BIDIRECTIONAL MULTI-SCALE ATTENTION NETWORKS FOR SEMANTIC SEGMENTATION OF OBLIQUE UAV IMAGERY

Ye Lyu[a]*, George Vosselman[a], Gui-Song Xia[b], Michael Ying Yang[a]

[a]University of Twente, The Netherlands - (y.lyu, george.vosselman, michael.yang)@utwente.nl
[b]Wuhan University, China - guisong.xia@whu.edu.cn

KEY WORDS: Semantic Segmentation, Multi-Scale, Attention, Oblique View, UAV, Deep Learning

ABSTRACT:

Semantic segmentation for aerial platforms has been one of the fundamental scene understanding task for the earth observation. Most of the semantic segmentation research focused on scenes captured in nadir view, in which objects have relatively smaller scale variation compared with scenes captured in oblique view. The huge scale variation of objects in oblique images limits the performance of deep neural networks (DNN) that process images in a single scale fashion. In order to tackle the scale variation issue, in this paper, we propose the novel bidirectional multi-scale attention networks, which fuse features from multiple scales bidirectionally for more adaptive and effective feature extraction. The experiments are conducted on the UAVid2020 dataset and have shown the effectiveness of our method. Our model achieved the state-of-the-art (SOTA) result with a mean intersection over union (mIoU) score of 70.80%.

1. INTRODUCTION

Semantic segmentation has been one of the most fundamental research tasks for scene understanding. It is to assign each pixel within an image with the class label it belongs to. There have been many works for semantic segmentation on the remote sensing images and the aerial images (Demir et al., 2018, Rottensteiner et al., 2014), which are captured in nadir view style. The spatial resolutions in such images are approximately the same for all pixels. Oblique views have a much larger land coverage if the platforms are at the same flight height. For example, the unmanned aerial vehicle (UAV) platform has been used to for urban scene observation (Lyu et al., 2020, Nigam et al., 2018). The images of different viewing directions are shown in Figure 1. The left image of nadir view is from the Vaihingen dataset (Rottensteiner et al., 2014), while the right image of oblique view is from the UAVid2020 dataset (Lyu et al., 2020). Compared with the images in nadir view style, the images in oblique view have very large spatial resolution variation across the entire image.

The state-of-the-art methods for semantic segmentation all rely on powerful deep neural networks, which can effectively extract high-level semantic information to determine the class types for all pixels. Deep neural networks serve as non-linear functions, which map an image input to a label output. Due to its non-linear property, the label output will not scale linearly as the image input scales. When designing the deep neural networks, there is usually a performance trade-off for objects in different scales. For example, the semantic segmentation of a small car in a remote sensing image is better handled in higher resolution where finer details can be observed, such as wheels. For larger objects like roads and buildings, it is better to have more global context to recognize the objects since their whole shapes can be observed for semantic segmentation.

When objects in an image dataset have very large scale variation, the semantic segmentation performance of deep neural networks will drop if this multi-scale problem is not considered in the network design. A simple strategy is to apply multi-scale inference (Zhao et al., 2017), i.e., a well-trained deep neural networks predict the score maps of the same image in multiple different scales, and the score maps are averaged to determine the final label prediction. Such strategy generally provides better performance. However, a good prediction from a proper scale could be undermined by those worse predictions from other scales, which limits the model performance. Max-pooling selects one score map prediction of multiple scales for each pixel, but the optimal output could be the interpolation of the prediction of multiple scales. A smarter way of fusing the output score maps is to leverage on an attention model (Chen et al., 2016), which determines the weights when fusing the score maps of different scales. The strategy has been extended to a hierarchical structure for better performance (Tao et al., 2020).

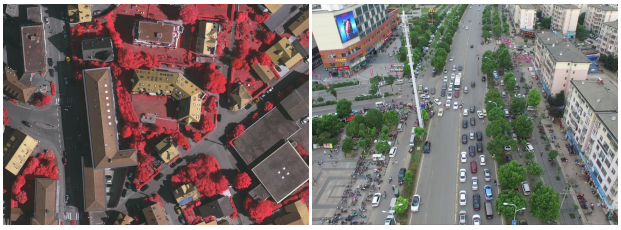

Figure 1. Example of images in different viewing style. The left image from Vaihingen dataset(Rottensteiner et al., 2014) is captured in nadir view. The right image from UAVid2020 dataset(Lyu et al., 2020) is captured in oblique view.

With respect to the design of deep neural networks, there are several strategies to relieve the multi-scale problem. The first strategy is to gradually refine features from coarse scales to fine scales (Long et al., 2015, Ronneberger et al., 2015). The second strategy is to design a multi-scale feature extractor module in the middle of the deep neural networks (Zhao et al., 2017, Chen et al., 2017, Chen et al., 2018, Yuan and Wang, 2018, Lyu et al., 2020). Self-attention (Fu et al., 2019, Huang et al., 2019, Yuan

* Corresponding author

et al., 2020) and graph networks (Liang et al., 2018, Li and Gupta, 2018) have also been applied to aggregate information globally to reinforce the features for each pixel.

