Efficient Hybrid Transformer: Learning Global-local Context for Urban Sence Segmentation

Libo Wang 1, Shenghui Fang 1*, Ce Zhang 2,3, Rui Li 1 and Chenxi Duan 4,5

1) School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China.

2) Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

3) UK Centre for Ecology & Hydrology, Library Avenue, Lancaster LA1 4AP, UK.

4) Faculty of Geo-Information Science and Earth Observation (Lyons et al.), University of Twente, Enschede, the Netherlands

5) The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China.

*Corresponding author.
Abstract—Semantic segmentation of fine-resolution urban scene images plays a vital role in extensive practical applications, such as land cover mapping, urban change detection, environmental protection and economic assessment. Driven by rapid developments in deep learning technologies, convolutional neural networks (CNNs) have dominated the semantic segmentation task for many years. Convolutional neural networks adopt hierarchical feature representation and have strong local context extraction. However, the local property of the convolution layer limits the network from capturing global information that is crucial for improving fine-resolution image segmentation. Recently, Transformer comprise a hot topic in the computer vision domain. Vision Transformer demonstrates the great capability of global information modelling, boosting many vision tasks, such as image classification, object detection and especially semantic segmentation. In this paper, we propose an efficient hybrid Transformer (EHT) for semantic segmentation of urban scene images. EHT takes advantage of CNNs and Transformer, learning global-local context to strengthen the feature representation. Extensive experiments demonstrate that EHT has higher efficiency with competitive accuracy compared with state-of-the-art benchmark methods. Specifically, the proposed EHT achieves a 67.0% mIoU on the UAVid test set and outperforms other lightweight models significantly. The code will be available soon.

Index Terms—Semantic Segmentation, Vision Transformer, Global-local Context, Hybrid Structure.
1. Introduction

Benefiting from the advance of sensor technology, more and more urban scene images are captured at fine spatial resolutions. Fine-resolution urban scene images, which contain rich characteristics of geo-objects and massive spatial details, play a crucial role in the field of semantic segmentation. Semantic segmentation is a pixel-level classification task, which has extensive urban applications, including land cover mapping, change detection, environment protection, road and building extraction, and other practical applications (Li et al., 2021a; Li et al., 2021b; Li et al., 2021c; Wang et al., 2021a; Wang et al., 2021b; Yang et al., 2021). Driven by the rapid development of deep learning, convolutional neural networks (CNNs) have become the mainstream approach for semantic segmentation. In comparison with the traditional machine learning methods, such as support vector machine (SVM), random forest, and conditional random field (CRF), CNNs have demonstrated stronger feature representation with the hierarchical structure and local information modelling. Representative networks include the fully convolutional network (FCN), UNet and its variants and context aggregation networks. The accuracy of CNNs, although encouraging, meets bottlenecks in global information modelling due to the local property of the convolution layer.

Recently, Transformer comprise a hot topic in the computer vision domain. Transformer is originally designed for natural language processing (NLP) tasks. The tremendous breakthrough in the natural language domain inspires researchers to explore the potential and feasibility of Transformer for vision tasks. Vision Transformer demonstrates the great capability of global information modelling, boosting many fundamental vision tasks, such as image classification,
objection detection and semantic segmentation. Vision Transformer can be split into two
categories: the pure self-attention based Transformer and hybrid Transformer. The last one
demonstrates strong discrimination capability with great generalization.

In this paper, we propose a hybrid structure, named Global-local Transformer (EHT), for
semantic segmentation of remotely sensed images. EHT utilizes multi-head self-attention to
capture global context while introduces convolution layers to extract local context, then,
aggregates the global and local context to achieve stronger feature representation. Fig. 1 illustrates
the difference between our hybrid Transformer and the standard Transformer. Fig. 2 illustrates the
difference between the global context generated by multi-head self-attention and local context
produced by the convolution layer and demonstrates the advantages of our global-local context.

In addition, we propose a cross-scale fusion module (CFM) to aggregate layer-by-layer features.
The design of our EHT is lightweight, which could further increase segmentation accuracy while
ensuring the efficiency of semantic segmentation simultaneously. A thorough benchmark
comparison was undertaken against the state-of-the-art to demonstrate the effectiveness of the
proposed EHT.
Fig. 1. Illustration of the difference between the (a) standard Transformer block and (b) hybrid Transformer block.

