2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222768

SPATIAL-FREQUENCY NETWORK FOR SEGMENTATION OF REMOTE SENSING IMAGES

Tony Zhang and Robert P. Dick

Electrical and Computer Engineering Department
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

We describe a deep learning system for satellite image segmentation. Our CNN model embeds contextual feature dependencies in both spatial and frequency domains. Its Spatial Weighting Module uses a multi-scale pooling layer to represent correlations at longer length scales in the spatial domain. Its Frequency Weighting Module uses frequency-domain information to better discriminate between object classes. Experimental results on the Potsdam dataset demonstrate that our model has a 1.9% higher average F1 accuracy than previous methods.

Index Terms— Remote sensing segmentation, spatial, frequency

1. INTRODUCTION

Remote sensing technologies have enabled the collection of numerous optical satellite images. Identifying land use patterns from satellite imagery is an important problem, and requires that each pixel be precisely classified. Past work used CNNs for for this problem due to their general applicability.

Ding et al. incorporated patch attention to enhance the feature extraction of context and leveraging multi-layer fusion [1]. Yu et al. use multiscale feature extraction via the pyramid pooling module for semantic segmentation on aerial images [2]. Liu et al. used boundary losses to improve edge extraction in satellite images [3].

Satellite image-based land use detection is challenging because aerial images are high-resolution with many diverse objects. In particular, understanding scene context is important to process high-resolution satellite images by extracting the relationships of each pixel with surrounding pixels. This is essential for distinguishing among spatial areas and modeling the relationship between different semantic classes.

Past computer vision research has found that texture information can be used to improve CNN accuracy. Other research found that frequency-domain information can denote texture, noise, and low-level information in images [4, 5]. Generally, sharp edges are best captured using higher frequencies and smooth gradations with lower frequencies. Past work learns identical parameters for all frequency components, whereas learning different parameters for different frequency levels can enhance feature representation.

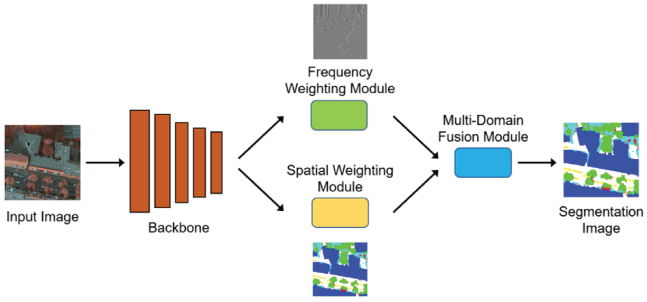

Fig. 1. Architectural overview of satellite image segmentation network, depicting frequency and spatial weighting modules.

Satellite image segmentation requires learning expressive features for intricate scene understanding in both spatial and frequency domains. Learning features at various frequencies reduces confusion among semantic classes. This paper describes a spatial-frequency CNN for aerial segmentation.

We introduce a Frequency Weighted Module to regularize the network using frequency-domain features to improve segmentation. We also develop a Spatial Weighting Module that determines which spatial areas the network should focus on. Finally, we develop a Multi-Domain Fusion Module to aggregate complementary features from the different domains.

2. METHODOLOGY

2.1. Overview

This section describes a remote sensing segmentation model for images (see Figure 1). The input is a remote sensing image of an aerial view of a historic city. The output is a segmentation map indicating land use patterns.

We adopt a ResNet-50 as our backbone network to extract multi-level features from the input image, i.e., f_i $(i = 1, 2, ..., 5)$. The backbone has five stages, each with several residual blocks. We next extract more informative features from both spatial and frequency domains. We define features

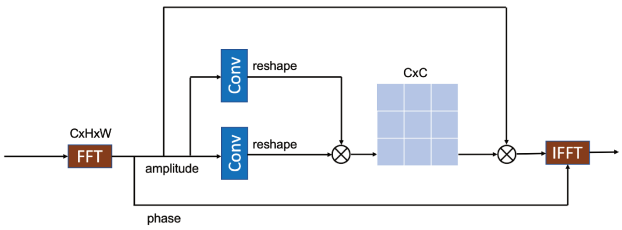

Fig. 2. Frequency Weighting Module (FWM).