In this paper, we propose the bidirectional multi-scale attention networks (BiMSANet) to address the multi-scale problem in the semantic segmentation task. Our method is inspired by the multi-scale attention strategy (Chen et al., 2016, Tao et al., 2020) and the feature level fusion strategy (Chen et al., 2017, Zhao et al., 2017), and jointly fuses the features guided by the attention of different scales in bidirectional pathways, i.e, coarse-to-fine and fine-to-coarse. Our method is tested on the new UAVid2020 dataset (Lyu et al., 2020). One of its challenges is the huge inter-class and intra-class scale variance for different objects due to its oblique viewing style. Our method achieves a new state-of-the-art result with a mIoU score of 70.8%. Compared with the currently top ranked method (Tao et al., 2020), which features on handling the multi-scale problem, our methods outperforms by almost 0.8%.

The contributions of this paper are summarized as follows,

- We have proposed a novel bidirectional multi-scale attention networks to handle the multi-scale problem for the semantic segmentation task.
- We have visualized in multiple perspectives and analyzed the bidirectional multi-scale attentions in details.
- We have achieved state-of-the-art result on the UAVid2020 benchmark, and the code will be made public.

2. RELATED WORK

In this section, we will discuss other works that are related to our paper. In order to handle the multi-scale problem for the semantic segmentation, a number of deep neural networks have been designed.

Multi-scale feature fusion. The first basic type of method is to aggregate features of multiple scales from deep neural networks. FCN (Long et al., 2015) and U-Net (Ronneberger et al., 2015) have adopted skip connections between encoder and decoder to gradually fuse the information from multiple scales. MSDNet (Lyu et al., 2020) has extended the connection across scales to further increase the performance. ZipZagNet (Di Lin, 2019) uses a more complex zip-zag architecture between the backbone and the decoder for intermediate multi-scale feature aggregation. HRNet (Wang et al., 2019) proposes a multi-scale backbone to exchange information between branches of coarse scale and fine scale. BiSeNet (Yu et al., 2018) proposes a dual branch structure for better performance, one branch for higher spatial resolution, while the other for richer semantic features.

Multi-scale context extraction. Another method is to aggregate multi-scale context from the same feature maps with a module. PSPNet (Zhao et al., 2017) has adopted pyramid pooling module, which has pooling modules of multiple scales to pool context features for the object recognition. DeepLabv3 (Chen et al., 2017, Chen et al., 2018) has utilized atrous spatial pyramid pooling module, which assembles multi-scale features with convolutions of multiple atrous rates. OCNet (Yuan and Wang, 2018) proposes pyramid object context (Pyramid-OC) module and atrous spatial pyramid object context (ASP-OC) module to extract object context in multiple scales.

Context by relations. With the creation of self-attention mechanism (Vaswani et al., 2017) for natural language processing, better semantic segmentation results have also been achieved when self-attention is applied to reason the relation between pixels. Self-attention refines the features in a non-local style, which aggregates information for each pixel globally. DANet (Fu et al., 2019) has utilized dual attention module, position attention and channel attention, to extract information globally. CCNet (Huang et al., 2019) has applied the criss-cross attention module to reduce the computational complexity of the self-attention. OCRNet (Yuan et al., 2020) has used explicit class attention to reinforce the features. However, these types of methods are normally intensive in memory and computation as there are too many pixels, resulting in very dense connections between them. Graph reasoning is another way to include relations among objects. Instead of adopting dense pixel relations, sparse graph structure makes the context relation reasoning less intensive in memory and computation. Symbolic graph reasoning (SGR) layer has been proposed (Liang et al., 2018) for context information aggregation through knowledge graph. 2D images have been transformed into a graph structure (Li and Gupta, 2018), whose vertices are clusters of pixels. Context information is propagated across all vertices on the graph.

Inference in multi-scale. Multi-scale inference is widely used to provide more robust prediction, which is orthogonal to previously discussed methods as those networks can be regarded as a trunk for multi-scale inference. Average pooling and max pooling on score maps are mostly used, but they limit the performance. Attention has been applied for fusing score maps across multiple scales (Chen et al., 2016) for better performance. The method is more adaptive to objects in different scales as the weights for fusing score maps across multiple scales can vary. It has been further improved by introducing a hierarchical structure (Tao et al., 2020), which allows different network structures during training and testing to improve the model design. Our paper also focuses on the multi-scale inference. We have further improved the multi-scale attention mechanism by introducing feature level bidirectional fusion.

3. PRELIMINARY

In this section, we first go through some network architecture design to better help understand the newly proposed bidirectional multi-scale attention networks.

3.1 Multi-Scale-Dilation Net

The multi-scale-dilation net (Lyu et al., 2020) is proposed as the first attempt to tackle the multi-scale problem for the UAVid dataset. The basic idea shares the philosophy of multi-scale image inputs, where the input images are scaled by the scale to batch operation and batch to scale operation. The intermediate features are concatenated from coarse to fine scales, which are used to output the final semantic segmentation output. The structure is shown in Figure 2. The feature extraction part is named as Trunk, features as Feat, and segmentation head as Seg in the following figures.