Fig. 2. Illustration of the difference of extracted context between local branch and global branch.
2. Efficient Hybrid Transformer

The proposed efficient hybrid Transformer is demonstrated in Fig. 3. We attach the Global-Local Transformer block on the top of the ResNet18 backbone like BottleNeck Transformer. Three cross-scale fusion modules equipped with three cross-scale connections are utilized to aggregate multi-layer features.

1) Global-local Transformer Block

The details of the proposed Global-Local Transformer block (GLTB) are shown in Fig. 4. The main module Global-Local attention block is a hybrid structure, which employs linear multi-head self-attention to capture global context while uses convolution layers to extract local context. Finally, an addition operation is applied to the global context and local context to extract the global-local context.
**Fig. 4.** The structure of the Global-local Transformer block. H and W represent the resolution of the feature map and N=H×W. D and h denote the number of channels and heads, respectively, 
\[ d = D/h. \]

**Linear multi-head self-attention:** Our previous research proposed a linear attention mechanism to replace the softmax function with the first-order approximation of the Taylor expansion. In this paper, we improved the linear attention to linear multi-head self-attention for higher efficiency and stronger sequence modelling. The specific formula derivation process is as follows.

If the normalization function is set as softmax, the \( i \)-th row of the result matrix generated by the self-attention attention can be written as:

\[
D(Q, K, V)_i = \frac{\sum_{j=1}^{N} e^{q_i^T k_j} v_j}{\sum_{j=1}^{N} e^{q_i^T k_j}},
\]

where \( v_j \) is \( j \)-th value feature. According to the Taylor expansion:
\[ e^{q_i^T k_j} \approx 1 + q_i^T k_j. \]  

(2)

To guarantee the above approximation to be nonnegative, \( q_i \) and \( k_j \) are normalized by the \( l_2 \) norm, thereby ensuring \( q_i^T k_j \geq -1 \):

\[ \text{sim}(q_i, k_j) = 1 + \left( \frac{q_i}{\|q_i\|_2} \right)^T \left( \frac{k_j}{\|k_j\|_2} \right). \]  

(3)

Thus, equation (1) can be rewritten as equation (4) and simplified as equation (5):

\[
D(Q, K, V)_i = \frac{\sum_{j=1}^N \left( 1 + \left( \frac{q_i}{\|q_i\|_2} \right)^T \left( \frac{k_j}{\|k_j\|_2} \right) \right) v_j}{\sum_{j=1}^N \left( 1 + \left( \frac{q_i}{\|q_i\|_2} \right)^T \left( \frac{k_j}{\|k_j\|_2} \right) \right)}.
\]  

(4)

\[
D(Q, K, V)_i = \frac{\sum_{j=1}^N v_j + \left( \frac{q_i}{\|q_i\|_2} \right)^T \sum_{j=1}^N \left( \frac{k_j}{\|k_j\|_2} \right) v_j^T}{N + \left( \frac{q_i}{\|q_i\|_2} \right)^T \sum_{j=1}^N \left( \frac{k_j}{\|k_j\|_2} \right)}.
\]  

(5)

Equation (5) can be turned into a vectorized form:

\[
D(Q, K, V) = \frac{\sum_j V_{i,j} + \left( \frac{Q}{\|Q\|_2} \right)^T \left( \frac{K}{\|K\|_2} \right)^T V}{N + \left( \frac{Q}{\|Q\|_2} \right)^T \sum_j \left( \frac{K}{\|K\|_2} \right)_{i,j}}.
\]  

(6)

Since \( \sum_{j=1}^N \left( \frac{k_j}{\|k_j\|_2} \right) v_j^T \) and \( \sum_{j=1}^N \left( \frac{k_j}{\|k_j\|_2} \right) \) can be calculated and reused for each query, the time and memory complexity of the attention based on equation (6) is \( O(dN) \). More detailed information on the proposed attention mechanism, as well as its validity and efficiency, are in our previous work.

Please note that a trainable scale factor is deployed on the output of linear multi-head self-attention for stable context aggregation.

**Locality-enhanced module**: We adopt two parallel convolution layers followed by a Batch Normalization operation to extract local context.

