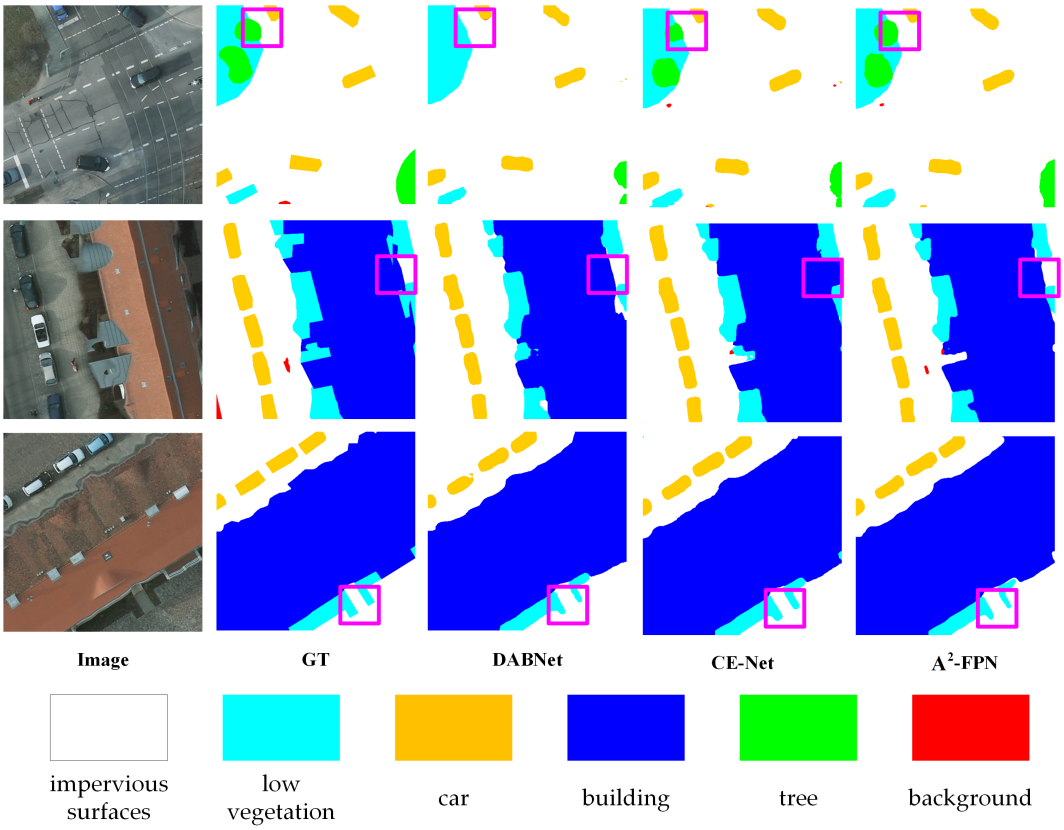


**Figure 7.** Visualization of results on the Potsdam dataset.

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 $A^2$-FPN, benefiting from the utilization of multi-scale contextual information.

### 4.7. *Results on the UAVid Dataset*

As illustrated in Table 4, the proposed $A^2$-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 $A^2$-FPN. Since the ground truth of the test set is not available now, we visualize and compare the results generated by our $A^2$-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 $A^2$-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 $A^2$-FPN correctly recognizes all moving cars.