3.2 Hierarchical Multi-Scale Attention Net

The hierarchical multi-scale attention net (Tao et al., 2020) is proposed to learn to fuse semantic segmentation outputs of adjacent scales by a hierarchical attention mechanism. The deep neural networks learn to segment the images while predicting

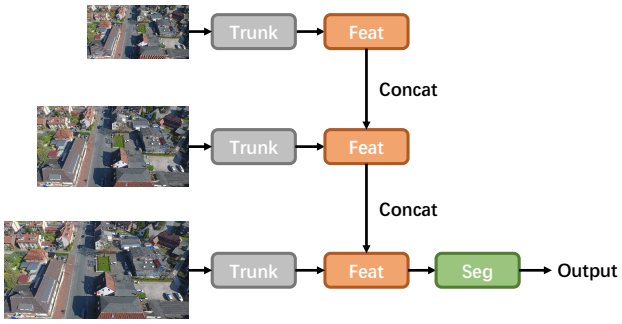

Figure 2. Architecture of the multi-scale dilation net. Features are aggregated from coarse to fine scales with concatenation.

the weighting masks for fusing the score maps. This method ranks as the top method in the Cityscapes pixel-level semantic labeling task (Cordts et al., 2016), which focuses on the multi-scale problem. The hierarchical mechanism allows different network structures during training and inference, e.g., the networks have only two branches of two adjacent scales during training, while the networks could have three branches of three adjacent scales during testing as shown in Figure 3. up and down stand for bilinear upsampling and downsampling, respectively.

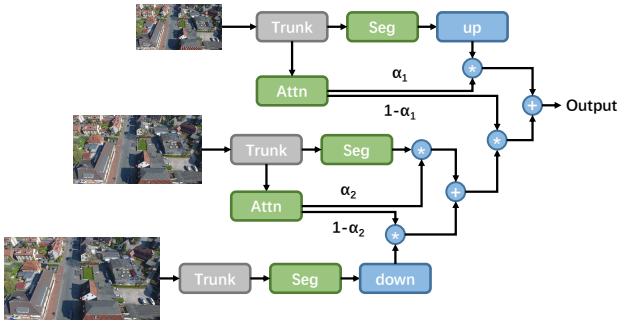

Figure 3. Architecture of the hierarchical multi-scale attention networks. In addition to the predicted score maps, extra weighting masks are predicted from the attention sub-networks for fusing the score maps of adjacent scales. $\oplus$, $\circledast$ stand for element-wise addition and multiplication, respectively.

3.3 Feature Level Hierarchical Multi-Scale Attention Net

One limitation of the hierarchical multi-scale attention networks is that the fused score maps are the linear interpolation of the score maps in adjacent scales, whereas the best score maps could be acquired with the interpolated features instead. A simple solution that we propose is to move the segmentation head to the end of the fused features as shown in Figure 4.

4. BIDIRECTIONAL MULTI-SCALE ATTENTION NETWORKS

In this section, the structure of the proposed bidirectional multi-scale attention networks will be introduced.

4.1 Overall Architecture

Our design also takes the hierarchical attention mechanism and the feature level fusion into account. The overall architecture is shown in Figure 5. For the input image I of size $H \times W$,

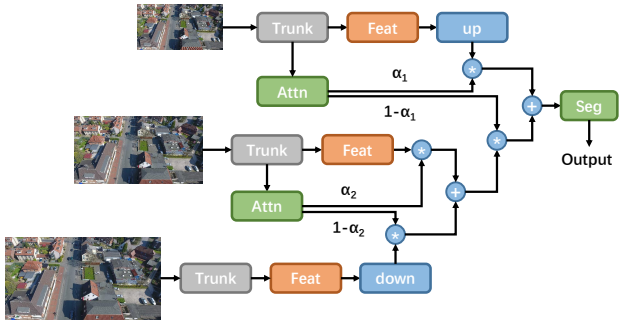

Figure 4. Architecture of the hierarchical multi-scale attention networks with feature level fusion. Segmentation head is moved to the end of the fused features. $\oplus$, $\circledast$ stand for element-wise addition and multiplication, respectively.

the image pyramid is built by adding two extra images $I_{2\times}$ and $I_{0.5\times}$, which are acquired by bi-linear up-sampling I to size of $2H \times 2W$ and bi-linear down-sampling I to $\frac{1}{2}H \times \frac{1}{2}W$. The bidirectional multi-scale attention networks have two pathways for feature fusing in a hierarchical manner. For each pathway, the structure is the same as the feature level hierarchical multi-scale attention nets. The design of the two pathways allows the feature fusion from both directions, and the fusion weights can be better determined in a better scale. The reason to use feature level fusion is that we need distinct features for two pathways. If the score maps are used for fusion, the feat1 and the feat2 in the two pathways would be the same, which limits the representation power of the two pathways. The two pathways take advantage of their own attention branches and features. Attn1 branch and Feat1 are for the coarse to fine pathway, while Attn2 branch and Feat2 are for the fine to coarse pathway. The Feat1 and the Feat2 from two pathways are fused hierarchically across scales, and the final feature is the concatenation of the features from the two pathways.

The Feat1 and Feat2 are reduced to the half number of channels as the Feat in feature level hierarchical multi-scale attention net. This setting is to provide fair comparisons between these two types of networks, since it leads to features with the same number of channels before the segmentation head.

The parameter sharing is also applied in the design. Three branches corresponding to the three scales share the same network parameters for Trunk, Attn1 and Attn2. Feat1 and Feat2 in the three branches are different as they are the output of different image inputs through the same trunk.

4.2 Module Details

In this section, we will illustrate the details of each component.

