mathematics

MDPI

Article

Semantic Segmentation of UAV Images Based on Transformer Framework with Context Information

Satyawant Kumar, Abhishek Kumar and Dong-Gyu Lee *

Department of Artificial Intelligence, Kyungpook National University, Daegu 41566, Republic of Korea
* Correspondence: dglee@knu.ac.kr

Abstract: With the advances in Unmanned Aerial Vehicles (UAVs) technology, aerial images with huge variations in the appearance of objects and complex backgrounds have opened a new direction of work for researchers. The task of semantic segmentation becomes more challenging when capturing inherent features in the global and local context for UAV images. In this paper, we proposed a transformer-based encoder-decoder architecture to address this issue for the precise segmentation of UAV images. The inherent feature representation of the UAV images is exploited in the encoder network using a self-attention-based transformer framework to capture long-range global contextual information. A Token Spatial Information Fusion (TSIF) module is proposed to take advantage of a convolution mechanism that can capture local details. It fuses the local contextual details about the neighboring pixels with the encoder network and makes semantically rich feature representations. We proposed a decoder network that processes the output of the encoder network for the final semantic level prediction of each pixel. We demonstrate the effectiveness of this architecture on UAVid and Urban Drone datasets, where we achieved mIoU of 61.93% and 73.65%, respectively.

Keywords: semantic segmentation; UAV street scene images; transformer; global and local context

MSC: 68U10; 68T01; 68T07; 68T45

Citation: Kumar, S.; Kumar, A.; Lee, D.-G. Semantic Segmentation of UAV Images Based on Transformer Framework with Context Information. *Mathematics* **2022**, *10*, 4735. https://doi.org/10.3390/math10244735

Academic Editor: Teng Li

Received: 31 October 2022
Accepted: 11 December 2022
Published: 13 December 2022

1. Introduction

Semantic segmentation of images, i.e., assigning labels to each pixel, has been widely studied in the field of computer vision. Majorly its applications are in medical imaging [1,2], autonomous driving cars [3–5], satellite image segmentation [6], and point cloud scene segmentation [7]. Due to the rapid development of smart technologies, the application of Unmanned Aerial Vehicles (UAVs), i.e., drones, has significantly increased. UAV devices can capture photos even in remote areas where it is difficult for a human. Aerial images captured by UAV devices contain rich information [8] that can be utilized for different tasks such as traffic density estimation [9], building extraction [10], and flooded area monitoring [11].

Recently, semantic segmentation of UAV images has shown decent performance in a variety of applications [12]. Dutta et al. [13] designed a segmentation framework for UAV images for the early detection of disease in cruciferous crops. Song et al. [14] designed an image fusion-based segmentation network for sunflower lodging recognition using UAV images. Lobo Torres et al. [15] tested Fully Convolutional Network (FCN) based approaches for conducting endangered tree species analyses using UAV urban scene areas. The semantic segmentation of UAV street scene images has opened a new direction of work for researchers [16]. Pixel-level dense prediction results of objects in street scene images are beneficial for land use monitoring. Convolution-based architectures are widely exploited for segmentation in computer vision. The majority of the segmentation models [17–20] adopt FCN-based [21] encoder network and design different decoder architectures. However, the FCN-based encoder backbone results in coarse predictions [22], and the receptive

field becomes constant after a certain threshold limit [4]. Due to local receptive fields in Convolutional Neural Networks (CNNs), it maintains local context properly but fails to capture global contexts [23]. For UAV scene segmentation, Lyu et al. [16] constructed a segmentation dataset and designed a segmentation model, which showed satisfactory results for UAV image segmentation. Girisha et al. [24] adopted Long Short-Term Memory (LSTM) or optical flow modules for UAV urban scene segmentation.

UAV images normally contain very complex backgrounds, with lots of variations in object appearance and scale, which poses a significant challenge for semantic segmentation [12]. Even though the existing segmentation frameworks show satisfactory results, including the ones specially designed for UAV scene segmentation, their feature extractor struggles to capture the inherent features of aerial images. Segmenting minority classes, especially small objects such as humans, which have the least number of pixels in the whole image, becomes challenging. For dense prediction tasks on complex UAV images, i.e., semantic segmentation, global and local context information is an important component [22,25]. Considering both, the global and local contextual details can precisely give better performance [8,12,26]. Pixels belonging to the same class but far apart from each other represent the global context, whereas the neighboring pixels of the same class represent the local context. Global and local context modeling is shown in Figure 1. Yellow and red arrows in the figure represent the global and local contextual information modeling.

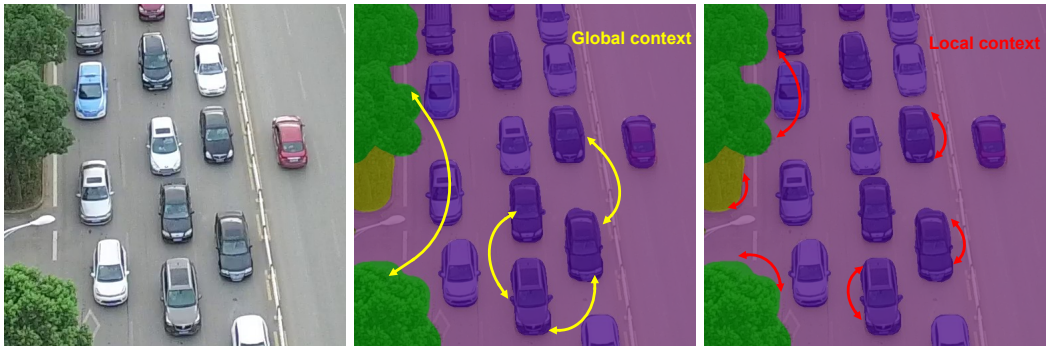

Figure 1. Illustrating global and local context information for UAV image semantic segmentation. (**a**) Input image. (**b**) Global context modeling. The yellow arrows represent the global contextual information. (**c**) Local context modeling. The red arrows represent the local contextual information.

The focus of this work is to achieve precise semantic segmentation of UAV street scene images. Inspired by the transformer-based design paradigms for computer vision tasks [23,27], an encoder-decoder framework is proposed to address such issues in this work. It incorporates a self-attention-based encoder network, which captures long-range information in the UAV images via maintaining global receptive fields. This allows the network to maintain global contextual details. For modeling low-level contextual details, using the advantage of CNNs in capturing local information, a convolution-based element is introduced in the encoder network, i.e., the Token Spatial Information Fusion (TSIF) module. Capturing local context information lets the network maintain the proper shape and size of the objects in the segmentation results. The output of the powerful self-attention-based encoder network contains semantically very rich information, both global and local contextual details. A decoder network is proposed for final pixel-level predictions, which processes this output information from the encoder network.

In short, the main contributions of this paper are as follows:

1. We propose a new encoder-decoder-based framework for semantic segmentation of UAV street scene images;
2. The encoder network is designed using the transformer-based module to capture global and local context information;

3. A decoder network is constructed to upsample the output of the encoder network to the original size image without losing the semantically rich information;
4. Experiments on UAVid and Urban Drone Dataset (UDD-6) datasets demonstrate that the proposed method outperforms the state-of-the-art methods.