The generated global-local context is further proceeded by a depth-wise convolution, a batch
normalization operation and a 1x1 convolution to enhance generalization ability.

2) Cross-scale fusion module

Cross-scale connection: We adopt two parallel convolution layers followed by a Batch Normalization operation to extract local context. The details of the cross-scale connection are shown in Fig. 5. The scale factor of the upsampled operation is 2. L denotes the number of repetitions. Three cross-scale connections are corresponding to three cross-scale fusion modules. The dilated rates of Atrous convolution of three cross-scale connections are 6, 12 and 18, respectively.

![Fig. 5. The cross-scale connection.](image)

Weighted feature fusion: The three generated by cross-scale connections are aggregated with the corresponding residual feature and the upsampled global-local semantic feature by a weighted element-wise sum operation to strengthen the generalization capability. The equation is as follows:

$$\mathcal{G}_i = \begin{cases} \mathcal{G}_i, & \text{if } i = 4 \\ \alpha_1 \cdot f_{\mu}(\mathcal{G}_i \mathcal{F}_{i+1}) + \alpha_2 \cdot f_5(R_i) + \alpha_3 \cdot \mathcal{CSF}_i, & \text{if } i \in \{1,2,3\} \end{cases}$$ (7)

where $f_{\mu}$ is a resize operation to unify the shape of $\mathcal{G}_i \mathcal{F}_{i+1}$ and $\mathcal{CSF}_i$, while $f_5$ is a standard 1x1 convolution to unify the channels of $R_i$ and $\mathcal{CSF}_i$. $\alpha_1, \alpha_2, \alpha_3$ denote the weight coefficients and always satisfy $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

We further aggregate $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \mathcal{G}_4$ as the input of the Head for final
3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Experimental setting

3.1.1 Implementation Details

All models in the experiments were implemented with PyTorch framework on a single NVIDIA GTX 2080ti GPU. For fast convergence, we deployed the AdamW optimizer to train all models in the experiments. The base learning rate was set to 1e-4 and the weight decay value was 0.05. The early stopping technique was applied to control the training time for preventing overfitting. Cross-entropy loss with online hard example mining was chosen as the loss function. Random flip and Random Brightness are used for data augmentation.

3.1.2 Models for comparison

The comparative benchmark methods selected included the contextual information aggregation methods designed initially for natural images, such as pyramid scene parsing network (PSPNet) (Zhao et al., 2017) and dual attention network (DANet) (Fu et al., 2019), the multi-scale feature aggregation models proposed for remote sensing images, including multi-stage attention ResU-Net (MAResU-Net) (Li et al., 2021b) and edge-aware neural network (EaNet) (Zheng et al., 2020), as well as lightweight networks developed for efficient semantic segmentation, including depth-wise asymmetric bottleneck network (DABNet) (Li et al., 2019), efficient residual factorized convNet (ERFNet) (Romera et al., 2017), bilateral segmentation network V1 (BiSeNetV1) (Yu et al., 2018) and V2 (BiSeNetV2) (Yu et al., 2020), fast attention network (FANet) (Hu et al., 2020),
ShelfNet (Zhuang et al., 2019) and SwiftNet (Oršić and Šegvić, 2021). In the inference stage, we also utilized the data augmentation operation including random rotation and horizontal as well as vertical flipping which is also known as test-time augmentation (TTA). Besides, ablation studies were conducted with the following model design:

**Baseline**: The baseline can be constructed by the ResNet-18 backbone and layer-by-layer feature fusion like UNet.

**Baseline + CFM**: We replace the layer-by-layer feature fusion with our cross-scale fusion module to construct a simple variant. The variant is utilized to demonstrate the effectiveness of the cross-scale fusion module.

**Baseline + CFM + GLTB**: We inserted the Global-Local Transformer block into Baseline + CFM to generate the entire EHT, which could demonstrate the effectiveness of the proposed Global-Local Transformer block.

### 3.2 Experiments I: results on the UAVid dataset

As a fine-resolution Unmanned Aerial Vehicle (UAV) semantic segmentation dataset, the UAVid dataset is focusing on urban street scenes with two resolutions 3840 × 2160 and 4096 × 2160. UAVid is a challenging benchmark since the large resolution of images, large-scale variation, and complexities in the scenes. To be specific, there are 420 images in the dataset where 200 are for training, 70 for validation, and the remaining 150 for testing.