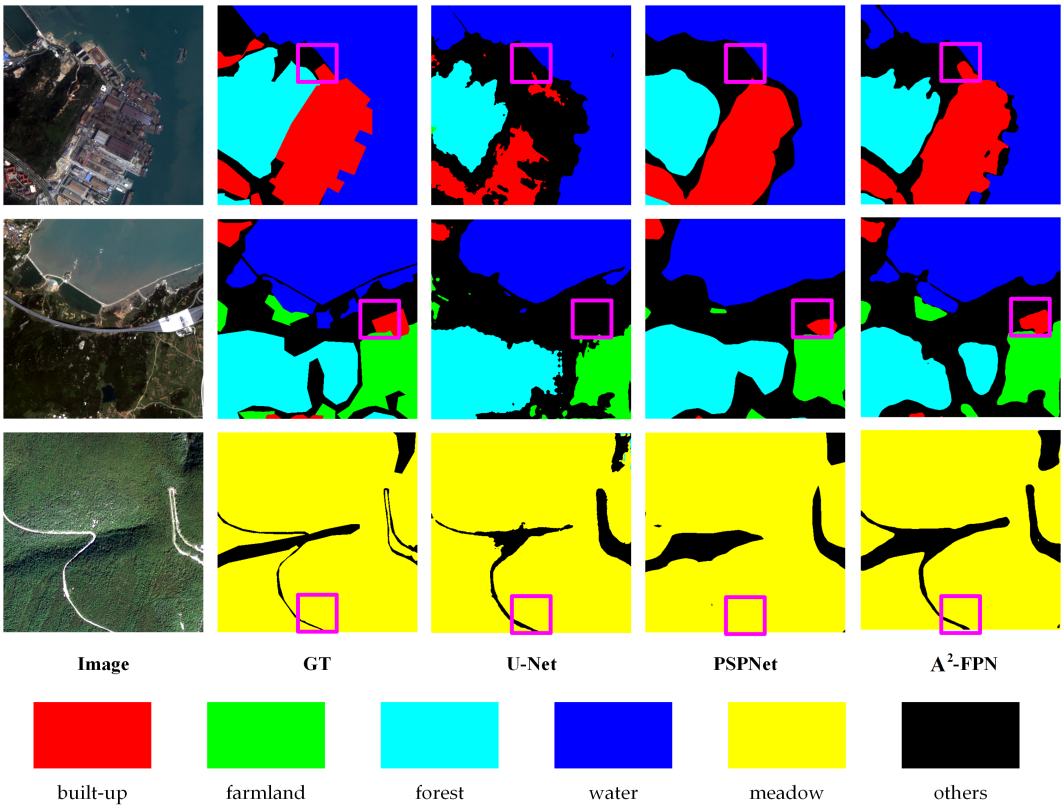


**Figure 8.** Visualization of results on the GID dataset.

## 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 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.

**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 |
| $A^2$-FPN | ResNet-34 | **86.3** | **91.0** | **92.4** | **87.9** | **96.8** | **79.9** | **89.1** | **88.3** | **80.7** |

**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 | **77.3** | **63.4** | 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 | **67.4** | **81.2** | 63.1 | 48.4 | 69.8 | 13.9 | 63.8 |
| $A^2$-FPN | ResNet-18 | **87.2** | **80.1** | **67.4** | 80.2 | 63.7 | 53.3 | **70.1** | **23.4** | **65.7** |