The rest of the paper is presented as follows: Section 2 discusses the related work. In Section 3, we briefly discuss the proposed design. We describe the experimental results in Section 4. Based on our findings, we discuss the advantage of the proposed method in Section 5. At last, in Section 6, we conclude the work of this paper.

2. Related Work

2.1. Semantic Segmentation

In the direction of semantic image segmentation, a Fully Convolutional Network (FCN) [21] is the first work that used Convolutional Neural Networks (CNNs). Following this work, many different methods have been designed to improve its segmentation results. DeconvNet [28] used a progressive upsampling strategy to upsample the coarse output of the FCN. SegNet [29] upsampled a coarse feature map by using max pool indices transferred from the FCN encoder. U-Net [1] used skip connections and introduced contracting and expanding paths. It fused the information from the encoder network to the decoder network during each upsampling stage. GCN [30] and DeepLAbV3+ [17] adopted a dilation rate greater than 1 to increase the strength of the receptive field. DANet [18], CCNet [31], and ISNet [3] used attention-based operations to capture long-range global context details. Li et al. [32] used the Multitask Low-rank Affinity Pursuit (MLAP) method to annotate and segment images. Li et al. [33] used automatic labeling of infra-red images for UAVs using a probabilistic framework.

2.2. Transformer

Transformer-based [34] frameworks have transformed the entire field of Natural Language Processing (NLP). With its huge success in NLP, researchers have also started exploring its application for computer vision tasks. Vision Transformer (ViT) [23] is the first work in this direction, and it used a fully transformer-based design for image classification. After that, many researchers [35–38] followed its work to improve the classification accuracy. DeiT [35] introduced a teacher-student framework to provide easy learning for the network.

ViT split an input image into a sequence of tokens, and the sequence is processed by a set of stacked transformer encoder blocks. It uses a self-attention-based mechanism to maintain the global receptive field, which CNN-based architectures lack [4,23]. After its introduction, many architectures were designed to solve dense prediction tasks such as image segmentation [4,39] and object detection [40,41]. SETR [4] used ViT as an encoder and developed different variants of decoder designs. On top of a ViT-based encoder, Segmenter [42] designed a mask-based decoder network. PVT [27], DPT [43], and Segformer [44] adopted different design choices to generate multi-scale feature representations like CNNs. DPT generated hierarchical feature representation by assembling tokens from different stages of the ViT. PVT and Segformer adopted a framework to progressively shrink the pyramid to output varying scale features. Yuan et al. [45], Xie et al. [44], Wu et al. [46], and Chu et al. [37] adopted different convolution-based variants to fuse positional information.

2.3. Semantic Segmentation of Aerial Images

Due to an increased application of UAV devices, recent semantic segmentation of aerial images [8,25] has given a new research opportunity to researchers. UNetFormer [25] used a ResNet-based encoder module and designed a decoder network using a transformer framework to model image-level and semantic-level details. MFNet [8] employed a framework to maintain low-level details and inter-class discriminative characteristics. UAVFormer [26] used aggregation window-based self-attention modules to model complex features in the UAV scenes. ABCNet [22], CANet [47], and BANet [48] used a dual path approach to capture long-range and fine-grained information. Liu et al. [49] designed a lightweight

attention network, which uses spatial and channel attention modules. Iqbal et al. [50] followed a weakly-supervised domain adaption framework for the segmentation of aerial and satellite images. It handles cross-domain discrimination issues between aerial and satellite imagery. Yeung et al. [6] adopted steerable-filter-based and transfer-learning-based paradigms for segmenting satellite images. It addressed the issue of the least availability of labeled data for satellite scenes. Gebrehiwot et al. [51] segmented flooded regions in UAV images, where the flooded water is differentiated from buildings, vegetation, and roads. Ichim et al. [11] used a decision fusion-based strategy to segment the flooded areas and vegetation in UAV scenes. Zhang et al. [52] combined a series of residual U-Net modules for segmenting plants in UAV images. Dutta et al. [13] designed a framework to semantically segment unhealthy leaves in aerial images. ICENet [53] used a positional and channel-wise fusion of attentive features for segmenting ice in rivers. Gevaert et al. [54] used the semantic classification method to manage infrastructural development in informal settlement regions.

3. Proposed Method

3.1. Encoder Network

The encoder network is constructed using a self-attention-based transformer [23] framework. Overall, it consists of an image tokenization module, a linear projection module, transformer encoder modules, and a Token Spatial Information Fusion (TSIF) module. Figure 2a describes an overview of the encoder network. It uses an image as input and splits it into a sequence of tokens using the image tokenization module.

The image tokenization module transforms the two-dimensional input data, i.e., $x \in I^{H \times W \times 3}$ into a sequence of two-dimensional tokens, i.e., $x_t \in I^{S \times (t^2 \cdot 3)}$. (H, W) and (t, t) represent the input image size and the token size, respectively. $S = \frac{HW}{t^2}$ represents the number of tokens coming out from the input image. The linear projection module projects the sequence to a projection dimension, i.e., $x_t \in I^{S \times P}$. P represents the projection dimension. Here, the values of t and P are set to 16 and 1024, respectively.

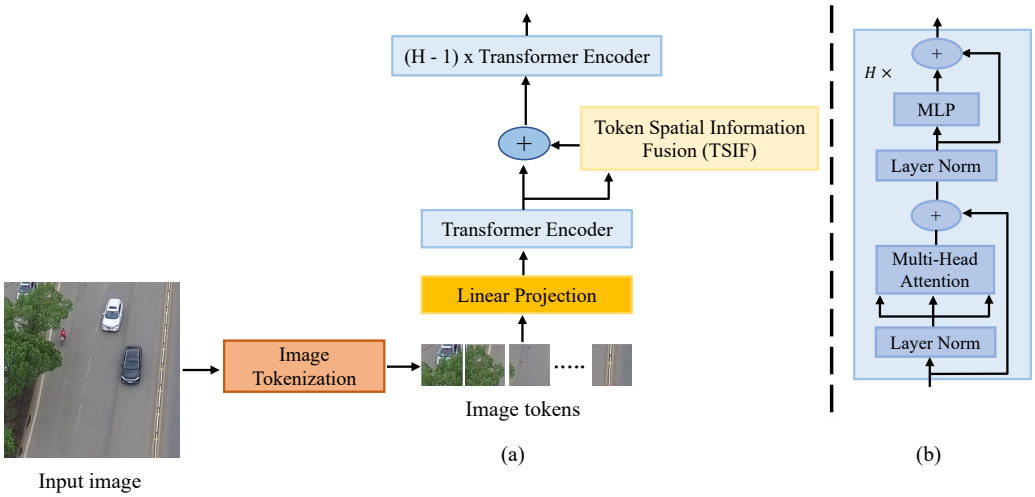

Figure 2. (**a**) Overview of an encoder network design. It consists of linear projection, transformer encoder, and Token Spatial Information Fusion (TSIF) modules. (**b**) Transformer encoder module.

