Modulation Learning on Fourier-Domain for Road Extraction From Remote Sensing Images

Jing Yang and Huajun Liu, *Member, IEEE*

***Abstract*— Extraction road from remote sensing (RS) images is a challenging topic because of the inhomogeneous intensity, nonconsistent contrast, and very cluttered background of satellite images. Most previous approaches, relying on convolutions or self-attention, are built on the local operation or global modeling on the spatial domain but are difficult to capture weak and continuous road objects. The spectral representation of road image features and modulation learning on it provides a novel long-range-dependent and fine-grained feature representation mechanism. Based on it, we propose a novel road extraction network on RS images, called an adaptive Fourier filtered U-shaped network (AFU-Net) in this letter, which relies on modulation learning on the Fourier domain. The AFU-Net is composed of modulation learner (MoL) basic blocks and follows the pipeline of the classical U-Net model. The basic MoL block includes a global MoL (GML) block for global spectral modulation learning and an attentive MoL (AML) block which contains two parallel layers, i.e., phase-modulated filter (PMF) and magnitude-modulated filter (MMF), for fine-grained spectral modulation on the Fourier spectrum. The experiments on two public datasets, such as Massachusetts roads and DeepGlobe road datasets have shown the outstanding performance of AFU-Net on the metrics of accuracy, precision, recall, and mean intersection over union (mIoU).**

***Index Terms*— Fourier-domain, modulation learning (MoL), remote sensing (RS), road extraction, U-Net.**

I. Introduction

HIGH-QUALITY remote sensing (RS) images provide convenient data sources for geographic information systems (GISs) for their characteristics of high definition, high coverage, and easy access, and road extraction from RS images is an important and fundamental research topic, which is essential for urban road planning, autonomous driving, and geographic mapping. However, extracting road from RS images is a challenging task because of the in-homogeneous intensity, nonconsistent contrast, and very cluttered background of satellite images.

Traditional road extraction methods from RS images are usually based on handcrafted road features (e.g., shape, color, and texture), which are heavily relying on expert knowledge and cannot be generalized to complex scenarios. Since deep learning methods were applied to RS-related tasks, they have obtained overwhelming superiority over traditional methods on road extraction. Road extraction tasks are usually modeled as a segmentation problem. For example, classical convolutional neural networks (CNNs), such as fully convolutional network (FCN) [1], U-Net [2], a deep convolutional encoder-decoder architecture for image segmentation (SegNet) [3], and Deeplabv3 [4] have provided promising results for road extraction from RS images. Some novel solutions, such as D-LinkNet [5] and ConDinet++ [6], adopt additional dilated convolution layers to enlarge receptive field to extract detailed road features in RS images. And RoadDA [7] proposed a novel two-stage unsupervised domain adaptation framework via an adversarial self-training method for road segmentation. Furthermore, introducing self-learning in feature fusion module with self-attention (FFS) [8] has been proven successful to detect damaged buildings in RS images. And feature fusion and attention mechanism in clustering feature constraint multiscale attention network (CFC-Net) [9] can enhance the shadow extraction capability from RS images. The inherent locality of CNNs brings high efficiency, however, it also faces difficulties in capturing long and continuous road objects and weak edges in RS images.

Recently, vision transformer and its variants [10] have shown great potential for object detection and image segmentation due to their superiority in modeling long-range interaction by self-attention. Combining convolution and self-attention is a current popular scheme to simultaneously capture global and local features, for instance, TransUNet [11], SwinUNet [12], and CSwin transformer with dual resolution (DCS)-TransUperNet [13] are successful transformer models for medical image segmentation or road extraction from RS images. But over-stacked self-attention blocks behave like low-pass filters [14] and would suppress high-frequency signals and smooth detailed feature maps, which is essential for weak road object detection.

Besides the transformer, the spectral convolution theorem [15], [16] provides another path to design novel neural architectures with nonlocal receptive field by pointwise update in the spectral domain, and several operators on the Fourier-domain, such as fast Fourier convolution (FFC) [16], global filter network (GFNet) [17], and adaptive Fourier neural operator (AFNO) [18], have shown more flexibility and efficiency on diverse vision-related tasks. Inspired by recent Fourier neural operators [17], [18], a Fourier-domain modulation learning model, called adaptive Fourier filtered U-shaped network (AFU-Net), is built to provide a flexible feature representation method to detect more weak and continuous road features from RS images in this letter. Specifically, the main contributions of this work can be summarized as follows.