Trunk. In order to effectively extract information from each single scale, we have adopted the deeplabv3+ (Chen et al., 2018) as the trunk. We apply the wide residual networks (Zagoruyko and Komodakis, 2016) as the backbone, namely the WRN-38, which has been pre-trained on the imagenet dataset (Deng et al., 2009). The ASPP module in the deeplabv3+ has convolutions with atrous rate of $1, 6, 12$, and 18. The features f_b from the deeplabv3+ are further refined with a sequence of modules as follows, $Conv3 \times 3(256) -> BN -> ReLU -> Conv3 \times 3(256) -> BN -> ReLU -> Conv1 \times 1(nc)$ (numbers in the brackets are the numbers of output channels), which corresponds to the feature transformation in the Seg of

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2021-75-2021 | © Author(s) 2021. CC BY 4.0 License.

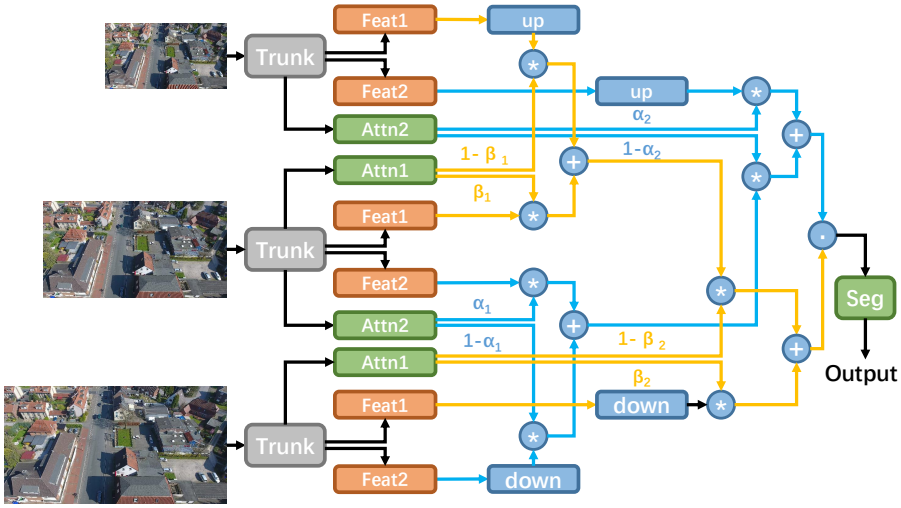

Figure 5. Architecture of the bidirectional multi-scale attention networks. The structure is the combination of two feature level hierarchical multi-scale attention nets corresponding to two pathways, where they share the same trunks. The coarse to fine pathway and the fine to coarse pathway are marked with the yellow and the blue arrows, respectively. $\oplus, \circledast$ stand for element-wise addition and multiplication, respectively. $\odot$ stands for concatenation in channel dimension.

the hierarchical multi-scale attention net before the final classification. $Conv, BN, ReLU$ stand for convolution, batchnorm, and rectified linear unit, respectively (Chen et al., 2018).

The trunk T transforms an image input I into feature maps f with nc channels, i.e., $f = T(I)$. $nc = n_{class} \times d$, where n_{class} is the total number of classes for the semantic segmentation task. d is the expansion rate for the channels. d is set to 4 in our case. The first $\frac{1}{2}nc$ channels are for the Feat1, while the second $\frac{1}{2}nc$ channels are for the Feat2.

Attention head. The Attn1 and the Attn2 share the same structure, but with different parameters. The attention heads map the features f_b from the deeplabv3+ to the attention weights α, β (ranging from 0.0 to 1.0 with $\frac{1}{2}nc$ channels) for the two pathways. For each attention head, the structure is comprised of a sequence of modules as follows, $Conv3\times3(256)->BN->ReLU->Conv3\times3(256)->BN->ReLU->Conv1\times1(\frac{1}{2}nc)->Sigmoid$ (numbers in the brackets are the output channels).

Segmentation head. The segmentation head Seg converts the fused input feature maps f_{fused} into score maps l ($8channels$ for the UAVid2020 dataset), which correspond to the class probabilities for all the pixels, i.e., $l = Seg(f_{fused})$. The segmentation head is simply a 1×1 convolution, $Conv1 \times 1(n_{class})$. Argmax operation along the channel dimension outputs the final class labels for all the pixels.

Auxiliary semantic head. As in (Tao et al., 2020), we apply auxiliary semantic segmentation heads for each branch during training, which consists of only a 1×1 convolution, $Conv1 \times 1(n_{class})$.

4.3 Training and inference

As our model follows the hierarchical inference mechanism, it allows our model to be trained with only 2 scales, while to infer with 3 scales ($0.5\times, 1\times, 2\times$). Such design makes it possible for our network to adopt a large trunk such as deeplabv3+ with WRN-38 backbone for better performance. We use RMI loss (Zhao et al., 2019) for the main semantic segmentation head and cross entropy loss for the auxiliary semantic head.

5. EXPERIMENTS

In this section, we will illustrates the implementation details for the experiments and compare the performance of different models on the UAVid2020 dataset. Our code is available on Github[1].

5.1 Dataset and Metric

Our experiments are conducted on the public UAVid2020 dataset[2] (Lyu et al., 2020). The UAVid2020 dataset focuses on the complex urban scene semantic segmentation task for 8 classes. The images are captured in oblique views with large spatial resolution variation. There are 420 high quality images of $4K$ resolutions (4096×2160 or 3840×2160) in total, split into training, validation and testing sets with 200, 70 and 150 images, respectively. The performance of different models are evaluated on the test set of the UAVid2020 benchmark. The performance for the semantic segmentation task is assessed based on the standard mean intersection-over-union metric (Everingham et al., 2015).