The output of the linear projection module is processed using H number of transformer encoder modules. Each transformer encoder module consists of a Multi-head Self-attention (MHA) block and a Multi-Layer Perceptron (MLP) block. A Layer Normalization (LN) is applied before each block, and a residual connection after each block is applied. Figure 2b demonstrates the overall architecture of the transformer encoder module. Here, H is set to 24.

The collection of multiple Self-Attention (SA) operations makes an MHA block. The input sequence $x_t = I^{S \times P}$ is projected into query (Q), key (K), and value (V):

$$Q \in I^{S \times P}, K \in I^{S \times P}, V \in I^{S \times P}, \tag{1}$$

then the SA operation is performed using Q, K and V vectors as follows:

$$SelfAttention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{C}})V, \tag{2}$$

feature space of Q, K, and V are split into k times, and the MHA operation is performed in parallel. Here, k denotes the number of heads and is set to 16. The output of each head is concatenated together and projected to the projection dimension. The self-attention-based transformer encoder module gives the ability to incorporate a global receptive field at each stage, and hence supports the network to capture global contextual information [23]. The MLP block consists of two linear layers with a non-linearity function in between them. It performs feature expansion and reduction operations using the layers. The first layer expands the feature space by a factor of e. The second layer restores the expanded dimension back to the original. Here, e represents the feature expansion/reduction ratio. The output of the MLP module is as follows:

$$MLP(x) = (\nabla(xW_1 + b_1))W_2 + b_2, \tag{3}$$

where $W_1 \in I^{P \times L}$ and $W_2 \in I^{L \times P}$ corresponds to the weight of the first and second layers, respectively. P is the projection dimension and L is the expanded dimension. Here, L is set to 4096. The $b_1 \in I^L$ and $b_2 \in I^P$ are the biases corresponding to the first and second layers, respectively. $\nabla(.)$ is the non-linearity function. The overall output of an individual transformer encoder is as follows:

$$y = LN(x + MHA(x)), \tag{4}$$

$$z = LN(y + MLP(y)). \tag{5}$$

The Token Spatial Information Fusion (TSIF) module helps to capture low-level spatial context details about neighboring pixels in image tokens. Since the receptive field of CNNs are local in nature, it can extract local information around the neighboring pixels well [45]. Hence, convolution keeps the ability to capture local context details for dense prediction tasks such as semantic segmentation [44]. TSIF operation can be represented as follows:

$$p' = \psi(Reshape(x)), \tag{6}$$

$$TSIF_{out} = Flatten(p') + x, \tag{7}$$

where ψ is the $R \times R$ convolution operation and it is set to 3 in this work. ψ can also be replaced with other convolution variants as well for fusing the local fine-grained details. Using this design paradigm of the TSIF module also increases the representation power of features.

3.2. Decoder Network

The decoder network consists of four upsampling stages. It takes the output feature map of the powerful transformer-based encoder as input for final segmentation. It uses a gradual upsampling policy to generate the original size feature map as the input image. Before feeding the output of the encoder network to the decoder network, the feature maps are reshaped into a $3D$ representation as follows:

$$G(x_{in}) = P \times \frac{H}{16} \times \frac{W}{16}, \tag{8}$$

where G is the reshape operation, which takes the output feature of the encoder network, i.e., x_{in} as the input. P is the projection dimension and $\frac{H}{16} \times \frac{W}{16}$ represents the output feature resolution of the encoder network. This reshaped $3D$ feature is fed to the decoder network. Figure 3 represents an overview of the decoder network. Each stage consists of bilinear upsampling with a scale factor of 2, followed by a 3×3 convolution operation and a Gaussian Error Linear Units (GELU) activation function. N_c represents the feature channels at each stage and is set to 256. The output feature maps of the last stage are of the same resolution as the original image, i.e., (H, W). The channel dimension of these features are projected to the number of semantic categories, i.e., N_{cls} using a 1×1 convolution operation.

The output feature of the encoder network contains semantically rich information in terms of both global and local context details. This design paradigm of the decoder network generates the segmentation output without losing those details.

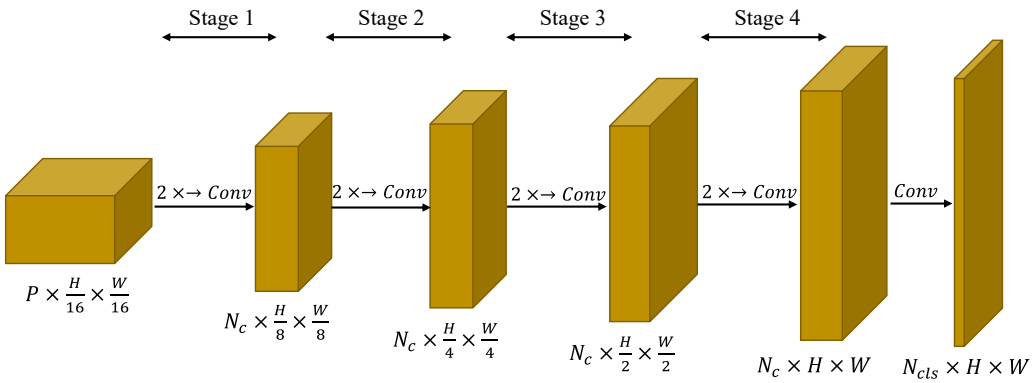

Figure 3. Overview of a decoder network. It consists of four upsampling stages.

4. Experiments

4.1. Dataset Overview

The performance of the proposed method was evaluated on two publicly available Unmanned Aerial Vehicles (UAVs) semantic segmentation datasets; UAVid [16] and Urban Drone Dataset (UDD-6) [55]. The UAVid dataset contains 42 sequences of urban street scene images. Each sequence consists of 10 RGB images of 3840×2160 or 4096×2160 size. Twenty sequences are for training, seven sequences are for validation, and the remaining are for testing. It consists of eight semantic categories; building, road, tree, vegetation, static car, moving car, humans, and clutter. The UDD-6 dataset is collected in multiple cities. It contains 141 RGB images that consist of training and validation sets of 106 and 35 images, respectively. Each image has 3840×2160 or 4096×2160 or 4000×3000 size. It contains six semantic labels; facade, road, vegetation, vehicle, roof, and others.

4.2. Implementation Details

Since it is difficult and computationally expensive to use high-resolution UAV images for training, we split the original image into patches of 512×512 resolution without overlap. The original image size is not divisible by 512, and there will be some extra image regions that will not fit in 512×512 size. For those extra regions, we extract 512×512 size patch with the required overlap. We merge the predictions on each patch into the original size image for evaluation of the metrics. Image attributes are normalized using mean and standard deviation of $(0.4914, 0.4824, 0.4467)$ and $(0.2471, 0.2436, 0.2616)$, respectively. Random horizontal flipping is used for augmentation. Complete experiments are conducted on NVIDIA RTX A6000 GPU, using the PyTorch library. Step learning rate schedule is used with an initial learning rate set to 1×10^{-3} and Stochastic Gradient Descent (SGD) is utilized as an optimizer with a momentum of 0.9 and weight decay of 1×10^{-4} on both the datasets. The network is trained for 120 and 180 epochs on UAVid and UDD-6 datasets respectively, using a batch size of 4. Cross-entropy loss is minimized during training.