**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 |

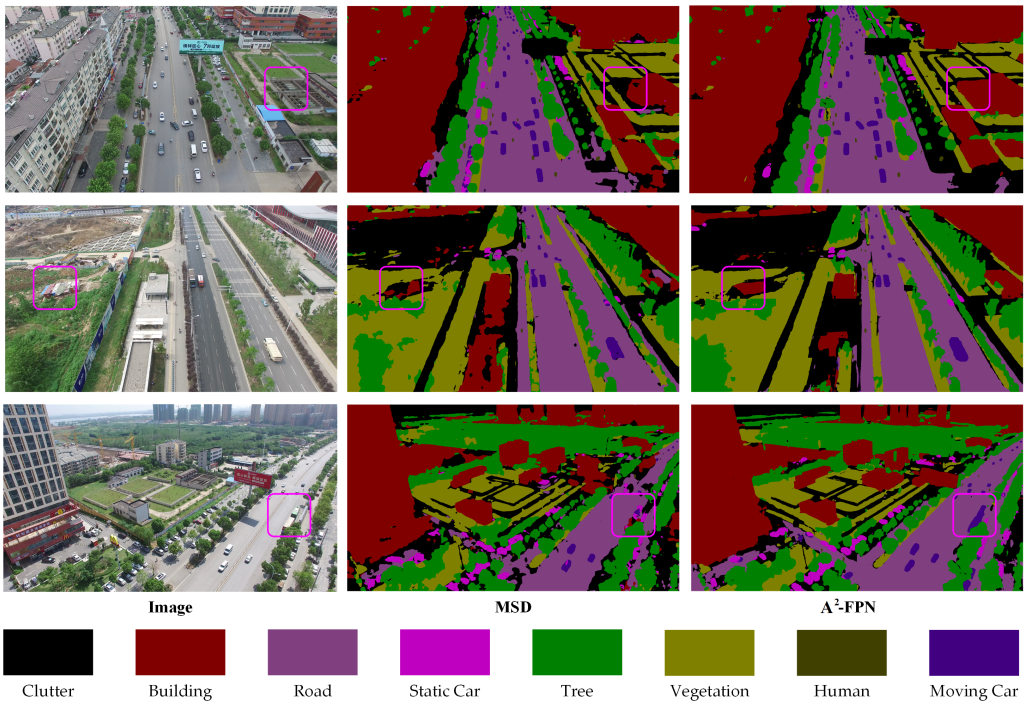


**Figure 9.** Visualization of results on the UAVid dataset.

### 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.

### 5.3. *Limitation*

Although the proposed $A^2$-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