In our experiments, the batch size is set to 2. All image tiles were cropped into 1152 × 2048 px patches to fit the image resolution and output stride of 32.
3.2.1 Ablation study

To evaluate the performance of the CFM and GLTB separately, we conduct ablation experiments. The results are illustrated in TABLE 1.

**Ablation study for the cross-scale fusion module**: With the employment of CFM, the average mIoU increases by 2.84%.

**Ablation study for global-local transformer block**: the deployment of GLTB produces an improvement of mIoU by 3.84%, demonstrating the effectiveness of GLTB for semantic labelling of remotely sensed images.

**TABLE 1.** Ablation study of each component of EHT on the UAVid test set.

| Dataset  | Method           | mIoU (%) |
|----------|------------------|----------|
| Baseline | Baseline + CFM   | 63.19    |
| UAVid    | Baseline + CFM + GLTB (EHT) | 67.03    |
3.2.2 Comparison with lightweight networks

TABLE 2. Quantitative comparison results on the UAVid test set with the lightweight networks. FPS is tested with a input size of 1024×1024 on the Nvidia 2080Ti GPU.

| Method  | Backbone | Clutter | Building | Road | Tree | Vegetation | Moving Car | Static Car | Human | mIoU (%) | Parameter (M) | FPS |
|---------|----------|---------|----------|------|------|------------|------------|------------|-------|----------|---------------|-----|
| BiSeNet | ResNet18 | 64.7    | 85.7     | 61.1 | 78.3 | 77.3       | 48.6       | 63.4       | 17.5  | 61.5     | -              | -   |
| SwiftNet| ResNet18 | 64.1    | 85.3     | 61.5 | 78.3 | 76.4       | 51.1       | 62.1       | 15.7  | 61.1     | -              | -   |
| MSD     | -        | 57.0    | 79.8     | 74.0 | 74.5 | 55.9       | 62.9       | 32.1       | 19.7  | 57.0     | -              | -   |
| ABCNet  | ResNet18 | 67.4    | 86.4     | 81.2 | 79.9 | 63.1       | 69.8       | 48.4       | 13.9  | 63.8     | 13.7           | 72.2|
| EHT     | ResNet18 | 68.8    | 88.2     | 81.6 | 80.1 | 63.9       | 73.0       | 56.6       | 24.2  | 67.0     | 12.6           | 83.4|

More results of lightweight models are on the way.

3.3 Experiments II: results on the Vaihingen dataset

The Vaihingen dataset consists of 33 very fine spatial resolution TOP image tiles at an average size of 2494×2064 pixels. Each TOP image tile has three multispectral bands (Near Infrared, Red, Green) as well as the digital surface model (DSM) and the normalized digital surface model (NDSM) with a 9 cm ground sampling distance (GSD). Only TOP image tiles were used in our experiments without DSM. The dataset involves five foreground classes (impervious surface, building, low vegetation, tree, car) and one background class (clutter). Following the recommendation by Liu et al., 16 image tiles were selected as the training set and the remaining 17 image tiles as the original Vaihingen test set.

In our experiments, the image tiles are cropped into 1024×1024 px patches, and the batch size is set to 4.
### TABLE 3. Quantitative comparison results on the Vaihingen test set with the lightweight networks.