4.3. Results on UAVid Dataset

In this subsection, we present the test results using the UAVid dataset. It is an UAV urban scene segmentation dataset, and due to many scale variations, it is very challenging to get high segmentation accuracy. We selected the Dilation Net [56], U-Net [1], MSD [16], ERFNet [57], BiSeNetV2 [58], Fast-SCNN [59], ShelfNet [60], SETR-PUP [4], and Segmenter [42] for comparison. Per class Intersection over Union (IoU) score, Mean IoU (mIoU), and Overall Accuracy (OA) are used as performance metrics. Appendix A.1 describes the evaluation protocol for these metrics. The quantitative results compared with competitive methods are mentioned in Table 1. The highest value of each metric in the table is marked in bold. The proposed framework shows mIoU and OA of 61.93% and 84.49%, respectively. It outperforms other methods at least by 2.25% in mIoU. Per class IoU also shows competitive performance with other methods.

ERFNet and BiSeNetV2 show mIoU of 59.28% and 59.68%, respectively, which is comparable to the performance of the proposed framework with a mIoU of 61.93%. However, the proposed method outperforms Dilation Net, U-Net, MSD, Fast-SCNN, ShelfNet, SETR-PUP, and Segmenter with a decent margin of 13.35%, 3.51%, 4.94%, 15.99%, 14.9%, 7.94%, and 3.18%, respectively, in mIoU. In per-class IoU, the proposed method outperforms Segmenter in five out of eight categories. Humans are the minority class in the whole dataset and also the smallest object to segment. Hence, it is very challenging to segment humans. For the human class, the proposed method shows decent performance, with a mIoU of 22.94%. The proposed method lags ERFNet by 0.2% in OA, but it outperforms by 2.68% in mIoU. It outperforms U-Net, BiSeNetV2, and SETR-PUP by 1.06%, 3.39%, and 4.63%, respectively, in OA.

Table 1. Quantitative comparison of the UAVid test dataset result with competitive methods.

Methods	Class IoU (%)								mIoU (%)	O.A (%)
	Building	Tree	Clutter	Road	Low Veg	Static Car	Moving Car	Human		
Dilation Net [56]	80.70	73.80	45.40	65.10	45.50	24.50	53.60	0.00	48.58	-
U-Net [1]	82.94	77.27	61.80	75.15	62.03	29.98	59.59	18.62	58.42	83.43
MSD [16]	79.80	74.50	57.00	74.00	55.90	32.10	62.90	19.70	56.99	-
ERFNet [57]	**85.58**	**77.87**	**64.50**	77.34	**62.21**	46.13	60.64	0.00	59.28	**84.69**
BiSeNetV2 [58]	81.62	75.97	61.18	77.11	61.30	38.51	**66.36**	15.40	59.68	81.10
Fast-SCNN [59]	75.70	71.50	44.20	61.60	43.40	19.50	51.60	0.00	45.94	-
ShelfNet [60]	76.90	73.20	44.10	61.40	43.40	21.00	52.60	3.60	47.03	-
SETR-PUP [4]	77.77	73.51	55.64	71.62	54.15	26.88	55.89	16.49	53.99	79.86
Segmenter [42]	84.40	76.10	64.20	**79.80**	57.60	34.50	59.20	14.20	58.75	-
Proposed Method	84.67	77.47	63.93	78.44	61.31	**47.66**	59.02	**22.94**	**61.93**	84.49

Qualitative results of the proposed method using the UAVid validation and test datasets are depicted in Figures 4 and 5, respectively. We can see that the proposed approach generates smooth segmentation of UAV images. It captures the global and local contextual information well. Cars are well segmented while maintaining their proper shape and size. In the third and fifth images of Figure 5, small objects such as humans are well segmented.

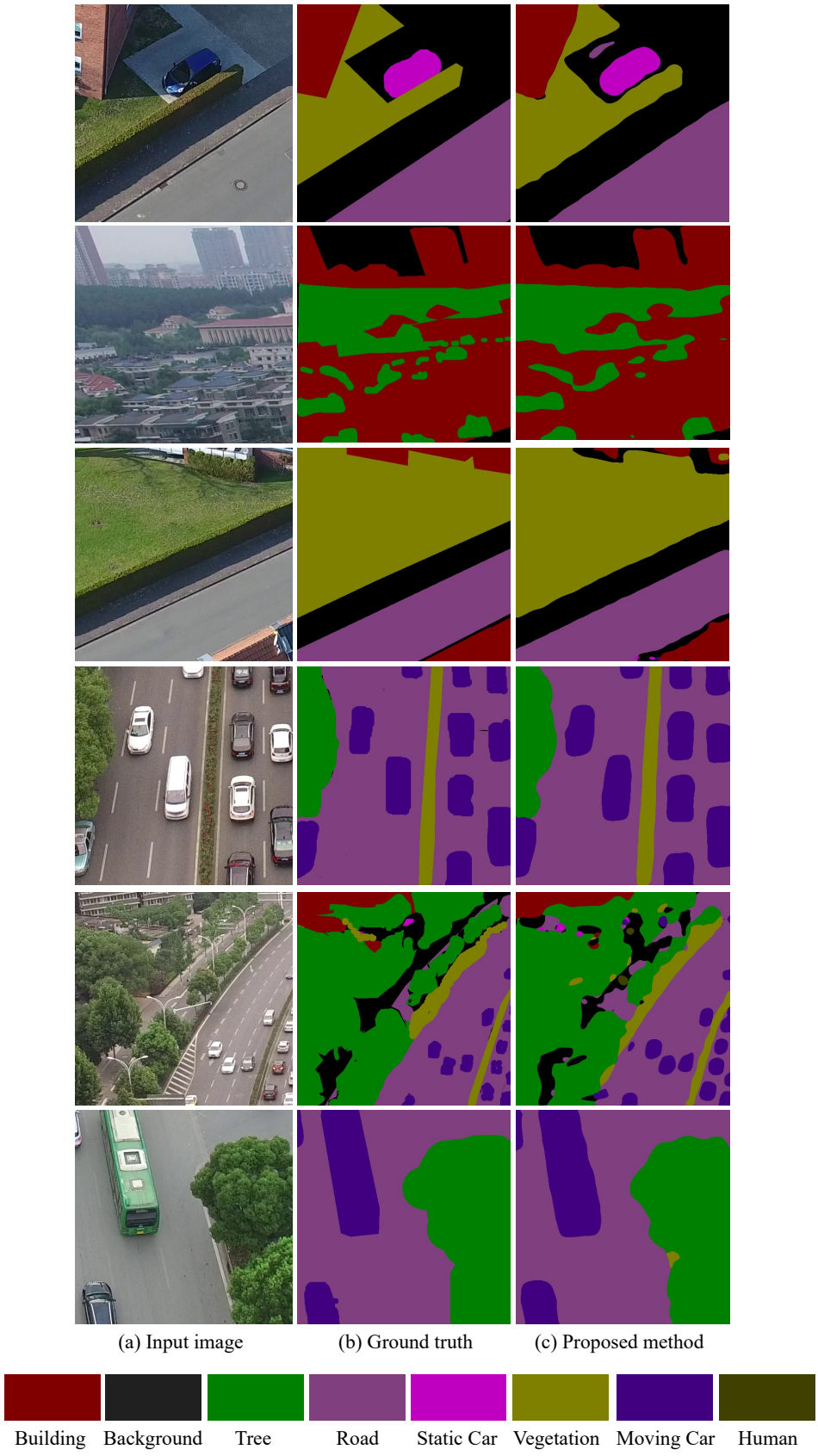

Figure 4. Qualitative prediction results using the UAVid validation dataset.