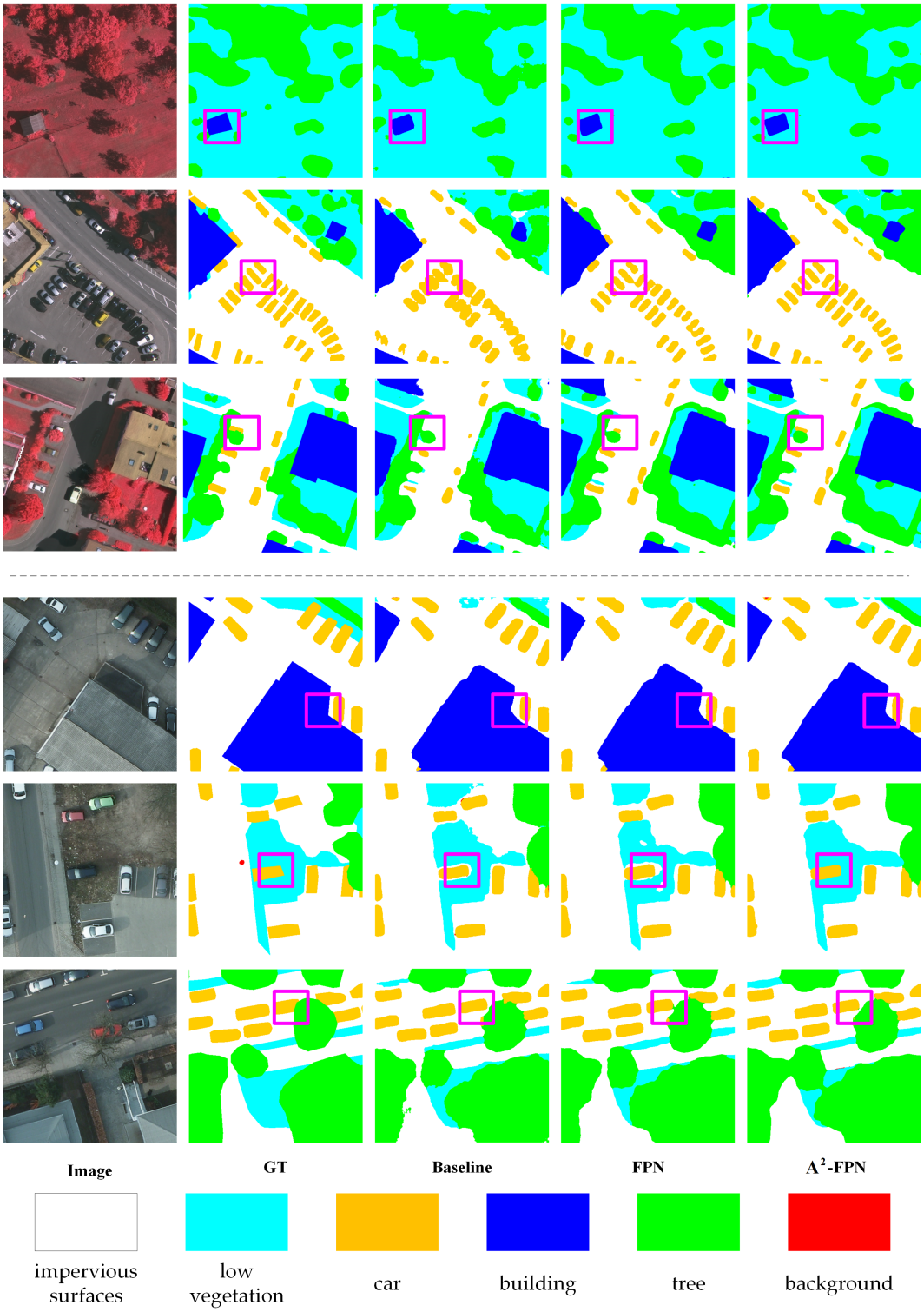


**Figure 10.** Visualization of ablation study on (top) the Vaihingen dataset and (bottom) the Potsdam dataset.

**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 $A^2$-FPN and other methods. 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.

| 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 |
| $A^2$-FPN | 22.93 | 22.27 | 107.12 | 65.44 | 16.87 | 4.60 |

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. 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.

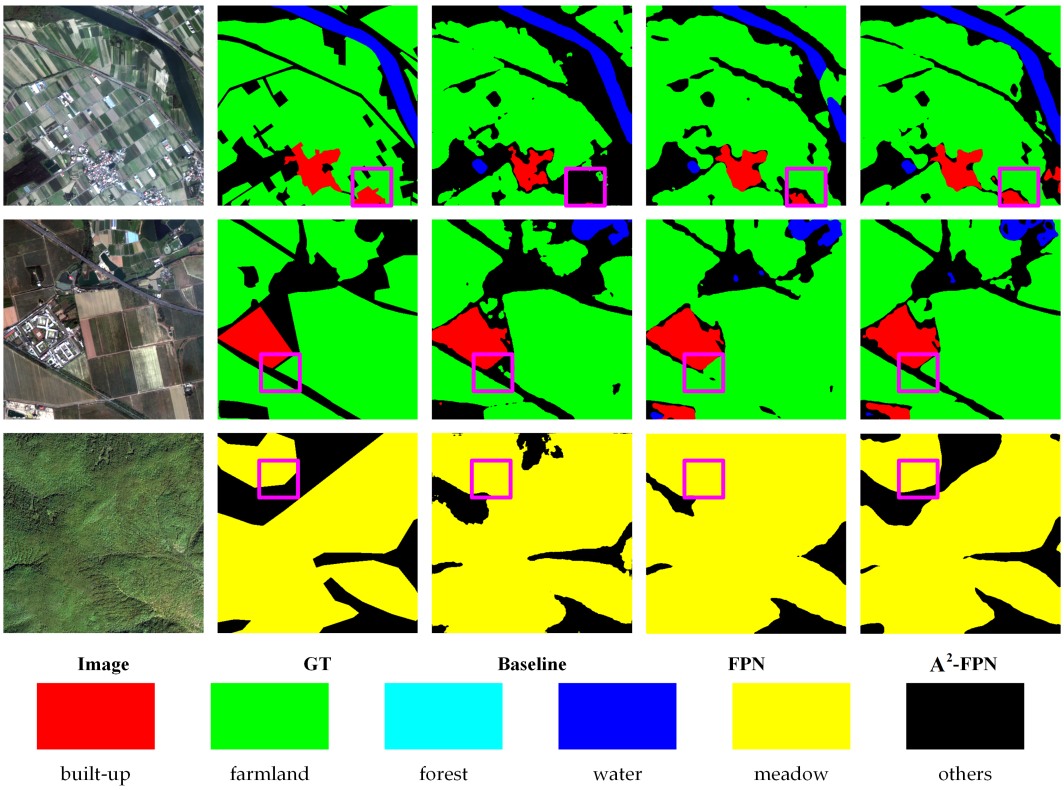


**Figure 11.** Visualization of ablation study on the GID dataset.

## Conflict of Interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

## Author's Contributions

This work was conducted in collaboration with all author. 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.

## Funding

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

## References


Audebert, Nicolas, Bertrand Le Saux, and Sébastien Lefèvre. 2018. "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 140: 20–32.