1) Proposing the composite modulation learning mechanism on the Fourier domain and implementing a global modulation learner (GML) and an attentive modula-

Manuscript received 18 August 2022; revised 2 November 2022 and 3 December 2022; accepted 19 December 2022. Date of publication 18 January 2023; date of current version 24 February 2023. *(Corresponding author: Huajun Liu.)*

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: liuhj@njust.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3237930

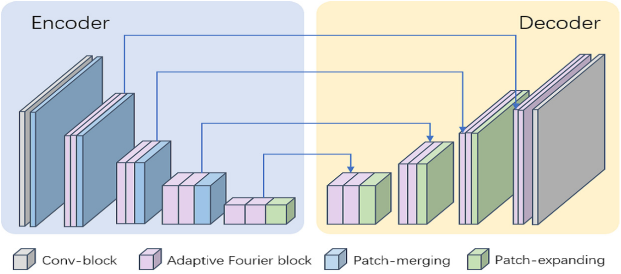

Fig. 1. Architecture of AFU-Net.

tion learner (AML) for long-range dependent and fine-grained road features extraction in RS images.

2) Proposing a parallel phase modulated filter (PMF) and magnitude modulated filter (MMF) in the AML for fine-grained spectral feature extraction.

3) The proposed AFU-Net achieved the state-of-the-art (SOTA) performance on two publicly available datasets for road extraction.

II. Proposed Method

A. Overall Architecture

The architecture of the proposed AFU-Net is presented in Fig. 1. The layout of blocks and configuration of channels and layers are followed by the classical U-Net network [2] and the basic block is our proposed adaptive Fourier block (AF block). On the encoder, a stem block implemented by a 3 × 3 convolution starts the feature encoding from an input image to high-dimensional features. The patch merging [10] and patch expanding [12] blocks are used for downsampling and upsampling to expand and shrink the feature spatial dimension, respectively. Hierarchical feature maps are generated by the four-stage encoder. The resolution in subsequent encoders will be reduced to $(H/2) \times (W/2)$ on the previous $H \times W$ dimension in each stage, and finally reduced to $(H/16) \times (W/16)$ after the fourth stage encoder, meanwhile channel dimension will be expanded from 64, to 128, to 256, and eventually to 512. The symmetrical four-stage decoder is configured with the same AF blocks. The resolution will be gradually expanded to $(H/8) \times (W/8)$, $(H/4) \times (W/4)$, $(H/2) \times (W/2)$, and eventually to the original size after each decoder. At the beginning of each stage, the skip-connection between features in each encoder and upsampled features in the corresponding decoder is used. The standard AF block is shown in Fig. 3(a) which consists of a modulation learner (MoL), a multilayer perceptron (MLP), and layer norms (LNs). The MoL consists of a GML for global spectral representation and an AML for fine-grained spectral feature extraction. Specifically, the AF block can be formulated as

$$\hat{\mathbf{X}}_l = \mathrm{MoL}(\mathrm{LN}(\mathbf{X}_{l-1})) + \mathbf{X}_{l-1} \tag{1}$$

$$\mathbf{X}_l = \mathrm{MLP}\big(\mathrm{LN}\big(\hat{\mathbf{X}}_l\big)\big) + \hat{\mathbf{X}}_l \tag{2}$$

where $\mathrm{LN}(\cdot)$ denotes the layer normalization operator and $\mathbf{X}_l$ is the lth layer feature representation.

B. Spectral Representation for Image Features

The spectral density of the data on the Fourier spectrum is nonuniform with respect to frequency. The spectral power of satellite images with rich boundaries always locates at low-frequency parts [15]. The magnitude and phase spectrum are two vital components after Fourier transformation on image feature maps, and they are frequency characteristic response of image features. Specifically, the noisy background and the continuous road have different spectral features. Intuitively, as for the image feature spectrum, its magnitude carries much of the intensity information, such as the contrast or the difference between the brightest and darkest peaks of images, and its phase crucially reflects the spatial structure of the image. Fig. 2 shows the original image, its corresponding magnitude spectrum, and phase spectrum. The spectrum is a complex-value feature $\mathbf{X}_F \in \mathbb{C}^{C\times H\times W}$, and its phase and magnitude components are calculated as follows:

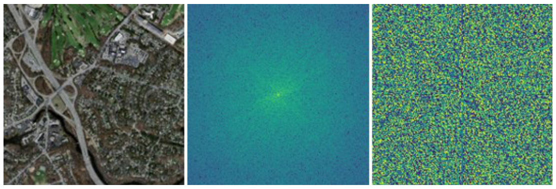

Fig. 2. Original image, magnitude, and phase spectrum.

$$\mathrm{Pha}(\mathbf{X}_F) = \mathrm{Arctan}(\mathbf{X}_F.\mathrm{img}/\mathbf{X}_F.\mathrm{real}) \tag{3}$$

$$\mathrm{Mag}(\mathbf{X}_F) = \mathrm{Abs}(\mathbf{X}_F.\mathrm{img}, \mathbf{X}_F.\mathrm{real}) \tag{4}$$

where $\mathbf{X}_F.\mathrm{img} \in \mathbb{R}^{C\times H\times W}$ is the imaginary part and $\mathbf{X}_F.\mathrm{real} \in \mathbb{R}^{C\times H\times W}$ is the real part of complex number. $\mathrm{Arctan}(\cdot)$ is the arc-tangent function, and $\mathrm{Abs}(\cdot)$ computes the absolute value. $\mathrm{Pha}(\cdot)$ and $\mathrm{Mag}(\cdot)$ are phase and magnitude functions that transform spectrum from complex domain to real domain.

C. Global MoL

The GML is a large-kernel convolution without bias on the Fourier domain, which aims to learn to modulate the spectrum on entire frequency space to capture coarse-grained image features. Given a feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ in real space, we first conduct 2-D FFT to convert $\mathbf{X}$ to the frequency domain $\mathbf{X}_F \in \mathbb{C}^{C\times H\times W}$. A large-kernel learnable filter $\mathbf{K} \in \mathbb{C}^{C\times H\times W}$ is utilized for modulation learning and learning global features by element-wise multiplication with $\mathbf{X}_F$ as

$$\mathbf{X}'_F = \mathbf{K} \odot \mathbf{X}_F \tag{5}$$

where $\odot$ is the complex-valued element-wise multiplication and $\mathbf{X}'_F \in \mathbb{C}^{C\times H\times W}$ is the preliminary coarse-grained spectral features modulated by GML. It has been proven that GML can capture long-term spatial dependencies in the frequency domain with log-linear time complexity [17].

D. Attentive MoL

Inspired by the attention architecture in convolutional block attention module (CBAM) [19], which generates two attention maps along the channel and spatial dimension, respectively, we propose an AML, specifically including a PMF and an MMF to generate phase-modulated masks $\mathbf{M}_{\mathrm{PMF}} \in \mathbb{R}^{C\times 1\times 1}$ and magnitude-modulated masks $\mathbf{M}_{\mathrm{MMF}} \in \mathbb{R}^{1\times H\times W}$ for fine-grained spectral feature extraction. It is well-known that sparse and salient components of the spectrum always locate at the high-frequency parts and reflect image's detailed features. Attentive modulation on salient and sparse spectral components can capture its fine-grained features. Based on the distribution characteristics on spectral magnitude and phase, two parallel attentive modulation filters along phase and magnitude dimension are proposed for fine-grained feature