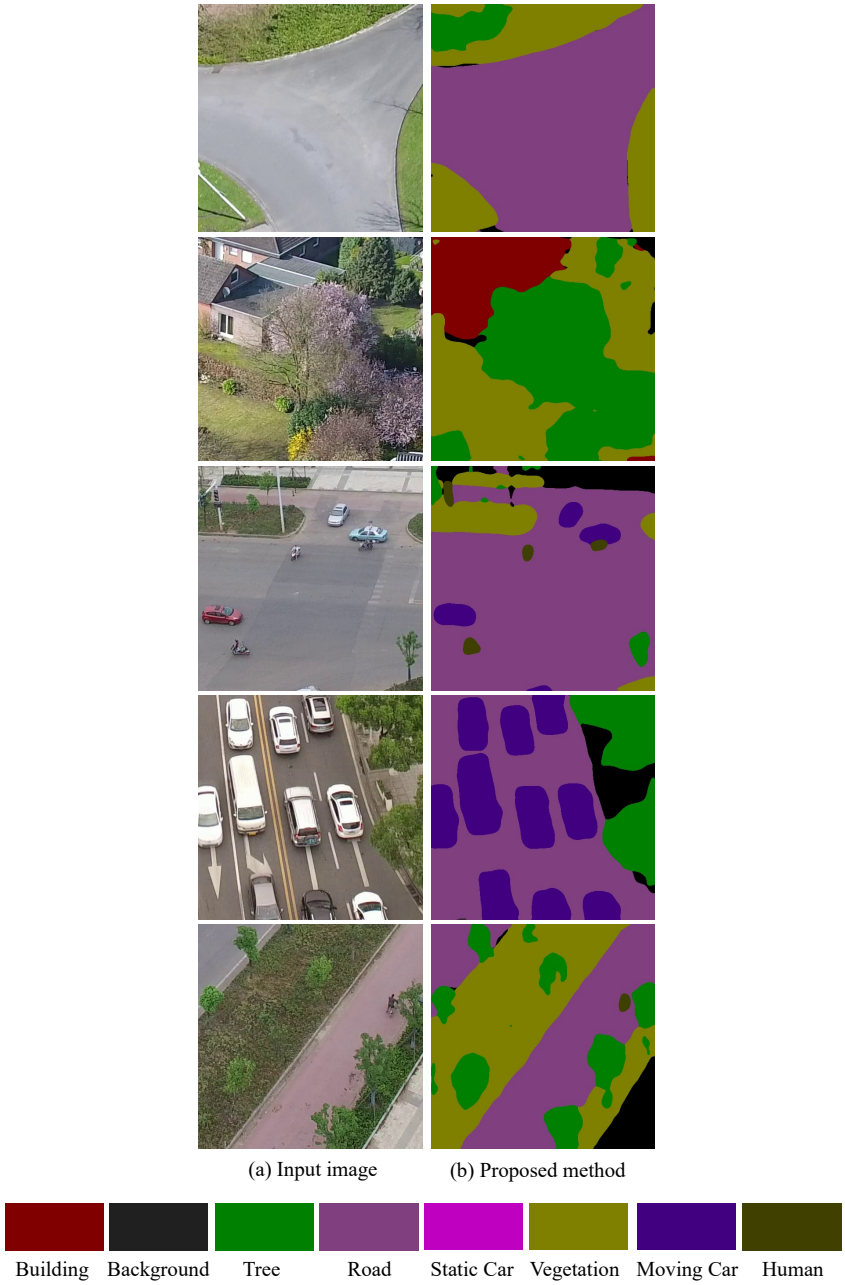

Figure 5. Qualitative prediction results using the UAVid test dataset.

4.4. Results on UDD-6 Dataset

In this subsection, we present the validation results using the UDD-6 dataset. FCN-8s [21], U-Net [1], GCN [30], ENet [61], ERFNet [57], BiSeNetV2 [58], DeepLab V3+ [17], and SETR-PUP [4] are selected for comparison. Per class IoU, mIoU, mean F1 score, and OA are used as performance metrics. Appendix A.1 describes the evaluation protocol for these metrics. The quantitative results compared with the competitive methods are mentioned in Table 2. The highest value of each metric in the table is marked in bold. The proposed method shows mIoU, mean F1 score, and OA of 73.65%, 84.40%, and 86.98% respectively. It outperforms other methods in most metrics by a significant margin.

Performance of ERFNet and DeepLab V3+ with mIoU of 73.34% and 73.18%, respectively, are comparable with the proposed method with mIoU of 73.65%. Similarly, their OA of 86.91% and 86.90%, respectively, are comparable with the proposed method, with an OA of 86.98%. However, the proposed method outperforms FCN-8s, U-Net, GCN, ENet, BiSeNetV2, and SETR-PUP by 1.88%, 2.81%, 1.07%, 4.21%, 2.43%, and 7.04%, respectively,

in mIoU. It also outperforms them in OA by 1.43%, 1.71%, 0.44%, 1.78%, 0.88%, and 3.73%, respectively. The mean F1 score of the proposed method is comparable with the ERFNet. However, it outperforms FCN-8s, U-Net, BiSeNetV2, and SETR-PUP by 1.26%, 1.96%, 1.75%, and 5.15%, respectively.

Qualitative results of the proposed method using the UDD-6 validation dataset are depicted in Figures 6 and 7. It preserves the local and global context details well and produces a fine segmentation output. Red and yellow boxes in Figure 7 represent the local and global contextual details, respectively, in an image. SETR-PUP captures global contexts well but it struggles to capture local details.

Table 2. Quantitative comparison of UDD-6 validation dataset result with competitive methods.

Method	Class IoU (%)						mIoU (%)	Mean F1 (%)	O.A (%)
	Other	Façade	Road	Veg	Vehicle	Roof			
FCN-8s [21]	88.38	67.70	66.12	66.22	83.53	58.65	71.77	83.14	85.55
U-Net [1]	89.34	67.31	65.18	62.56	82.54	58.08	70.84	82.44	85.27
GCN [30]	-	-	-	-	-	-	72.58	-	86.54
ENet [61]	-	-	-	-	-	-	69.44	-	85.20
ERFNet [57]	89.09	**72.28**	67.40	65.37	**86.68**	59.24	73.34	84.17	86.91
BiSeNetV2 [58]	**89.78**	67.99	66.84	58.74	84.44	59.51	71.22	82.65	86.10
DeepLab V3+ [17]	-	-	-	-	-	-	73.18	-	86.90
SETR-PUP [4]	54.58	60.27	63.09	88.19	53.24	80.29	66.61	79.25	83.25
Proposed Method	59.72	71.28	**69.38**	**89.21**	66.16	**86.14**	**73.65**	**84.40**	**86.98**