Bai, Yunkun, Guangmin Sun, Yu Li, Peifeng Ma, Gang Li, and Yuanzhi 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.

Duan, Chenxi, Jun Pan, and Rui Li. 2020. "Thick Cloud Removal of Remote Sensing Images Using Temporal Smoothness and Sparsity Regularized Tensor Optimization." *Remote Sensing* 12 (20): 3446.

Fu, Jun, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. "Dual attention network for scene segmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Gu, Xiaowei, Plamen Angelov, Ce Zhang, and Peter Atkinson. 2020. "A Semi-Supervised Deep Rule-Based Approach for Complex Satellite Sensor Image Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Gu, Zaiwang, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. 2019. "Ce-net: Context encoder network for 2d medical image segmentation." *IEEE transactions on medical imaging* 38 (10): 2281–2292.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kampffmeyer, Michael, Arnt-Borre Salberg, and Robert 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*, 1–9.

Kampffmeyer, Michael, Arnt-Børre Salberg, and Robert 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): 1758–1768.

Katharopoulos, Angelos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. "Transformers are rnns: Fast autoregressive transformers with linear attention." In *International Conference on Machine Learning*, 5156–5165. PMLR.

Kirillov, Alexander, Ross Girshick, Kaiming He, and Piotr Dollár. 2019. "Panoptic feature pyramid networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6399–6408.

Li, Gen, Inyoung Yun, Jonghyun Kim, and Joongkyu Kim. 2019. "Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation." *arXiv preprint arXiv:1907.11357* .

Li, Huapeng, Ce Zhang, Shuqing Zhang, and Peter 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.

Li, Rui, Chenxi Duan, Shunyi Zheng, Ce Zhang, and Peter M Atkinson. 2021a. "MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images." *IEEE Geoscience and Remote Sensing Letters* .

Li, Rui, Shunyi Zheng, Chenxi Duan, Jianlin Su, and Ce Zhang. 2021b. "Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* .

Li, Rui, Shunyi Zheng, Chenxi Duan, Yang Yang, and Xiqi Wang. 2020b. "Classification of hyperspectral image based on double-branch dual-attention mechanism network." *Remote Sensing* 12 (3): 582.

Li, Rui, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkin-

son. 2021c. "Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 1–13.

Li, Rui, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter 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: 84–98.

Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lyu, Ye, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. 2020. "UAVid: A semantic segmentation dataset for UAV imagery." *ISPRS journal of photogrammetry and remote sensing* 165: 108–119.

Maggiori, Emmanuel, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. 2017. "High-resolution aerial image labeling with convolutional neural networks." *IEEE Transactions on Geoscience and Remote Sensing* 55 (12): 7092–7103.

Marmanis, Dimitrios, Konrad Schindler, Jan Dirk Wegner, Silvano Galliani, Mihai Datcu, and Uwe Stilla. 2018. "Classification with an edge: Improving semantic image segmentation with boundary detection." *ISPRS Journal of Photogrammetry and Remote Sensing* 135: 158–172.

Oršić, Marin, and Siniša Šegvić. 2021. "Efficient semantic segmentation with pyramidal fusion." *Pattern Recognition* 110: 107611.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Shamsolmoali, Pourya, Masoumeh Zareapoor, Huiyu Zhou, Ruili Wang, and Jie Yang. 2020. "Road segmentation for remote sensing images using adversarial spatial pyramid networks." *IEEE Transactions on Geoscience and Remote Sensing* .

Sherrah, Jamie. 2016. "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery." *arXiv preprint arXiv:1606.02585* .

Shoeiby, Mehrdad, Ali Armin, Sadegh Aliakbarian, Saeed Anwar, and Lars Petersson. 2020. "Mosaic Super-resolution via Sequential Feature Pyramid Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 84–85.

Su, Zhongbin, Wei Li, Zheng Ma, and Rui Gao. 2021. "An improved U-Net method for the semantic segmentation of remote sensing images." *Applied Intelligence* 1–13.

Tong, Huilin, Zhijun Fang, Ziran Wei, Qingping Cai, and Yongbin Gao. 2021. "SAT-Net: a side attention network for retinal image segmentation." *Applied Intelligence* 51 (7): 5146–5156.