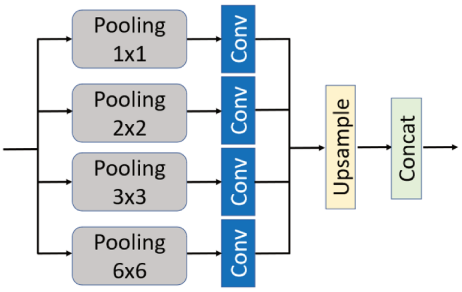

Fig. 3. Spatial Pooling Module used in the Spatial Weighting Module.

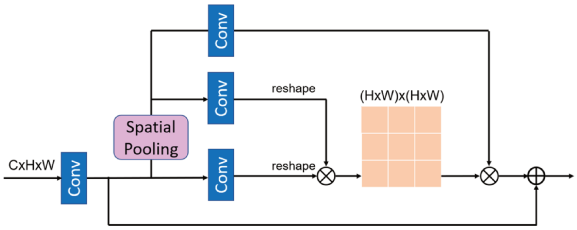

Fig. 4. Spatial Weighting Module (SWM).

obtained from spatial information compared to features obtained from frequency information at different domains.

To enhance contextual information in the spatial domain, we use a Spatial Weighting Module to determine relationships among distant pixels. This can help discern fine-grained spatial areas, especially for confusing areas and boundaries. We then apply a Frequency Weighted Module to encode context based on the Fourier transform. Since remote sensing images contain information about textures and outlines, and also suffer from noise, the model learns to selectively combine useful information from different frequency bands.

2.2. Frequency Weighting Module

Semantic segmentation of satellite should be robust to both intra-class and inter-class variations. Discriminating among many objects in remote sensing images is difficult because decisions are affected by both texture and the context. To solve this problem, we describe a Frequency Weighting Module (FWM) that enhances important information in the extracted features based on frequency.

Remote sensing (satellite) images are typically large and contain fine-grained data. They contains contextual information and it is important to evaluate semantic information at different frequencies to distinguish between object classes. High-frequency features tend to provide texture information, and low-frequency features tend to provide shape information.

As a result, we adjust extracted features in the frequency domain (dynamic frequency modulation), in contrast with past work that treated frequency levels equally. This approach facilitates information flow and learning complementary representations of features. Moreover, this mechanism can help suppress noise in feature representations.

We use the Fourier transform $\mathcal{F}$ to convert the features from the spatial domain to the frequency domain, and the inverse Fourier transform $\mathcal{F}^{-1}$ to convert the features from the frequency domain to the spatial domain. The Fourier transform outputs both amplitude and phase components, and the Fourier transform and inverses are computed independently on each channel of feature maps. In particular, the amplitude component tends to contain low-level statistics of the original image [4, 5].

We now describe the architecture of the Frequency Weighting Module (FWM) shown in Figure 2. We apply the Fourier transform to the output of the backbone network, and feed the amplitude component $F \in R^{C\times H\times W}$ into the FWM. We reshape F to two dimensions $R^{C\times(H\times W)}$ and obtain the weights $W \in R^{C\times C}$ by doing a matrix multiplication of F with F', and then applying the softmax operation:

$$w_{ji} = \frac{\exp(F_i \cdot F_j)}{\sum_{i=1}^{C} \exp(F_i \cdot F_j)}. \quad (1)$$

Afterward, the transpose of the weighting map W is multiplied by the amplitude feature map F. Then we reshape the result to $R^{C\times H\times W}$ to obtain the amplitude-based weighted features. Then we multiply the result by a parameter β and perform an element-wise sum with F to obtain the amplitude-based weighted features:

$$F_j = \beta \sum_{i=1}^{C} (w_{ji} F_i) + F_j. \quad (2)$$

β is a Pytorch parameter learned by backpropagation. Finally, we apply the inverse Fourier transform to the modified amplitude and phase components to obtain the spatial feature maps.

2.3. Spatial Weighting Module

Past work in segmentation used convolutional layers with strong inductive biases toward spatially local relationships

due to their constrained receptive fields. This made it difficult to learn long-range relationships among spatial details. We developed the Spatial Weighting Module (SWM) to overcome this limitation.

Images typically exhibit different attributes at different length scales. To enhance spatial details, we introduce a multi-scale pooling layer (see Figure 3) that uses average pooling operations with different bin sizes to capture contextual information. In our pooling layer, we use bin sizes of 1×1, 2×2, 3×3, and 6×6, and then upsample the pooled feature maps to the original size. After that, we concatenate the feature maps.