| Method       | Backbone | Imp. surf. | Building | Low veg. | Tree   | Car    | Mean F1 | OA (%) | mIoU (%) |
|--------------|----------|------------|----------|----------|--------|--------|---------|--------|----------|
| DABNet       | -        | 87.8       | 88.8     | 74.3     | 84.9   | 60.2   | 79.2    | 84.3   | 70.2     |
| ERFNet       | -        | 88.5       | 90.2     | 76.4     | 85.8   | 53.6   | 78.9    | 85.8   | 69.1     |
| BiSeNetV1    | ResNet18 | 89.1       | 91.3     | 80.9     | 86.9   | 73.1   | 84.3    | 87.1   | 75.8     |
| PSPNet       | ResNet18 | 89.0       | 93.2     | 81.5     | 87.7   | 43.9   | 79.0    | 87.7   | 68.6     |
| BiSeNetV2    | -        | 89.9       | 91.9     | 82.0     | 88.3   | 71.4   | 84.7    | 88.0   | 75.5     |
| DANet        | ResNet18 | 90.0       | 93.9     | 82.2     | 87.3   | 44.5   | 79.6    | 88.2   | 69.4     |
| FANet        | ResNet18 | 90.7       | 93.8     | 82.6     | 88.6   | 71.6   | 85.4    | 88.9   | 75.6     |
| EaNet        | ResNet18 | 91.7       | 94.5     | 83.1     | 89.2   | 80.0   | 87.7    | 89.7   | 78.7     |
| ShelfNet     | ResNet18 | 91.8       | 94.6     | 83.8     | 89.3   | 77.9   | 87.5    | 89.8   | 78.3     |
| MAResU-Net   | ResNet18 | 92.0       | 95.0     | 83.7     | 89.3   | 78.3   | 87.7    | 90.1   | 78.6     |
| SwiftNet(Oršić and Šegvić, 2021) | ResNet18 | 92.2       | 94.8     | 84.1     | 89.3   | 81.2   | 88.3    | 90.2   | 79.6     |
| ABCNet       | ResNet18 | **92.7**   | 95.2     | 84.5     | 89.7   | 85.3   | 89.5    | 90.7   | 81.3     |
| EHT          | ResNet18 | 92.6       | **95.3** | **84.6** | **90.3** | **87.4** | **90.0** | **90.8** | **82.1** |
**TABLE 4.** Quantitative comparison results on the Vaihingen test set with the state-of-the-art networks.

| Method       | Backbone | Imp. surf. | Building | Low veg. | Tree | Car | Mean F1 | OA (%) | mIoU (%) |
|--------------|----------|------------|----------|----------|------|-----|---------|--------|----------|
| DeepLabV3+   | ResNet101| 92.4       | 95.2     | 84.3     | 89.5 | 86.5| 89.6   | 90.6   | 81.5     |
| PSPNet       | ResNet101| 92.8       | 95.5     | 84.5     | 89.9 | **88.6**| 90.3   | 90.9     | **82.6** |
| DANet        | ResNet101| 91.6       | 95.0     | 83.3     | 88.9 | 87.2 | 89.2   | 90.4   | 81.3     |
| EaNet        | ResNet101| 93.4       | 96.2     | 85.6     | 90.5 | 88.3 | **90.8**| 91.2   | -        |
| DDCM-Net     | ResNet50 | 92.7       | 95.3     | 83.3     | 89.4 | 88.3 | 89.8   | 90.4   | -        |
| HUSTW5       | ResNet101| 93.3       | 96.1     | **86.4** | **90.8**| 74.6| 88.2   | 91.6   | -        |
| CASIA2       | ResNet101| 93.2       | 96.0     | 84.7     | 89.9 | 86.7 | 90.1   | 91.1   | -        |
| V-FuseNet    | FuseNet  | 91.0       | 94.4     | 84.5     | 89.9 | 86.3 | 89.2   | 90.0   | -        |
| DLR_9        | -        | 92.4       | 95.2     | 83.9     | 89.9 | 81.2 | 88.5   | 90.3   | -        |
| ABCNet       | ResNet18 | 92.7       | 95.2     | 84.5     | 89.7 | 85.3 | 89.5   | 90.7   | 81.3     |
| EHT          | ResNet18 | 92.6       | 95.3     | 84.6     | 90.3 | 87.4 | 90.0   | 90.8   | 82.1     |

### 3.4 Experiments III: results on the Potsdam dataset

The Potsdam dataset contains 38 very fine resolution TOP image tiles (GSD 5cm) at a size of 6000 × 6000 pixels and involves the same category information as the Vaihingen dataset. Four multispectral bands (Red, Green, Blue, and Near Infrared), as well as a DSM and NDSM, are provided in the dataset. The 24 image tiles were chosen as the training set, and the remaining tiles were selected as the original Potsdam test set. We utilized only TOP image tiles with three
multispectral bands (Near Infrared, Red, Green) in the experiments. Notably, the experimental settings are the same as that of Viahingen.