Tong, Xin-Yi, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. 2020. "Land-cover classification with high-resolution remote sensing images using transferable deep models." *Remote Sensing of Environment* 237: 111322.

Tu, Ying, Bin Chen, Tao Zhang, and Bing Xu. 2020. "Regional mapping of essential urban land use categories in China: A segmentation-based approach." *Remote Sensing* 12 (7): 1058.

Wambugu, Naftaly, Yiping Chen, Zhenlong Xiao, Mingqiang Wei, Saifullahi Aminu Bello, José Marcato Junior, and Jonathan Li. 2021. "A hybrid deep convolutional neural network for accurate land cover classification." *International Journal of Applied Earth Observation and Geoinformation* 103: 102515.

Wang, Libo, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. 2021a. "A Novel Transformer based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images." *arXiv preprint arXiv:2104.12137* .

Wang, Libo, Rui Li, Dongzhi Wang, Chenxi Duan, Teng Wang, and Xiaoliang 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, Yao, Kai Zhang, and Shunping 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): 8919–8931.

Yu, Changqian, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang.

2020. "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation." *arXiv preprint arXiv:2004.02147* .

Yu, Changqian, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. "Bisenet: Bilateral segmentation network for real-time semantic segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Zhang, Ce, Peter M Atkinson, Charles George, Zhaofei Wen, Mauricio Diazgranados, and France 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: 280–291.

Zhang, Ce, Paula A Harrison, Xin Pan, Huapeng Li, Isabel Sargent, and Peter M Atkinson. 2020b. "Scale Sequence Joint Deep Learning (SS-JDL) for land use and land cover classification." *Remote Sensing of Environment* 237: 111593.

Zhang, Ce, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, and Peter M Atkinson. 2019. "Joint Deep Learning for land cover and land use classification." *Remote sensing of environment* 221: 173–187.

Zhang, Xiu-Ling, Bing-Ce Du, Zhao-Ci Luo, and Kai Ma. 2021. "Lightweight and efficient asymmetric network design for real-time semantic segmentation." *Applied Intelligence* 1–16.

Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. "Pyramid scene parsing network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, Xianwei, Linxi Huan, Gui-Song Xia, and Jianya 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: 15–28.

Zhu, Xiao Xiang, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. 2017. "Deep learning in remote sensing: A comprehensive review and list of resources." *IEEE Geoscience and Remote Sensing Magazine* 5 (4): 8–36.

Zhuang, Juntang, Junlin Yang, Lin Gu, and Nicha Dvornek. 2019. "Shelfnet for fast semantic segmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