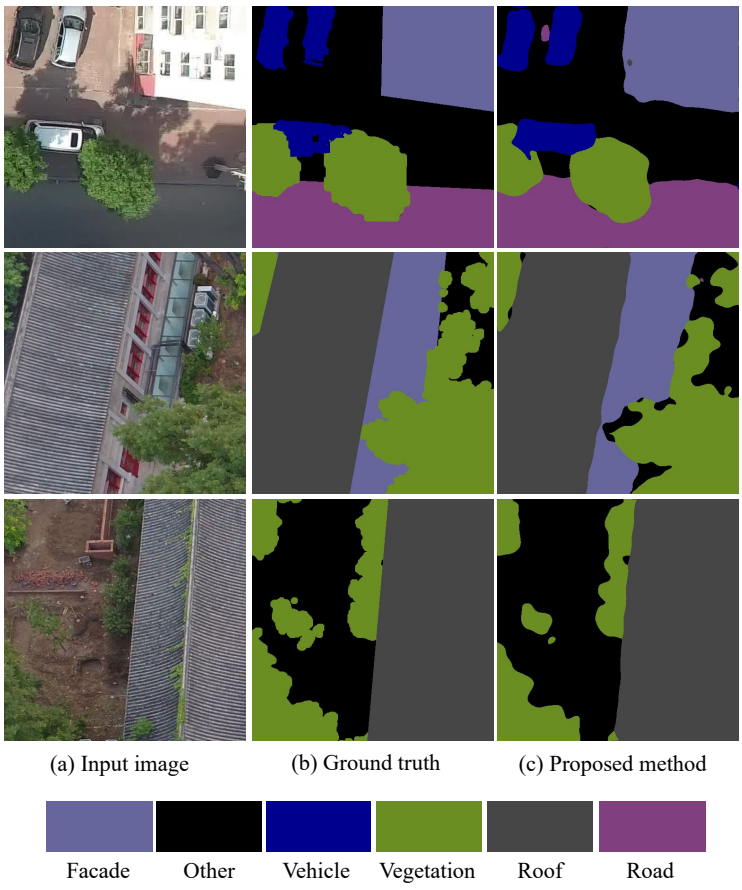

Figure 6. Qualitative prediction results using the UDD-6 validation dataset.

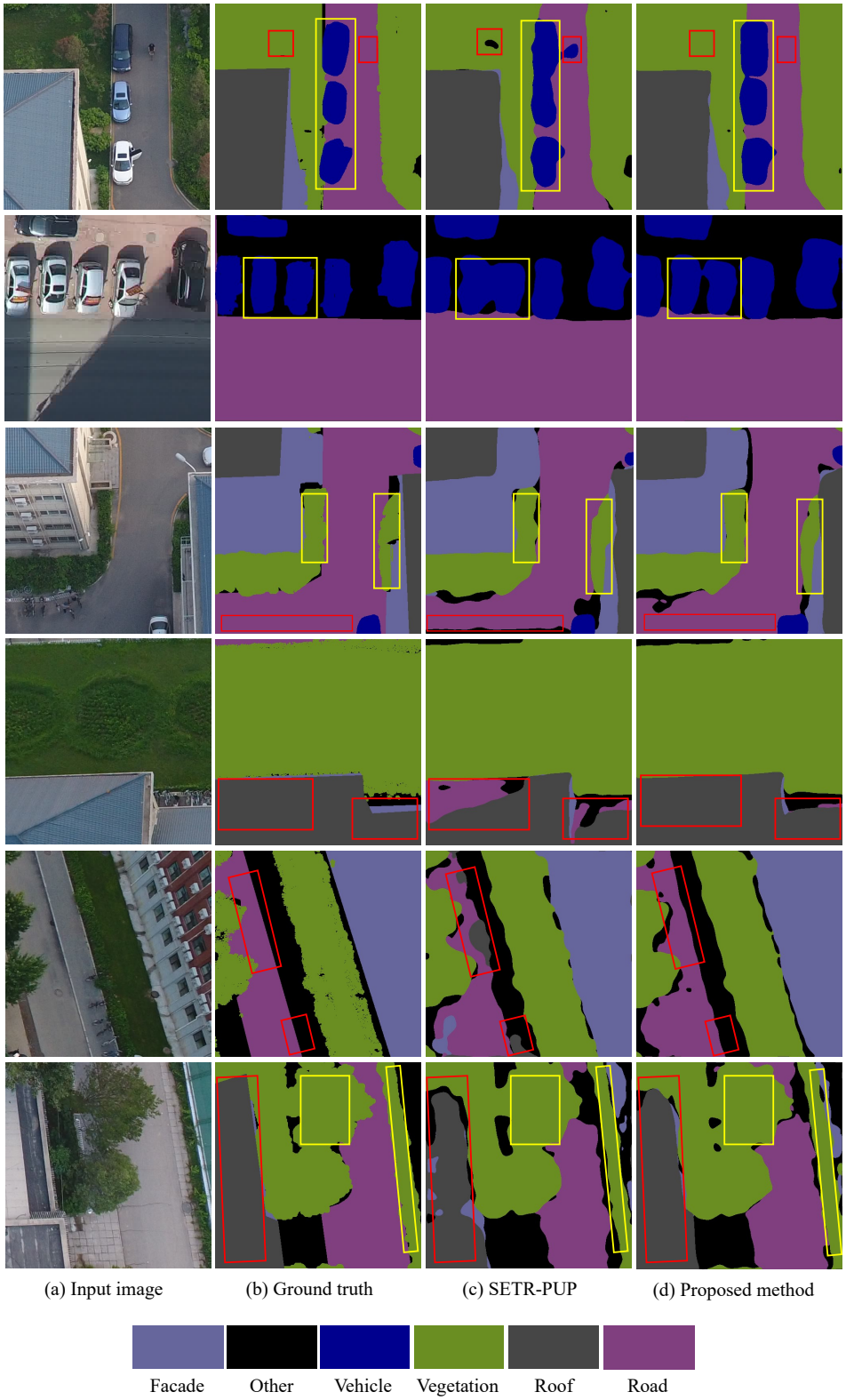

Figure 7. Qualitative prediction results using the UDD-6 validation dataset. The red and yellow boxes represent the local and global context, respectively, in the images.

4.5. Ablation Studies

In this subsection, we discuss the results of the sensitivity test for the adopted approach using the UDD-6 validation dataset. We investigate the effect of different convolution variants for ψ in the TSIF module. We tested 3×3 convolution and 3×3 dilated convolution

operations for this. A dilation rate of 2 is used for the dilated convolution. It can be seen from Table 3 that 3×3 convolution gives better results. It outperforms the dilated convolution by 0.56%, 0.41%, and 0.006% in mIoU, mean F1 score, and OA, respectively.

Table 3. The effect of different convolution variants for ψ in the TSIF module using the UDD-6 dataset.

ψ	mIoU (%)	Mean F1 (%)	O.A (%)
3×3 dilated convolution	73.09	83.99	86.92
3×3 convolution	**73.65**	**84.40**	**86.98**

We tested the nearest neighbor and bilinear upsampling policies to study the effect of distinct upsampling operations in the decoder network. It can be seen from Table 4 that the bilinear upsampling approach performs best. It outperforms the nearest neighbor by 0.81%, 0.59%, and 0.81% in mIoU, mean F1 score, and OA, respectively.