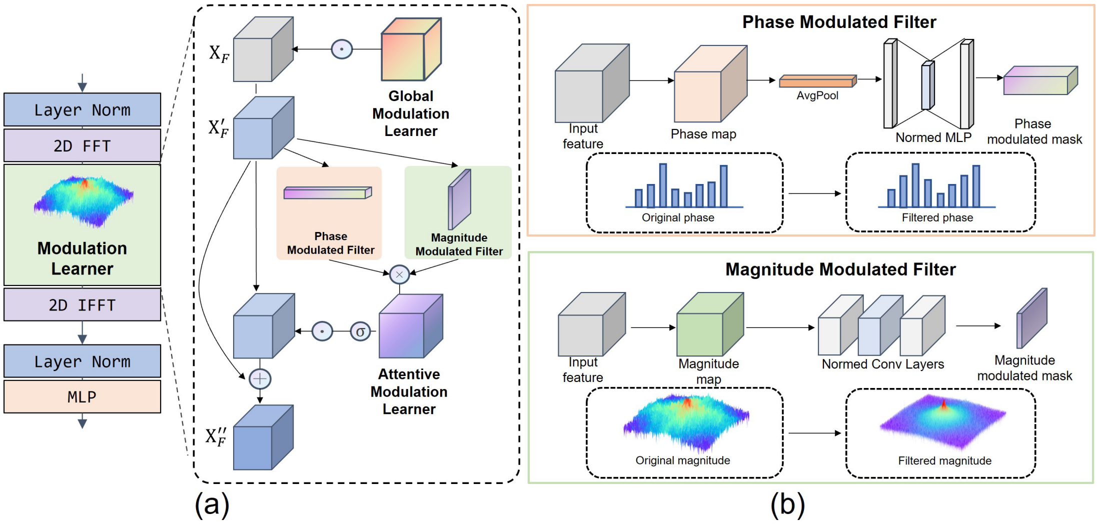

Fig. 3. Structure of (a) AF block and (b) PMF and MMF.

extraction. The diagram of AML is shown in Fig. 3(a), which can be formulated as

$$\mathbf{M}_{\text{Att}} = \mathbf{M}_{\text{PMF}} \otimes \mathbf{M}_{\text{MMF}} \tag{6}$$

$$\mathbf{X}''_F = \mathbf{X}'_F + \mathbf{X}'_F \odot (\sigma(\mathbf{M}_{\text{Att}})) \tag{7}$$

where $\odot$ denotes complex-valued element-wise multiplication, $\sigma(\cdot)$ denotes sigmoid activation function, $\otimes$ denotes tensor broadcasting mechanism, the output of $\mathbf{M}_{\text{PMF}}$ and $\mathbf{M}_{\text{MMF}}$ will be converted to $\mathbf{M}_{\text{Att}} \in \mathbb{R}^{C\times H\times W}$. Thus, $\mathbf{X}''_F \in \mathbb{C}^{C\times H\times W}$ is the fine-grained feature modulated by AML.

1) Phase-Modulated Filter: The phase of the spectrum varies on different channels, drawing the detailed difference of feature maps, which is vital for extremely thin boundary feature extraction. The PMF was proposed to generate a mask $\mathbf{M}_{\text{PMF}}$ on the phase component of the spectrum to extract the thin and weak boundaries in an efficient way. Specifically, we first convert the complex-valued feature maps into its phase components, then apply an average-pooling (AvgPool) to obtain the global phase feature with the shape of $C \times 1 \times 1$, and an MLP layer is used to learn the phase relationship between each channel. Finally, a PMF matrix $\mathbf{M}_{\text{PMF}}$ can be generated.

The diagram of the PMF is shown in the upper part of Fig. 3(b), and the learnable filter PMF$(\cdot)$ is concisely defined as

$$\mathbf{M}_{\text{PMF}} = \text{PMF}(\mathbf{X}'_F) \triangleq \text{MLP}(\text{BN}(\text{AvgPool}(\text{Pha}(\mathbf{X}'_F)))) \tag{8}$$

where Pha$(\cdot)$ denotes the phase function in (3) and MLP$(\cdot)$ and BN$(\cdot)$ denote an MLP layer and BatchNorm operator, the ReLU activation function follows the BN function.

2) Magnitude-Modulated Filter: The MMF focuses on high-energy regions mostly contributed by detailed features, such as texture, edge, and contour. Specifically, we first convert the complex-values feature maps into a magnitude spectrum. The features in the frequency domain are more sensitive since each point represents global information and we apply the BatchNorm operation to balance the weights. A large-kernel convolution is applied to modulate energy distribution in the magnitude spectrum. Finally, an MMF matrix $\mathbf{M}_{\text{MMF}}$ can be generated.

The diagram of MMF is shown in the lower part of Fig. 3(b). And the learnable filter MMF$(\cdot)$ is calculated as

$$\mathbf{M}_{\text{MMF}} = \text{MMF}(\mathbf{X}'_F) \triangleq \text{Conv}(\text{BN}(\text{Mag}(\mathbf{X}'_F))) \tag{9}$$

where Mag$(\cdot)$ denotes the magnitude function in (3), Conv$(\cdot)$ represents a convolution operation with 7 × 7 kernel size and the parameters of padding and stride are set to 3 and 1, respectively, to keep the same resolution as the input features, and a ReLU activation function follows the BN function.

The modulated features based on (6) are inversely transformed from the frequency domain to the spatial domain by 2-D inverse fast Fourier transform (IFFT). The GML and AML together contribute to our proposed MoL, and all modulation filtering parameters are learned in the frequency domain.