## Appendix A. More visual results

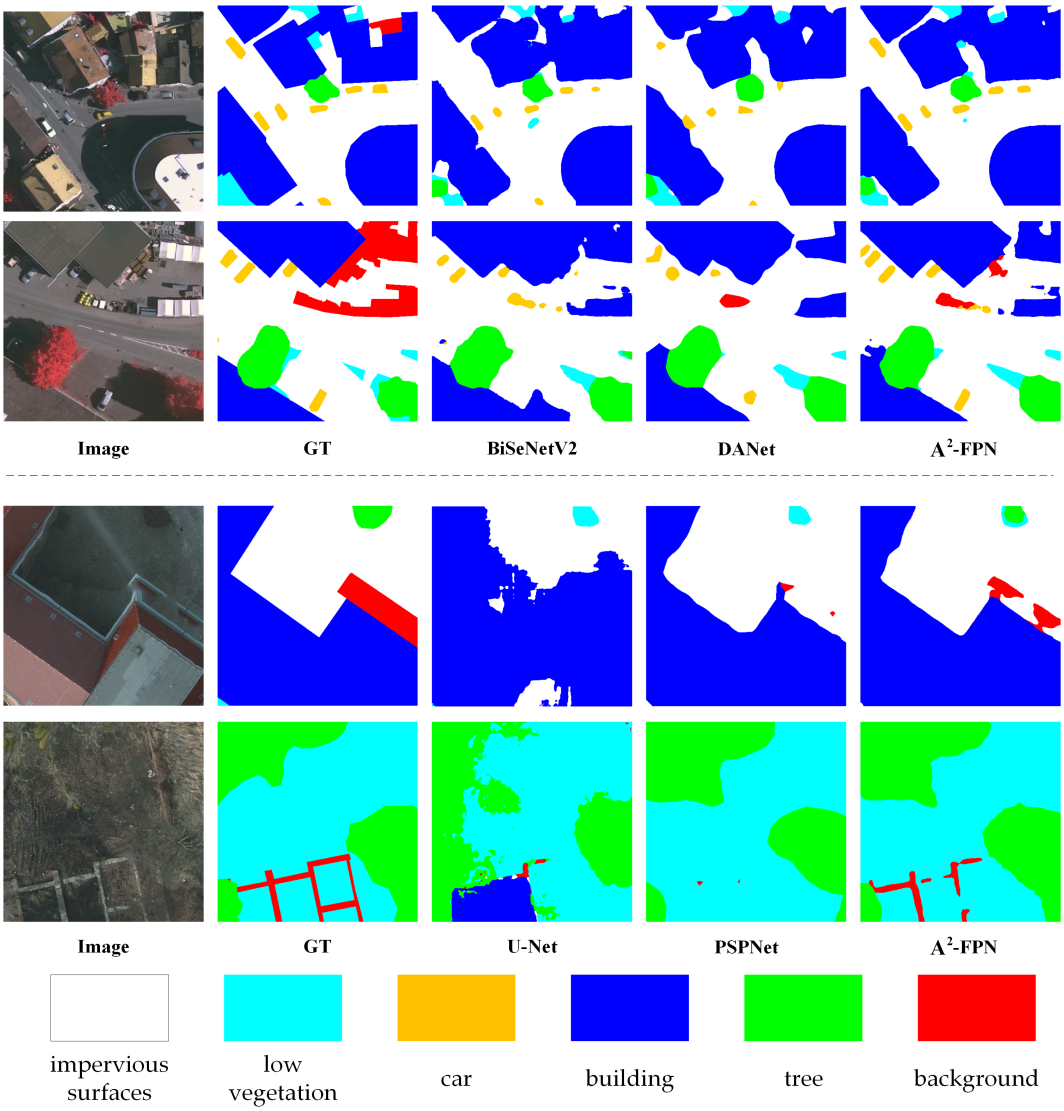


**Figure A1.** The failure cases in the (top) Vaihingen dataset and (bottom) Potsdam dataset.

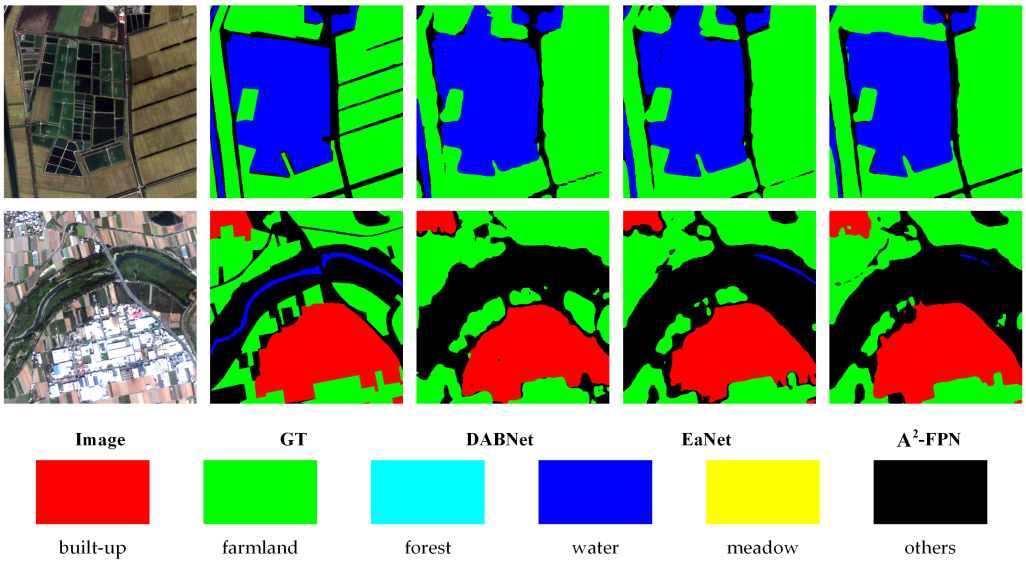