**TABLE 5** Quantitative comparison results on the Potsdam test set with the lightweight networks.

| Method     | Backbone | Imp. surf. | Building | Low veg. | Tree | Car | Mean F1 | OA (%) | mIoU (%) |
|------------|----------|------------|----------|----------|------|-----|---------|--------|----------|
| ERFNet     | -        | 88.7       | 93.0     | 81.1     | 75.8 | 90.5| 85.8    | 84.5   | 76.2     |
| DABNet     | -        | 89.9       | 93.2     | 83.6     | 82.3 | 92.6| 88.3    | 86.7   | 79.6     |
| PSPNet     | ResNet18 | 89.1       | 94.5     | 84.0     | 85.8 | 76.6| 86.0    | 87.2   | 75.9     |
| BiSeNetV1  | ResNet18 | 90.2       | 94.6     | 85.5     | 86.2 | 92.7| 89.8    | 88.2   | 81.7     |
| BiSeNetV2  | -        | 91.3       | 94.3     | 85.0     | 85.2 | 94.1| 90.0    | 88.2   | 82.3     |
| EaNet      | ResNet18 | 92.0       | 95.7     | 84.3     | 85.7 | 95.1| 90.6    | 88.7   | 83.4     |
| MAResU-Net | ResNet18 | 91.4       | 95.6     | 85.8     | 86.6 | 93.3| 90.5    | 89.0   | 83.9     |
| DANet      | ResNet18 | 91.0       | 95.6     | 86.1     | 87.6 | 84.3| 88.9    | 89.1   | 80.3     |
| SwiftNet   | ResNet18 | 91.8       | 95.9     | 85.7     | 86.8 | 94.5| 91.0    | 89.3   | 83.8     |
| FANet      | ResNet18 | 92.0       | 96.1     | 86.0     | 87.8 | 94.5| 91.3    | 89.8   | 84.2     |
| ShelfNet   | ResNet18 | 92.5       | 95.8     | 86.6     | 87.1 | 94.6| 91.3    | 89.9   | 84.4     |
| EHT        | ResNet18 | **93.2**   | **96.5** | **87.6** | **88.5** | **96.2** | **92.4** | **91.0** | **86.2** |
### Conclusion

In this paper, we propose a novel efficient hybrid Transformer for efficient semantic segmentation of fine-resolution remote sensing images. Extensive experiments on the ISPRS Vaihingen and Potsdam datasets as well as the UAVid dataset demonstrate the effectiveness and efficiency of the proposed GLHT.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal
relationships that could have appeared to influence the work reported in this paper.

**References**

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146-3154.

Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K., Sclaroff, S., 2020. Real-time semantic segmentation with fast attention. IEEE Robotics and Automation Letters 6, 263-270.

Li, G., Yun, I., Kim, J., Kim, J., 2019. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arXiv preprint arXiv:1907.11357.

Li, R., Duan, C., Zheng, S., Zhang, C., Atkinson, P.M., 2021a. MACU-Net for semantic segmentation of fine-resolution remotely sensed images. IEEE Geoscience and Remote Sensing Letters.

Li, R., Zheng, S., Duan, C., Su, J., Zhang, C., 2021b. Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters.

Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2021c. Multitattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing.

Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J., Elith, J., 2018. A comparison of resampling methods for remote sensing classification and accuracy assessment. Remote Sensing of Environment 208, 145-153.

Oršić, M., Šegvić, S., 2021. Efficient semantic segmentation with pyramidal fusion. Pattern Recognition 110, 107611.

Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R., 2017. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems 19, 263-272.

Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., Fang, S., 2021a. A Novel Transformer based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. arXiv preprint arXiv:2104.12137.

Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X., 2021b. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. Remote Sensing 13, 3065.

Yang, M.Y., Kumaar, S., Lyu, Y., Nex, F., 2021. Real-time Semantic Segmentation with Context Aggregation Network. ISPRS Journal of Photogrammetry and Remote Sensing 178, 124-134.

Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N., 2020. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. arXiv preprint arXiv:2004.02147.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation, Proceedings of the European conference on computer vision (ECCV), pp. 325-341.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881-2890.

Zheng, X., Huan, L., Xia, G.-S., Gong, J., 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.

Zhuang, J., Yang, J., Gu, L., Dvornek, N., 2019. Shelfnet for fast semantic segmentation, Proceedings of
the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0-0.