5.2 Implementation

Training. All the models in the experiments are implemented with pytorch (Paszke et al., 2019), and trained on a single Tesla V100 GPU of $16G$ memory with a batch of 2 images. Mixed precision and synchronous batch normalization are applied for the model. Stochastic gradient decent with a momentum 0.9 and weight decay of $5e^{-4}$ is applied as the optimizer for training. "Polynomial" learning rate policy is adopted (Liu et al., 2015) with a poly exponent of 2.0. The initial learning rate is set to $5e{-}3$. The model is trained for 175 epochs by random image selection. We apply random scaling for the images from $0.5\times$ to $2.0\times$. Random cropping is applied to acquire image patches of size of 896×896.

Testing. As the $4K$ image is too large to fit into the GPU, we apply cropping during testing as well. The image is partitioned

[1] https://github.com/YeLyuUT/semantic_segmentation
[2] https://uavid.nl/

Methods	mIoU(%)	Clutter	Building	Road	Tree	Vegetation	Moving Car	Static Car	Human
MSDNet	56.97	57.04	79.82	73.98	74.44	55.86	62.89	32.07	19.69
DeepLabv3+	67.36	66.68	87.61	80.04	79.49	62.00	71.69	68.58	22.76
HMSANet	70.03	69.32	88.14	82.12	79.42	61.21	77.33	72.52	30.17
FHMSANet	70.33	69.36	87.95	82.69	80.06	62.66	76.88	72.90	30.12
BiMSANet	70.80	69.94	88.63	81.60	80.38	61.64	77.22	75.62	31.34

Table 1. Performance comparisons for different models in intersection-over-union metric. The top ranked scores are marked in colors. Red for the 1st place, green for the 2nd place, and blue for the 3rd place.

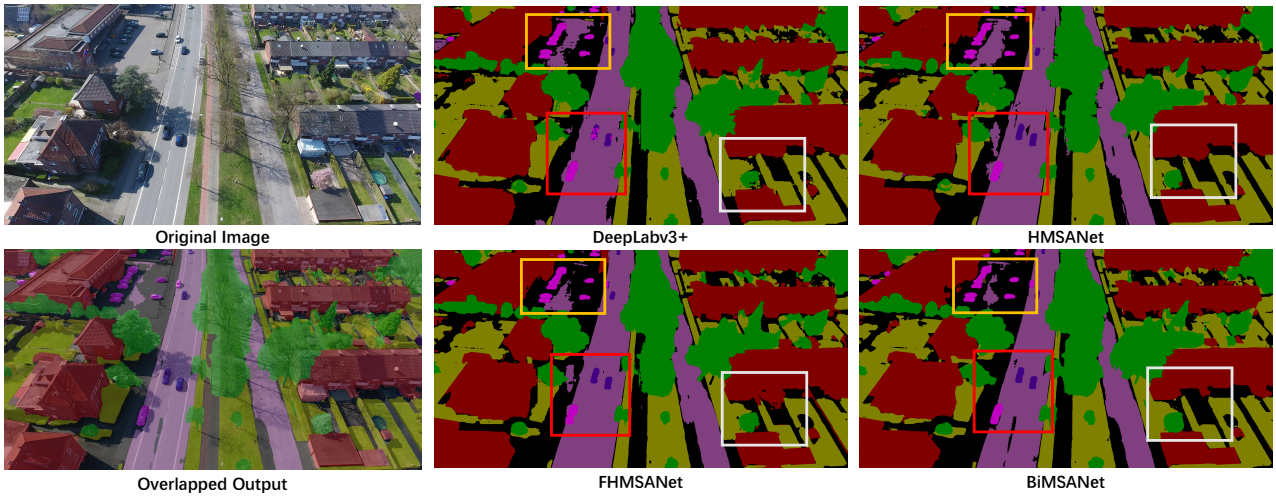

Figure 6. Qualitative comparisons of different models on the UAVid2020 test set. The example image is from the test set (seq30, 000400). Bottom left image shows the overlapped result of the BiMSANet output and the original image. Three example regions for comparisons are marked in red, orange, and white boxes.

into overlapped patches for evaluation as in (Lyu et al., 2020) and the average of the score maps are used for the final output in the overlapped regions. The crop size is set to 896×896 with an overlap of 512 pixels in both horizontal and vertical directions.

5.3 Model Comparisons

In this section, we will presents the semantic segmentation results on the test set of UAVid2020 dataset for multi-scale-dilation net (MSDNet) (Lyu et al., 2020), deeplabv3+ (Chen et al., 2018), hierarchical multi-scale attention net (HMSANet) (Tao et al., 2020), feature level hierarchical multi-scale attention net (FHMSANet), and our proposed bidirectional multi-scale attention networks (BiMSANet). The MSDNet is included as reference, which uses an old trunk FCN-8s (Long et al., 2015) in each scale. The major comparisons are among DeepLabv3+, HMSANet, FHMSANet, and BiMSANet.

The mIoU scores and the IoU scores for each individual class are shown in Table 1. Among all the compared models, the BiMSANet performs the best regarding the mIoU metric. Our BiMSANet has a more balanced prediction ability for both large and small objects.