E. Loss Function

Considering the small proportion of road objects in RS images, we adopt a weighted loss function to supervise the training process. A standard cross-entropy loss function L_{Cross} and dice loss function [20] L_{Dice} are defined as the two components of the loss function. The loss function is defined as

$$L = L_{\text{Cross}} + \alpha L_{\text{Dice}} \tag{10}$$

where α is the weight coefficient of two components.

III. Experiments

A. Datasets

Our proposed AFU-Net model is trained and evaluated on two public datasets: 1) Massachusetts roads dataset [21] and 2) DeepGlobe road dataset [22]. Massachusetts roads dataset consisted of aerial images of the state of Massachusetts. It contains 1108 images for training, 49 images for testing, and 14 images for validation. The resolution of each image is 1500 × 1500. DeepGlobe dataset comes from the 2018 conference on computer vision and pattern recognition (CVPR) Road Extraction Challenge, consisting of images captured over Thailand, Indonesia, and India, covering a variety of urban, suburban, and rural regions. This dataset is randomly divided into 4980 images for training, 624 for testing, and 626 for validation. The resolution of each image is 1024 × 1024.

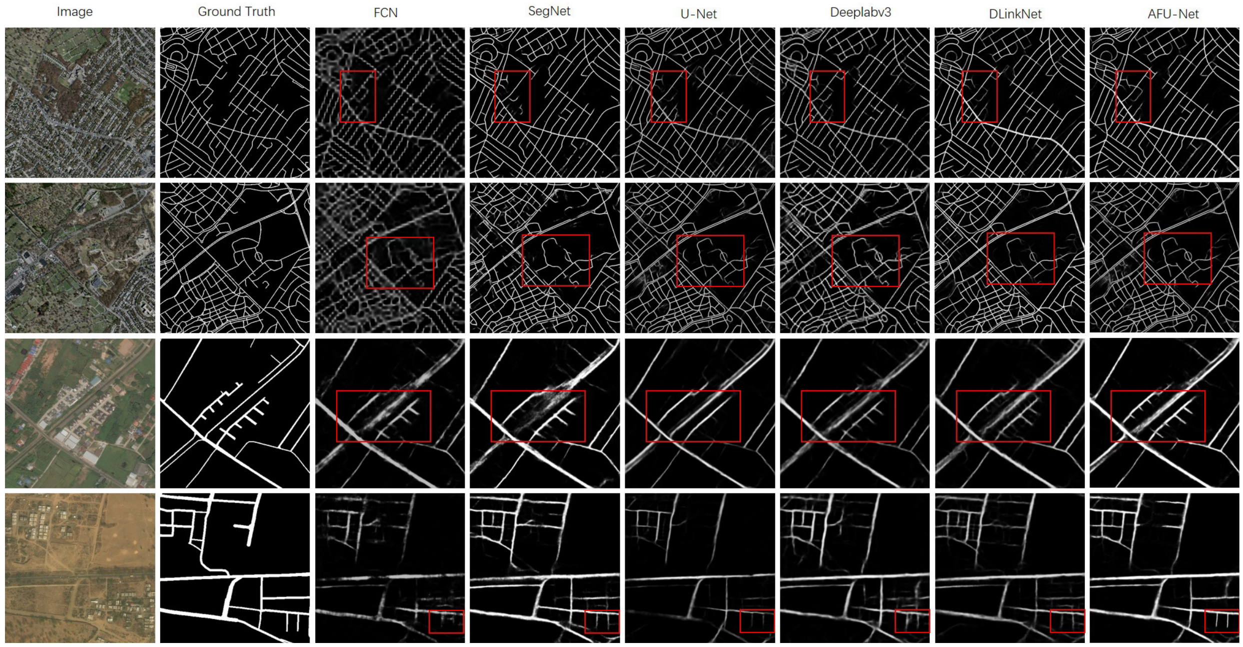

Fig. 4. Visualized results on two datasets. The first two rows of images are urban traffic roads selected from the Massachusetts roads dataset, and the last two rows of pictures are rural roads selected from the DeepGlobe dataset.