In the Spatial Weighting Module (see Figure 4), we feed the output of the backbone network into a 3×3 convolutional layer to obtain F. We then feed F into a multi-scale pooling layer described earlier to obtain F'. We reshape F to two dimensions $R^{(H\times W)\times C}$ and also F' to $R^{C\times(H\times W)}$. We obtain the weights $W \in R^{(H\times W)\times(H\times W)}$ by doing a matrix multiplication of F with its transpose, and then applying the softmax operation:

$$w_{ji} = \frac{\exp(F_i \cdot F'_j)}{\sum_{i=1}^{H\times W} \exp(F_i \cdot F'_j)}. \tag{3}$$

Afterward, the transpose of the weighting map W is multiplied by F'. Then we reshape the result to $R^{C\times H\times W}$, multiply it by a parameter λ, and perform an element-wise sum operation with F to obtain the position-based weighted features:

$$F_j = \lambda \sum_{i=1}^{H\times W} (w_{ji} F'_i) + F_j. \tag{4}$$

λ is a Pytorch parameter learned by backpropagation.

2.4. Multi-Domain Fusion Module

We now describe the Multi-Domain Fusion Module (see Figure 5) used to fuse cross-domain features. This block improves accuracy because it learns the complex relationships among features from different domains. While other methods have directly concatenated different feature vectors from different domains into one long vector, this does not fully extract the complementary information from spatial and frequency features.

We initially perform enhancement of features in both the spatial x_s and frequency x_f domains by boosting features in one domain into the other using a normalized weighted map. Initially, we feed the two kinds of features into a 3×3 conv layer to embed both into the same feature space. Next, we feed both features into a 3×3 conv layer and then a sigmoid activation layer. This produces normalized feature maps for both the spatial and frequency domains, w_s and w_f, respectively.

At this point, we weight the feature map of the spatial domain x_s by using the normalized feature map from the frequency domain w_f, and vice-versa, to represent the correlations between the feature domains. We also add a residual connection to retain the original information of each domain. The output x'_f is the cross-enhanced feature representation from w_s, and the output x'_s is the cross-enhanced feature representation from w_f.

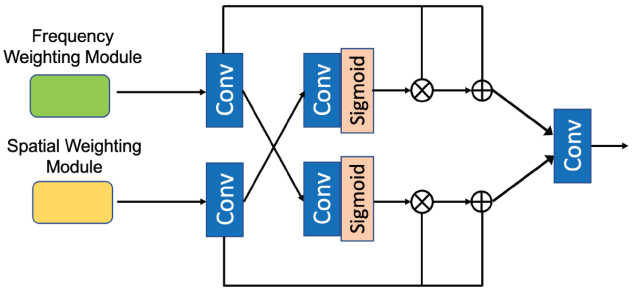

Fig. 5. Multi-Domain Fusion Module.

$$\begin{aligned} x'_f &= x_f + x_f \times w_s \text{ and} \\ x'_s &= x_s + x_s \times w_f. \end{aligned} \tag{5}$$

Afterward, the module integrates the features by concatenating and feeding them into a 3×3 convolutional layer. Finally, we obtain the output, which combines information from multiple domains.

The information in x_s and x_f is complimentary, so the multi-domain fusion module exploits relationships between the different features. The normalized feature maps can be regarded as feature-level attention maps to adaptively weight the feature representations of another domain. This improves leads to more discriminative features and improves segmentation accuracy for remote sensing images.

3. EXPERIMENTAL RESULTS

3.1. Dataset and Implementation Details

We evaluate our segmentation model for remote sensing images using the publicly available Potsdam dataset [12]. It contains 38 true orthophotos (TOPs) of size 6000×6000, consisting of satellite views of a historic city. The ground-truth contains six semantic categories: buildings, trees, cars, low-vegetation, impervious surfaces, and background/clutter. We select 24 RGB images for training and the remaining 14 images for testing. For both the training and testing datasets, we augment the dataset by randomly cropping 30 times for each image to size 224×224 to produce 1,140 images in total.

During the training phase, we set the learning rate to 5e-4, the batch size to 8, and the number of epochs to 100 for model training. Also, we set the momentum parameter to 0.9 and use Adam to optimize the parameters during training. We use the F1 score as a quality measure when comparing with past work.

Table 1. Results of aerial image segmentation with other segmentation methods.