For the evaluation of each individual class, the BiMSANet ranks the first for classes of clutter, building, tree, static car, and human. The most distinct improvement is for the static car, which is 2.72% higher than the second best score. With only the context information, our method could achieve decent scores for classes of both moving car and static car.

For human class, the scores of HMSANet, FHMSANet and BiMSANet are all significantly higher than the DeepLabv3+, which shows the superiority of multi-scale attention mechanism in handling the small objects. Thanks to the bidirectional multi-scale attention design, BiMSANet achieves the best performance for the human class.

Qualitative comparisons are shown in Figure 6. The example image is selected from the test set (seq30, 000400). As the ground truth label is reserved for benchmark evaluation, the overlapped output is shown instead in Figure 6. Three example regions are marked in red, orange, and white boxes.

In the red box region, it could be seen that the deeplabv3+ struggles to give coherent predictions for cars in the middle of the road, while the other three models have better results due to the multi-scale attention. The HMSANet and the FHMSANet wrongly classify part of the sidewalks, which is outside the road, as road class. BiMSANet handles better in this area. However, part of the road near the lane-mark are wrongly classified as clutter by the BiMSANet. In the orange box region, the parking lot, which belongs to the clutter class, is predicted as the road by all four models, and the BiMSANet makes the least error. In the white box region, the ground in front of the entrance door is wrongly classified as building by all models except the BiMSANet. This is benefited from the bidirectional multi-scale attention design.

We have also shown the performance for human class segmentation in Figure 7. The example image is from the test set (seq22, 000900). The zoomed in images in the middle and the right columns correspond to the patches in the white boxes of the overlapped output. The four patches are from different context, which is very complex in some local regions. Even though the humans in the image are quite small and in many different poses, such as standing, sitting, and riding, our model can still effectively detect and segment most of the humans in the image.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2021-75-2021 | © Author(s) 2021. CC BY 4.0 License.

Figure 7. Qualitative example of human class segmentation by the BiMSANet. The example image is from the test set (seq22, 000900). The left column shows the original full image and the overlapped output. The middle and the right columns show the image patches cropped from the overlapped output (marked by white boxes), which all focus on the human class. The red circles mark some missing segmentation.

Methods	mIoU(%)	mIoU Gains(%)	Trunk	Multi-Scale Attention	Feature Level Fusion	Bidirection
DeepLabv3+	67.36	-	✓	-	-	-
HMSANet	70.03	+2.67	✓	✓	-	-
FHMSANet	70.33	+0.30	✓	✓	✓	-
BiMSANet	70.80	+0.47	✓	✓	✓	✓

Table 2. Ablation study for models. The performance gains could be observed by gradually adding components.

5.4 Ablation Study

In this section, we will compare the performance gains by gradually adding the components. The results are shown in Table 2. It is easy to see that the multi-scale processing is useful for the oblique view UAV images. The mIoU score has increased by 2.67% by including the multi-scale attention into the networks. The feature level fusion is also proved to be useful as it helps the networks to improve the mIoU score by 0.3%. By further adding the bidirectional attention mechanism, the networks improve the mIoU score by another 0.47%.

5.5 Analysis of Learned Multi-Scale Attentions

In this section, we will analyze the learned multi-scale attentions from the BiMSANet to better understand how the attentions work. We explore from mainly three perspectives: attentions of different channels, different scales, and different directions. The example image is from the test set (seq25,000400). Attentions from both Attn1 branch and Attn2 branch are used, noted as α and β in Figure 5. α is for the fine to coarse pathway, while β is for the coarse to fine pathway.

5.5.1 Attention of different channels The multi-scale attentions in our BiMSANet have $\frac{1}{2}nc$ channels (16 in our case), which is different from the HMSANet (Tao et al., 2020), whose attention has only one single channel for all classes. The attentions guide the fusion of features across scales. Example attentions of different channels in $1\times$ scale branch are shown in Figure 8. Different channels have different attentions focusing on different parts of the image. It is obvious that different channels have different focus for different classes, e.g., $1th$ channel more focus on trees, $3th$ channel less focus on roads, and $7th$ channel have the most focus on moving cars.

5.5.2 Attention of different scales In order to analyze the difference of attentions in different scales, we have selected 4 attentions from each of the Attn1 branch and the Attn2 branch as shown in Figure 9. The superscripts are the channel index of the attentions. By comparing the α_1 with α_2, which are predicted in $1\times$ and $0.5\times$ scales, we could see that attentions in different scales have different focus. The difference of the same channel between α_1 and α_2 are more worth of comparisons. The same applies for β_1 and β_2.

From α_1^1 and α_2^1, it could be noted that the recognition of cars in closer distance are more based on context, since the values of α_2^1 are larger than α_1^1. The recognition of road that are closer to the camera also relies more on the coarser level features, which is reasonable as the road area is large and requires more context for recognition. It is also interesting to note that the middle lane-marks is even brighter than other parts of the road in α_2^1, which means the recognition requires more context. It is reasonable as the color and the texture of the lane-marks are quite different compared with other parts of the road. The distant buildings near the horizon rely more on the coarser level features as well.

We have also noticed that the α_2 ($0.5\times$ scale) and β_2 ($2\times$ scale) have larger values on average compared with α_1 and β_1 ($1\times$ scale), which means that features with context information and fine details are both valuable for object recognition.