TABLE I

RESULTS ON MASSACHUSETTS ROADS DATASET

Model	Accuracy(%)	Precision(%)	Recall(%)	mIoU(%)
FCN[1]	74.61	60.43	68.89	73.20
SegNet[3]	75.23	64.41	69.80	74.81
U-Net[2]	78.04	63.23	70.32	76.39
TransUNet[11]	78.10	68.35	71.23	77.89
Attention-UNet[23]	76.31	71.32	73.41	78.13
SwinUNet[12]	78.01	68.25	71.31	77.64
Deeplabv3[4]	77.82	68.41	72.48	77.43
D-LinkNet[5]	77.83	67.16	71.73	77.23
DCS-TransUperNet[13]	77.94	69.52	73.48	77.74
ConDinet++[6]	78.14	72.05	74.71	78.88
AFU-Net (Ours)	**78.93**	**76.83**	**75.17**	**80.32**

TABLE II

RESULTS ON DEEPGLOBE DATASET

Model	Accuracy(%)	Precision(%)	Recall(%)	mIoU(%)
FCN[1]	92.05	63.71	61.54	42.45
SegNet[3]	91.54	62.03	60.21	41.85
U-Net[2]	97.12	82.28	71.20	61.02
TransUNet[11]	93.21	64.45	62.34	46.10
Attention-UNet[23]	96.79	79.89	79.47	70.31
SwinUNet[12]	97.01	82.35	71.54	61.43
Deeplabv3[4]	95.87	65.83	64.48	48.31
D-LinkNet[5]	98.32	83.33	72.29	62.87
Deep FR TransNet[24]	98.70	87.30	81.15	72.44
AFU-Net (Ours)	**98.81**	**87.86**	**81.70**	**74.14**

B. Implementation Details

The Massachusetts roads dataset is first reshaped to 1024 × 1024. When training, each image in both dataset is cropped to sub-images with a fixed size of 512 × 512 under the stride of 512 pixels, meanwhile image transformations, including rotation, scaling, flipping, and mirroring, are also applied on both dataset for data augmentation. And when inference, output results will be mosaiced to the original size. Our proposed model is implemented by PyTorch 1.8 on a single NVIDIA RTX TITAN graphics processing unit (GPU). The training batch size is set to 4 and the stochastic gradient descent (SGD) optimizer is utilized as the learning rate scheme. The initial learning rate is $\lambda = 10^{-2}$ and linearly decayed to $\lambda = 10^{-6}$ after 160 000 iterations.

C. Evaluation Metrics

In this letter, in order to evaluate the performance of the proposed network in road extraction tasks, the widely accepted metrics as in [24]: precision, recall, F-score, and mean intersection over union (mIoU) are used for performance evaluation. The pixels in the segmentation image are divided into true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Then four metrics are calculated as

$$\text{mIoU} = \frac{1}{2} \times \left(\frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP} + \text{FN}} \right) \tag{11}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

D. Comparison With Other Deep Learning Models

The visualized results on road extraction segmentation are shown in Fig. 4. It can be seen that AFU-Net can successfully extract extremely weak and thin roads from two datasets. The area marked by red rectangles is an instance of significant improvement in road segmentation.

The comparison experiments of AFU-Net with other deep learning models are conducted. The baseline models, such as FCN [1], SegNet [3], Deeplabv3 [4], U-Net [2], TransUNet [11], Attention-UNet [23], SwinUNet [12], and D-LinkNet [5] are compared on the both benchmarks. In addition, some recent works have been added to our comparison experiments, such as DCS-TransUperNet [13], ConDinet++ [6],

TABLE III
ABLATION STUDY ON MoL COMPONENTS

GML	PMF	MMF	Accuracy↑	Precision↑	Recall↑	mIoU↑
✓			96.89	86.53	80.82	71.33
✓	✓		97.92	87.12	81.24	72.67
✓		✓	97.63	87.08	81.12	72.64
✓	✓	✓	**98.81**	**87.86**	**81.70**	**74.14**

and Deep feature-review (FR) TransNet [24], respectively. The performance comparison is listed in Tables I and II.

As shown in Table I on the Massachusetts dataset, AFU-Net achieved the highest recognition accuracy of 78.93%, precision of 76.83%, recall of 75.17%, and mIoU of 80.32, which obtained a 0.79%, 4.78%, 0.46%, and 1.44% gain compared with the SOTA method, i.e., ConDinet++ [6]. It means AFU-Net can accurately classify road pixels in RS images. Compared to the recentest transformer model DCS-TransUperNet [13], AFU-Net built on modulation learning on frequency domain demonstrates its superiority over the self-attention structure and outperforms it by 2.58% on mIoU.