**Figure A2.** The failure cases in GID dataset.

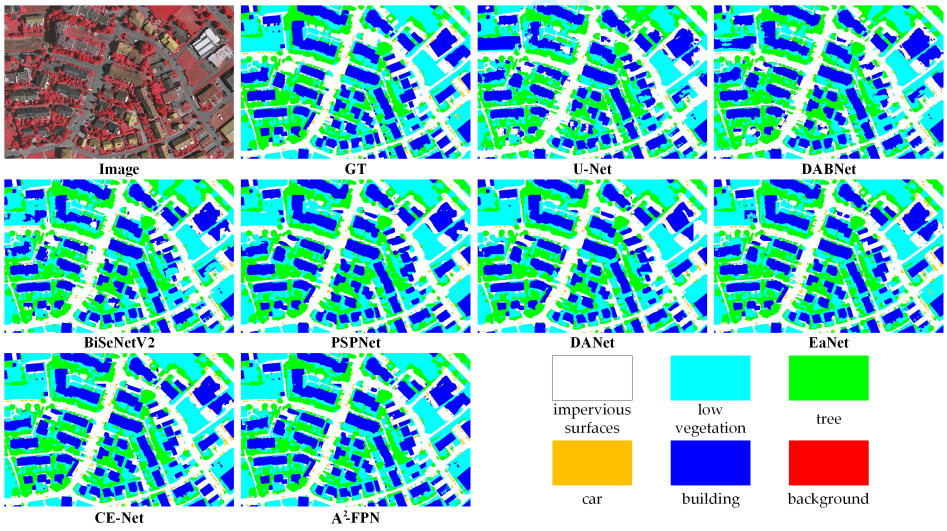


**Figure A3.** Visualization of tile-38 in the Vaihingen dataset.

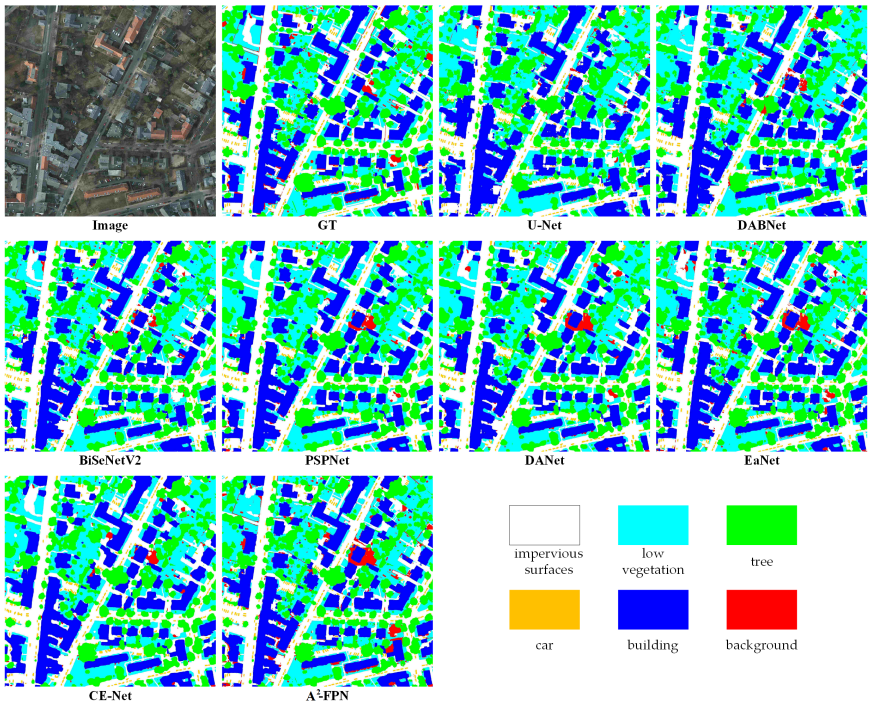


**Figure A4.** Visualization of tile-38 in the Potsdam dataset.