5.5.3 Attention of different directions In our bidirectional design, both the coarse to fine pathway and the fine to coarse pathway fuse the features from three scales ($0.5\times, 1\times, 2\times$). In this section, we analyze if the feature fusion in two pathways has the same attention pattern. Attention examples are shown in Figure 10. Attentions α_2 and $1-\beta_1$ from two pathways are both for the feature fusion across scale $0.5\times$ and $1\times$. Although

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2021-75-2021 | © Author(s) 2021. CC BY 4.0 License.

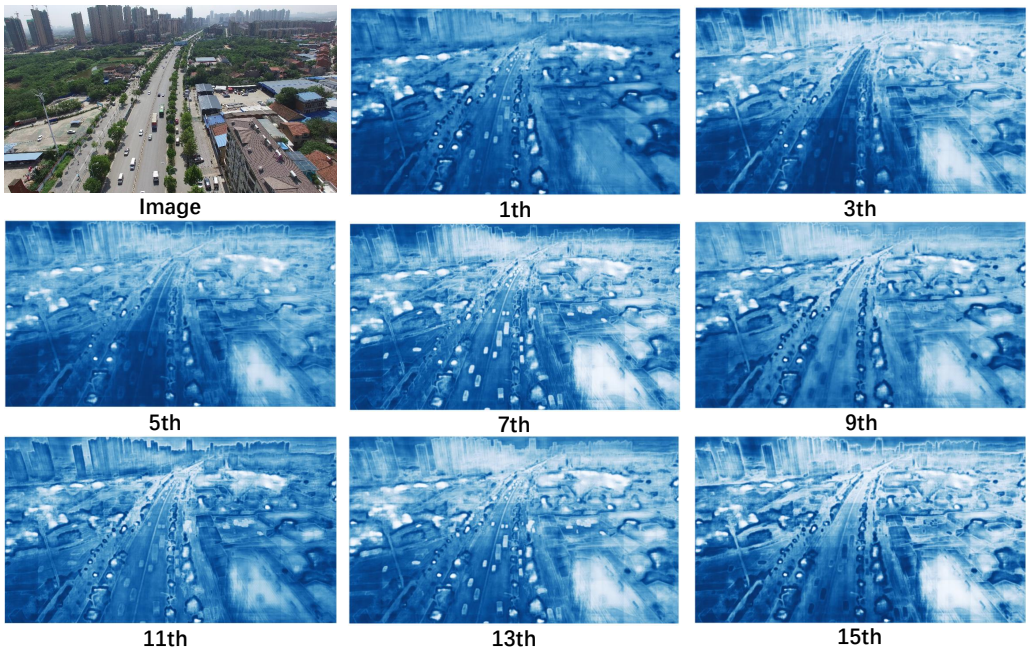

Figure 8. Attention analysis of different channels. Example attentions are of $1\times$ scale from Attn1 branch. The image on the top left shows the image adopted. The other 8 images are the attention maps from different channels. Channel indices are presented below the images. Brighter color means higher value. Best visualized with zoom in.

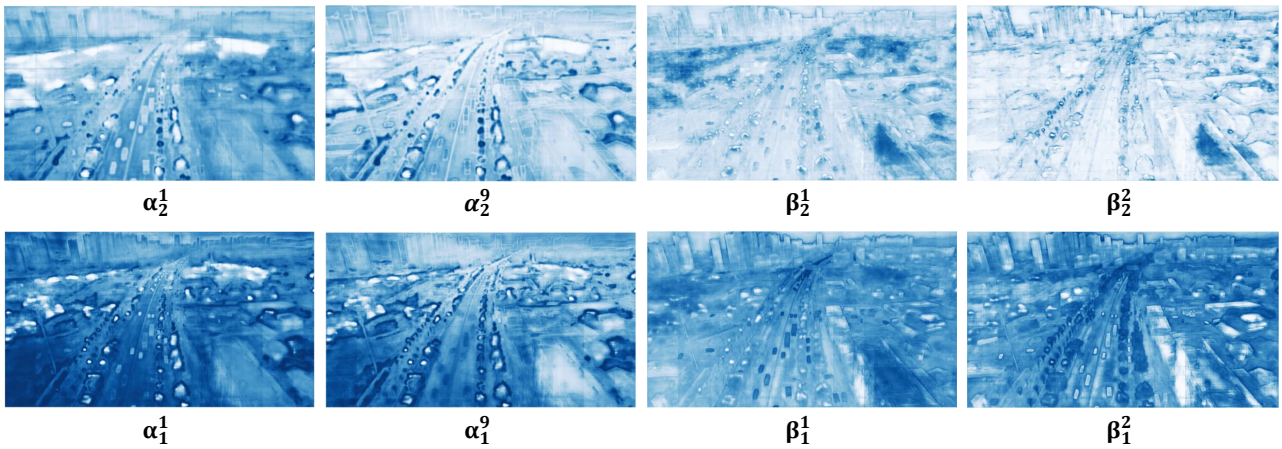

Figure 9. Attention analysis of different scales. We select 4 attentions from each of the Attn1 branch and the Attn2 branch. α, β are of the same meaning as in Figure 5. The superscripts are the channel index of the attentions. α_2, β_2 correspond to the attentions predicted in the $0.5\times$ scale and the $2\times$ scale. α_1, β_1 are predicted in $1\times$ scale. Brighter color means higher value. Best visualized with zoom in.

the attention values of same pixels can not be directly compared as the feature sources are different (Feat1 and Feat2), it is still evident that the attention densities on average are quite different. There are more activation in α_2 than $1 - \beta_1$, showing that the two pathways play different roles for feature fusion across same scales.