On the DeepGlobe dataset, it can be concluded from Table II that AFU-Net shows enough competitiveness. Our AFU-Net achieves the best accuracy of 98.81%, precision of 87.86%, recall of 81.70%, and mIoU of 74.14%, exceeding the SOTA method (i.e., deep FR TransNet [24]) by 1.7% on mIoU. In contrast to dilation convolutions and self-attentions, frequency domain operation has a larger receptive field based on the properties of the Fourier transform. Experimental results demonstrate that AFU-Net has better representation capability. It can be concluded that our proposed AFU-Net outperforms other comparable methods and can generate accurate feature maps in RS images with a cluttered background.

E. Ablation Study

To evaluate the performance of the AF block, we apply the classical U-Net as the baseline to perform an ablation study on the DeepGlobe dataset. Three components, i.e., GML, PMF and MMF in AML in our model are taken into account to verify their impacts to model performance. The experimental results are shown in Table III, which demonstrates that three components are critical to improving the performance of the AFU-Net model.

Specifically, ablation study results are listed in Table III. We gradually add PMF and MMF to the GML plus baseline individually or cooperatively. After introducing PMF and MMF, they can boost the model by 1.34% and 1.31% on mIoU individually, and both modules can boost the baseline by 2.81% on mIoU. It demonstrates that attentive modulation on the frequency domain from phase and magnitude can improve performance greatly. We also explored how to make choice on the weight coefficient α in the loss function, when α is set to 1–4, the mIoU, we can obtain 73.92, 74.01, 74.14, and 74.03, respectively, on DeepGlobe dataset, so $\alpha = 3$ is regarded as an optimized coefficient in our tasks.

IV. CONCLUSION

In this letter, we propose AFU-Net based on modulation learning on Fourier domain for road extraction in RS images, which follows the classical U-Net structure. A composite involving GML and AML can effectively extract fine-grained road features from RS images. Experimental results on two public benchmarks show that AFU-Net achieved more excellent performance than most previous deep learning models. In the future, we will extend AFU-Net to more complex road extraction scenarios and try to achieve lightweight and more efficient versions.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. MICCAI*, 2015, pp. 234–241.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder–decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2015.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with atrous separable convolution for semantic image segmentation,” in *Proc. ECCV*, 2018, pp. 801–818.

[5] L. Zhou, C. Zhang, and M. Wu, “D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[6] K. Yang, J. Yi, A. Chen, J. Liu, and W. Chen, “ConDinet++: Full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[7] L. Zhang, M. Lan, J. Zhang, and D. Tao, “Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609413.

[8] Y. Xie et al., “Damaged building detection from post-earthquake remote sensing imagery considering heterogeneity characteristics,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, Art. no. 4708417.

[9] Y. Xie et al., “Clustering feature constraint multiscale attention network for shadow extraction from remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4705414.

[10] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted Windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[11] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

[12] H. Cao et al., “Swin-UNet: UNet-like pure transformer for medical image segmentation,” 2021, *arXiv:2105.05537*.

[13] Z. Zhang, C. Miao, C. Liu, and Q. Tian, “DCS-TransUperNet: Road segmentation network based on CSwin transformer with dual resolution,” *Appl. Sci.*, vol. 12, no. 7, p. 3511, Mar. 2022.

[14] P. Wang, W. Zheng, T. Chen, and Z. Wang, “Anti-oversmoothing in deep vision transformers via the Fourier domain analysis: From theory to practice,” in *Proc. ICLR*, 2022, pp. 1–24.

[15] O. Rippel, J. Snoek, and R. P. Adams, “Spectral representations for convolutional neural networks,” in *Proc. NIPS*, 2015, pp. 1–9.

[16] L. Chi, B. Jiang, and Y. Mu, “Fast Fourier convolution,” in *Proc. NIPS*, 2020, pp. 4479–4488.

[17] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, “Global filter networks for image classification,” in *Proc. NIPS*, 2021, pp. 980–993.

[18] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, “Efficient token mixing for transformers via adaptive Fourier neural operators,” in *Proc. ICLR*, 2022, pp. 1–15.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. ECCV*, 2018, pp. 3–19.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[21] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto (Canada), 2013.

[22] I. Demir et al., “DeepGlobe 2018: A challenge to parse the Earth through satellite images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[23] O. Oktay et al., “Attention U-Net: Learning where to look for the pancreas,” 2018, *arXiv:1804.03999*.

[24] Z. Ge, Y. Zhao, J. Wang, D. Wang, and Q. Si, “Deep feature-review transmit network of contour-enhanced road extraction from remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.