Table 4. The effect of a different upsampling policy in the decoder network using the UDD-6 dataset.

Upsample	mIoU (%)	Mean F1 (%)	O.A (%)
Nearest neighbor	72.84	83.81	86.80
Bilinear	**73.65**	**84.40**	**86.98**

The number of feature channels in the decoder network, N_c, is a crucial hyperparameter. We investigate the sensitivity for its two values, 512 and 256. We used a bilinear upsampling operation and 3×3 convolution for ψ in the TSIF module for both tests. The test results can be depicted in Table 5. Both the feature channels show very close performance in OA. However, the N_c of 256 shows superior results compared to the 512 in the mIoU and mean F1 score. Overall, it outperforms N_c of 512 by 0.51%, 0.37%, and 0.02% in mIoU, mean F1 score, and OA, respectively.

Table 5. The effect of different feature channels, N_c, in the decoder network using the UDD-6 dataset.

Upsample	ψ	N_c	mIoU (%)	Mean F1 (%)	O.A (%)
Bilinear	3×3 convolution	512	73.14	84.03	86.96
Bilinear	3×3 convolution	256	**73.65**	**84.40**	**86.98**

5. Discussion

Due to the huge background clutter and variations in aerial images, it poses many challenges for its semantic segmentation. The proposed framework captures the inherent features in a global and local context for UAV images precisely. The qualitative visualization of results shows a smooth segmentation output. The global context about two similar objects in an image such as cars, low vegetation, and trees is well maintained. This can be observed in Figures 4 and 5. The self-attention mechanism in the encoder network captures these long-range details about similar objects. We can also see in Figure 6 that our method can handle global information about cars and vegetation precisely. Similarly, in the first and second rows of Figure 7, global details about the cars are well captured by our method. Long-range vegetation is also well-segmented in the third and sixth rows.

The shape and size of the objects are well preserved, and the boundary information between the two classes is maintained smoothly. This can also be seen in the figures. The Token Spatial Information Fusion (TSIF) module in the encoder network captures these local context details around the neighboring pixels. The roof and facade are smoothly segmented in Figure 6, while preserving the local details. Compared to the SETR-PUP [4], the proposed method gives better results. SETR-PUP struggles to preserve information

around the neighboring pixels belonging to the same class. In the fourth and sixth rows of Figure 7, our method segments the roof well. Similarly, in the first and third rows, our method smoothly handles the local context for the vegetation and road categories. These can be seen in the regions marked in red boxes.

The encoder network produces strong feature representation, which is semantically rich in both high and low-level details. The gradual upsampling strategy followed by a convolution operation in the decoder network helps to preserve this rich information, and hence it generates smooth predictions. The quantitative and qualitative findings using the UAVid and UDD-6 datasets show the advantage of the proposed framework in UAV image segmentation.

We performed sensitivity tests for various hyperparameters in our method. As shown in Table 3, the 3×3 convolution for ψ in the TSIF module gives superior results as compared to the dilated convolution. It captures the inherent features of the pixels belonging to the same class in a better way. Similarly, the bilinear upsampling policy in the decoder network gives the best performance. It preserves the overall semantic contextual details precisely for UAV images. A combination of 3×3 convolution for ψ, the bilinear upsampling approach, and the feature channels of 256 in the decoder network generates the best results in segmenting the aerial scenes, as presented in Table 5. Increasing the number of kernels leads to a decrease in performance.

Overall, the proposed design shows competitive performance but it still struggles to segment small objects such as humans. However, when compared to other methods, our method shows decent performance for the human category, but segmenting this category is very challenging. The UAVid and UDD-6 datasets were captured in high-visibility conditions. The performance of the proposed method may degrade for images captured in different weather conditions. The design of an effective framework to handle small objects in UAV scenes, which is also resilient to different weather conditions, can be considered in future work.

6. Conclusions

In this study, we proposed a transformer-based design for the semantic segmentation of UAV street scene images. We used an encoder-decoder framework that captures the inherent features in the global and local context for UAV images. The self-attention-based encoder network captures long-range information, therefore the global context about two similar objects in the image are well maintained. The convolution-based TSIF module fuses local contextual details in the network. This helps the network to segment neighboring pixels while maintaining the proper shape and size of the objects. The decoder network uses semantically rich feature representations from the encoder network for final pixel-level predictions. It generates smooth segmentation with well-preserved boundary information between two classes. We performed a set of ablation studies for the sensitivity test. Overall, the proposed framework shows a competitive result on the two public datasets; UAVid and UDD-6. However, it struggles to segment small objects in complex UAV images. The datasets used in this work were captured in high visibility conditions, so the performance of the framework may degrade for different weather condition images. In future research, we will explore more potential segmentation designs for handling small objects in aerial images, which are also resilient to changing weather conditions.

Author Contributions: Conceptualization, S.K.; methodology, S.K., A.K. and D.-G.L.; software, S.K.; validation, A.K. and D.-G.L.; formal analysis, S.K., A.K. and D.-G.L.; investigation, A.K.; resources, D.-G.L.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, A.K. and D.-G.L.; visualization, S.K.; supervision, D.-G.L.; project administration, D.-G.L.; funding acquisition, D.-G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No.2021R1C1C1012590), (No. NRF-2022R1A4A1023248), Project BK21 FOUR and the Information Technology Research Center (ITRC) support program supervised by the Institute of Information Communications and Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) (IITP-2022-2020-0-01808).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The UAVid dataset can be found at https://uavid.nl/ (accessed on 12 December 2022). The Urban Drone (UDD-6) dataset can be found at https://github.com/MarcWong/UDD (accessed on 12 December 2022).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Appendix A.1. Evaluation Metrics

In this subsection, we describe the metrics used for the evaluation in this work.

Appendix A.1.1. Intersection over Union (IoU)

It is represented as follows:

$$IoU = \frac{True\ Positive}{True\ Positive + False\ Positive + False\ Negative} \tag{A1}$$

Appendix A.1.2. Mean Intersection over Union (mIoU)

It is the mean of the IoU score of all the classes. It can be represented as follows:

$$mIoU = \frac{IoU_1 + IoU_2 + \ldots + IoU_N}{N}, \tag{A2}$$

where N represents the number of classes.

Appendix A.1.3. Mean F1 Score (mean F1)

It is the mean of the F1 score of all the classes. The mathematical expression for the F1 score is as follows:

$$F1\ score = 2 \times \frac{P \times R}{P + R}, \tag{A3}$$

where P and R are mathematically calculated as follows:

$$P = \frac{True\ Positive}{True\ Positive + False\ Positive}, \tag{A4}$$

$$R = \frac{True\ Positive}{True\ Positive + False\ Negative}, \tag{A5}$$

the mean F1 score can be represented as follows:

$$mean\ F1 = \frac{Score_1 + Score_2 + \ldots + Score_N}{N}, \tag{A6}$$

where $Score_1, Score_2, \ldots, Score_N$ denote the $F1$ scores of all the classes, i.e., N.