6. CONCLUSION

In this paper, we have proposed the bidirectional multi-scale attention networks (BiMSANet) for the semantic segmentation task. The hierarchical design adopted from (Tao et al., 2020) allows the usage of larger trunk for better performance. The feature level fusion and the bidirectional design allows the model to more effectively fuse the features from both of the adjacent coarser scale and the finer scale. We have conducted the experiments on the UAVid2020 dataset (Lyu et al., 2020), which have large variation in spatial resolution. The comparisons among different models have shown that our BiMSANet achieves better results by balancing the performance of small objects and large objects. Our BiMSANet achieves the state-of-art result with a mIoU score of 70.80% for the UAVid2020 benchmark.

REFERENCES

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR*, abs/1706.05587.

Chen, L.-C., Yang, Y., Wang, J., Xu, W., Yuille, A. L., 2016. Attention to scale: Scale-aware semantic image segmentation. *CVPR*.

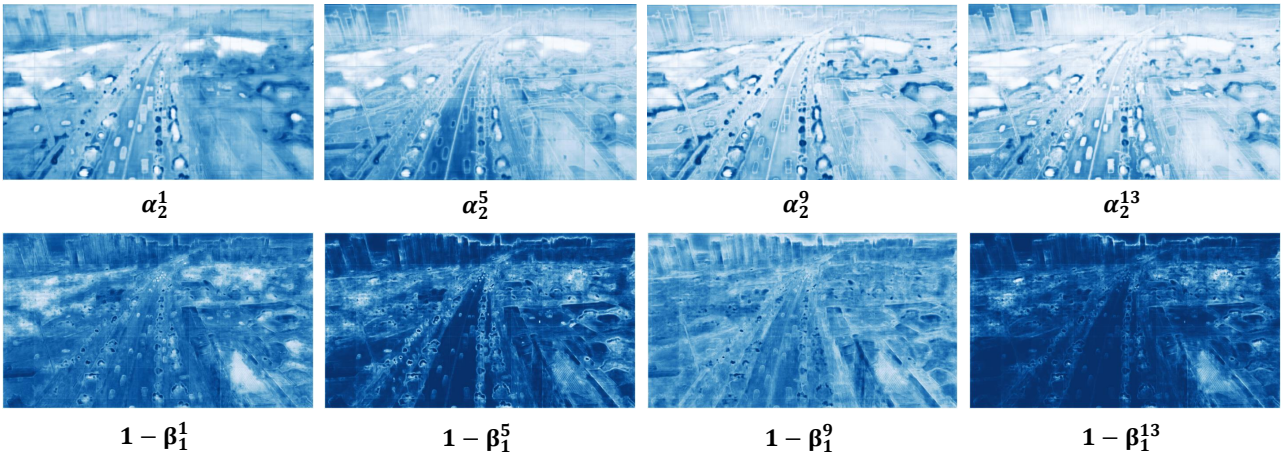

Figure 10. Attention analysis of different directions. The figure shows the attentions for fusing features of scale $0.5\times$ and scale $1\times$. α_2 is for the fine to coarse pathway, while $1-\beta_1$ is for the coarse to fine pathway. Brighter color means higher value. Best visualized with zoom in.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *ECCV*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *CVPR*.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *CVPRW*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR*.

Di Lin, Dingguo Shen, S. S. Y. J. D. L. D. C.-O. H. H., 2019. Zigzagnet: Fusing top-down and bottom-up context for object segmentation. *CVPR*.

Everingham, ., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 111(1), 98–136.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. *CVPR*.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. Ccnet: Criss-cross attention for semantic segmentation. *ICCV*.

Li, Y., Gupta, A., 2018. Beyond grids: Learning graph representations for visual recognition. *NeurIPS*.

Liang, X., Hu, Z., Zhang, H., Lin, L., Xing, E. P., 2018. Symbolic graph reasoning meets convolutions. *NeurIP*.

Liu, W., Rabinovich, A., Berg, A. C., 2015. ParseNet: Looking Wider to See Better. *CoRR*, abs/1506.04579.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *CVPR*.

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M. Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165, 108 - 119.

Nigam, I., Huang, C., Ramanan, D., 2018. Ensemble knowledge transfer for semantic segmentation. *WACV*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS journal of photogrammetry and remote sensing*, 93, 256–271.

Tao, A., Sapra, K., Catanzaro, B., 2020. Hierarchical Multi-Scale Attention for Semantic Segmentation. *CoRR*, abs/1910.12037.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is all you need. *NeurIPS*.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2019. Deep High-Resolution Representation Learning for Visual Recognition. *TPAMI*.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *ECCV*.

Yuan, Y., Chen, X., Wang, J., 2020. Object-contextual representations for semantic segmentation. *ECCV*.

Yuan, Y., Wang, J., 2018. Ocnet: Object context network for scene parsing. *CoRR*, abs/1809.00916.

Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. *BMVC*.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *CVPR*.

Zhao, S., Wang, Y., Yang, Z., Cai, D., 2019. Region mutual information loss for semantic segmentation. *NeurIPS*.