Methods	Impervious surface	Building	Low vegetation	Tree	Car	Mean F1
SegNet [6]	0.551	0.537	0.368	0.308	0.684	0.490
U-Net [7]	0.488	0.518	0.438	0.500	0.702	0.529
RefineNet [8]	0.578	0.587	0.469	0.502	0.746	0.576
LANet [1]	0.641	0.665	0.450	0.511	0.736	0.600
BiSeNetV2 [9]	0.627	0.673	0.458	0.435	0.790	0.597
MACUNet [10]	0.565	0.555	0.445	0.517	0.755	0.567
MA-Net [11]	0.626	0.678	0.479	0.531	0.720	0.607
Proposed	0.599	0.699	0.526	0.548	0.761	0.626

Table 2. Accuracy Implications of Removing Components

Methods	mean F1
Ours w/o FWM + Fusion	0.581
Ours w/o SWM + Fusion	0.611
Ours w/o Fusion	0.618
Ours	0.626

3.2. Model Comparison

We compare our model with past segmentation methods for aerial images in the Potsdam dataset, shown in Table 1. For a fair comparison, we calculate each method's accuracy with the same parameters and the cross-entropy loss function. Also, we use ResNet-50 pretrained on ImageNet as the backbone network for all previous methods.

Our spatial-frequency segmentation network has a higher F1 score than all the alternatives we evaluated. We assess a variety of methods including those containing multi-scale fusion and attention mechanisms. MA-Net is the most recent aerial image-based segmentation method and uses attention mechanisms based on the kernel operation and channel dimension. Our spatial-frequency segmentation network further improves accuracy by 1.9% in mean F1-score over MA-Net because our model has the ability to discern fine-grained spatial regions and discriminate between object classes.

3.3. Ablation Study

We conduct ablation experiments to determine the relative contributions of the proposed design components. Table 2 shows the results for the Potsdam data set. First, we remove the Multi-Domain Fusion Module from the network. Instead of the fusion module, we sum the outputs of the SWM and FWM and feed the output through a 3×3 convolutional layer. We observe that our model with the fusion module outperforms our model without it. Next, we measure the impact of removing the SWM module. From Table 2, we observe that our model with the SWM module outperforms our model without it.

Finally, we remove the FWM from the network. We observe that our model with the FWM outperforms our model without it. This reflects that the Frequency Weighting Module is necessary for improving the accuracy by using frequency levels capture richer features and discriminate between object classes. In summary, each proposed component contributes substantially to the overall accuracy improvement, although if it were necessary to eliminate a module to reduce complexity, SWM would be the best first choice.

4. CONCLUSION

This paper describes a novel deep learning framework that enhances feature representation in both the spatial and frequency domains. Two modules are proposed: 1) the Frequency Weighted Module enhances context information based on the frequency level of local descriptors to refine the segmentation details and 2) the Spatial Weighting Module encodes which pixels of the image are most significant by aggregating spatial context information. Finally, we develop a Multi-Domain Fusion Module to aggregate features from different domains, which can provide important complementary information. Each of these modules contributes to accuracy improvements, and the resulting F1 accuracy exceeds those of the existing approaches we compared against by 1.9%.

5. ACKNOWLEDGEMENTS

This work was supported, in part, by the National Science Foundation through grant CNS-2008151.

6. REFERENCES

[1] Lei Ding, Jing Zhang, and Lorenzo Bruzzone, "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, pp. 5367–5376, 2020.

[2] Bo Yu, Lu Yang, and Fang Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 3252–3261, 2018.

[3] Shuo Liu, Wenrui Ding, Chunhui Liu, Yu Liu, Yufeng Wang, and Hongguang Li, "Ern: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sensing*, vol. 10, pp. 1339, 2018.

[4] Bruce C. Hansen and Robert F. Hess, "Structural sparseness and spatial phase alignment in natural scenes.," *Journal of the Optical Society of America*, vol. 24 7, pp. 1873–85, 2007.

[5] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian, "A fourier-based framework for domain generalization," *2021 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14378–14387, 2021.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2481–2495, 2015.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[8] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5168–5177, 2017.

[9] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051–3068, 2021.

[10] Rui Li, Jianlin Su, Chenxi Duan, and Shunyi Zheng, "Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[11] Rui Li, Shunyi Zheng, Chenxi Duan, and Jianlin Su, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] ISPRS, "2d semantic labeling contest-potsdam," 2022.