Appendix A.1.4. Overall Accuracy (OA)

It is represented as follows:

$$OA = \frac{\#.\ Correct\ label\ predictions}{\#.\ Total\ label\ predictions} \tag{A7}$$

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
2. Singh, A.; Lall, B.; Panigrahi, B.; Agrawal, A.; Agrawal, A.; Thangakunam, B.; Christopher, D.J. Semantic segmentation of bone structures in chest X-rays including unhealthy radiographs: A robust and accurate approach. *Int. J. Med. Inf.* **2022**, *165*, 104831. [CrossRef] [PubMed]
3. Jin, Z.; Liu, B.; Chu, Q.; Yu, N. ISNet: Integrate image-level and semantic-level context for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7189–7198.
4. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
5. Lee, D.G. Fast Drivable Areas Estimation with Multi-Task Learning for Real-Time Autonomous Driving Assistant. *Appl. Sci.* **2021**, *11*, 10713. [CrossRef]
6. Yeung, H.W.F.; Zhou, M.; Chung, Y.Y.; Moule, G.; Thompson, W.; Ouyang, W.; Cai, W.; Bennamoun, M. Deep-learning-based solution for data deficient satellite image segmentation. *Expert Syst. Appl.* **2022**, *191*, 116210. [CrossRef]
7. Zeng, Z.; Xu, Y.; Xie, Z.; Tang, W.; Wan, J.; Wu, W. LEARD-Net: Semantic segmentation for large-scale point cloud scene. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102953. [CrossRef]
8. Su, Y.; Cheng, J.; Bai, H.; Liu, H.; He, C. Semantic Segmentation of Very-High-Resolution Remote Sensing Images via Deep Multi-Feature Learning. *Remote. Sens.* **2022**, *14*, 533. [CrossRef]
9. Guo, Y.; Wu, C.; Du, B.; Zhang, L. Density Map-based vehicle counting in remote sensing images with limited resolution. *ISPRS J. Photogramm. Remote Sens.* **2022**, *189*, 201–217. [CrossRef]
10. Hossain, M.D.; Chen, D. A hybrid image segmentation method for building extraction from high-resolution RGB images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 299–314. [CrossRef]
11. Ichim, L.; Popescu, D. Segmentation of vegetation and flood from aerial images based on decision fusion of neural networks. *Remote Sens.* **2020**, *12*, 2490. [CrossRef]
12. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; de Andrade Silva, J.; Matsubara, E.T.; Pistori, H.; Gonçalves, W.N.; Li, J. A review on deep learning in UAV remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456. [CrossRef]
13. Dutta, K.; Talukdar, D.; Bora, S.S. Segmentation of unhealthy leaves in cruciferous crops for early disease detection using vegetative indices and Otsu thresholding of aerial images. *Measurement* **2022**, *189*, 110478. [CrossRef]
14. Song, Z.; Zhang, Z.; Yang, S.; Ding, D.; Ning, J. Identifying sunflower lodging based on image fusion and deep semantic segmentation with UAV remote sensing imaging. *Comput. Electron. Agric.* **2020**, *179*, 105812. [CrossRef]
15. Lobo Torres, D.; Queiroz Feitosa, R.; Nigri Happ, P.; Elena Cué La Rosa, L.; Marcato Junior, J.; Martins, J.; Olã Bressan, P.; Gonçalves, W.N.; Liesenberg, V. Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution UAV optical imagery. *Sensors* **2020**, *20*, 563. [CrossRef] [PubMed]
16. Lyu, Y.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
17. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
18. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
19. Yin, M.; Yao, Z.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; Hu, H. Disentangled non-local neural networks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 191–207.
20. Lee, D.G.; Kim, Y.K. Joint Semantic Understanding with a Multilevel Branch for Driving Perception. *Appl. Sci.* **2022**, *12*, 2877. [CrossRef]
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
22. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
24. Girisha, S.; Verma, U.; Pai, M.M.; Pai, R.M. Uvid-net: Enhanced semantic segmentation of uav aerial videos by embedding temporal information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 4115–4127. [CrossRef]
25. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]

26. Yi, S.; Liu, X.; Li, J.; Chen, L. UAVformer: A Composite Transformer Network for Urban Scene Segmentation of UAV Images. *Pattern Recognit.* **2022**, *133*, 109019. [CrossRef]
27. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 568–578.
28. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
29. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
30. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
31. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
32. Li, T.; Cheng, B.; Ni, B.; Liu, G.; Yan, S. Multitask low-rank affinity graph for image segmentation and image annotation. *ACM Trans. Intell. Syst. Technol. (TIST)* **2016**, *7*, 1–18. [CrossRef]
33. Li, T.; Woo, J.; Kweon, I.S. Probabilistically Semantic Labeling of IR Image for UAV. In Proceedings of the MVA2007 IAPR Conference on Machine Vision Applications, Machine Vision and Application, Tokyo, Japan, 16–18 May 2007.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* .
35. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–14 August 2021; pp. 10347–10357.
36. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 357–366.
37. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional positional encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
38. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 558–567.
39. Zhang, W.; Huang, Z.; Luo, G.; Chen, T.; Wang, X.; Liu, W.; Yu, G.; Shen, C. TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 12083–12093.
40. Sun, Z.; Cao, S.; Yang, Y.; Kitani, K.M. Rethinking transformer-based set prediction for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3611–3620.
41. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
42. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 7262–7272.
43. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 12179–12188.
44. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
45. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating convolution designs into visual transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 579–588.
46. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 22–31.
47. Yang, M.Y.; Kumaar, S.; Lyu, Y.; Nex, F. Real-time semantic segmentation with context aggregation network. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 124–134. [CrossRef]
48. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
49. Liu, S.; Cheng, J.; Liang, L.; Bai, H.; Dang, W. Light-Weight Semantic Segmentation Network for UAV Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 8287–8296. [CrossRef]
50. Iqbal, J.; Ali, M. Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 263–275. [CrossRef]
51. Gebrehiwot, A.; Hashemi-Beni, L.; Thompson, G.; Kordjamshidi, P.; Langan, T.E. Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. *Sensors* **2019**, *19*, 1486. [CrossRef]

52. Zhang, C.; Atkinson, P.M.; George, C.; Wen, Z.; Diazgranados, M.; Gerard, F. Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 280–291. [CrossRef]
53. Zhang, X.; Jin, J.; Lan, Z.; Li, C.; Fan, M.; Wang, Y.; Yu, X.; Zhang, Y. ICENET: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. *Remote Sens.* **2020**, *12*, 221. [CrossRef]
54. Gevaert, C.M.; Persello, C.; Sliuzas, R.; Vosselman, G. Monitoring household upgrading in unplanned settlements with unmanned aerial vehicles. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *90*, 102117. [CrossRef]
55. Chen, Y.; Wang, Y.; Lu, P.; Chen, Y.; Wang, G. Large-scale structure from motion with semantic constraints of aerial images. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Guangzhou, China, 23–26 November 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 347–359.
56. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
57. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272. [CrossRef]
58. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comp. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
59. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
60. Zhuang, J.; Yang, J.; Gu, L.; Dvornek, N. Shelfnet for fast semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
61. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.