Frequency-Guided Spatial Adaptation for Camouflaged Object Detection

Shizhou Zhang, Dexuan Kong, Yinghui Xing, *Member, IEEE*, Yue Lu, Lingyan Ran, Guoqiang Liang, Hexu Wang, and Yanning Zhang, *Senior Member, IEEE*

***Abstract*—Camouflaged object detection (COD) aims to segment camouflaged objects which exhibit very similar patterns with the surrounding environment. Recent research works have shown that enhancing the feature representation via the frequency information can greatly alleviate the ambiguity problem between the foreground objects and the background. With the emergence of vision foundation models, like InternImage, Segment Anything Model etc, adapting the pretrained model on COD tasks with a lightweight adapter module shows a novel and promising research direction. Existing adapter modules mainly care about the feature adaptation in the spatial domain. In this paper, we propose a novel frequency-guided spatial adaptation method for COD task. Specifically, we transform the input features of the adapter into frequency domain. By grouping and interacting with frequency components located within non overlapping circles in the spectrogram, different frequency components are dynamically enhanced or weakened, making the intensity of image details and contour features adaptively adjusted. At the same time, the features that are conducive to distinguishing object and background are highlighted, indirectly implying the position and shape of camouflaged object. We conduct extensive experiments on four widely adopted benchmark datasets and the proposed method outperforms 26 state-of-the-art methods with large margins. Code will be released.**

***Index Terms*—Camouflaged object detection, frequency-guided, pretrained foundation model, spatial adaptation.**

Received 8 May 2024; revised 26 June 2024; accepted 9 July 2024. Date of current version 17 January 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62101453 and Grant 62201467, in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313088, in part by the Innovation Capability Support Program of Shaanxi under Grant 2024ZC-KJXX-043, in part by the China Postdoctoral Science Foundation under Grant 2022TQ0260 and Grant 2023M742842, in part by the Fundamental Research Funds for the Central Universities under Grant HYGJZN202331, and in part by the Natural Science Basic Research Program of Shaanxi Province under Grant 2022JC-DW-08. The associate editor coordinating the review of this article and approving it for publication was Prof. Qianqian Xu. *(Shizhou Zhang and Dexuan Kong contributed equally to this work.) (Correspondence author: Yinghui Xing.)*

Shizhou Zhang, Dexuan Kong, Yue Lu, Lingyan Ran, Guoqiang Liang, and Yanning Zhang are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China.

Yinghui Xing is with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China (e-mail: xyh_7491@nwpu.edu.cn).

Hexu Wang is with the School of Information and Technology, Northwest University, Xi'an 710127, China, and also with the Xi'an Key Laboratory of Human-Machine Integration and Control Technology for Intelligent Rehabilitation, Xijing University, Xi'an 710123, China.

Digital Object Identifier 10.1109/TMM.2024.3521681

I. Introduction

CAMOUFLAGED object detection (COD), which has wide downstream applications such as medical segmentation [1], [2] and recreational art [3], aims to segment the objects which are perfectly embedded in their surrounding environment. Recent years have witnessed the great progress of COD, it remains a challenging task due to the low contrast appearances between the concealed objects and the background. In addition, the semantic categories of the objects lie between a wide range from naturally camouflaged objects such as mammals or insects hiding themselves from their predators, to artificially camouflaged objects such as soldiers on the battlefields or human body painting arts. The diverse types of objects with various shapes, sizes and textures further increases the difficulties of the COD task.

From one hand, some recent methods try to design progressively coarse to fine feature enhancement process [4], [5] or to utilize extra edge information [6] to locate accurate boundaries from the spatial/RGB domain information of an image. While other works propose to introduce clues in frequency domain [7], [8], [9], as the frequency enhanced features are more discriminative between the concealed objects and background.

From the other hand, with the emergence of large scale pretrained vision foundation models, such as InternImage [10] and Segment Anything Model (SAM) [11], a promising research paradigm which is prevalent on almost all vision tasks is that adapting the foundation model on the downstream tasks with a small portion of extra trainable parameters or architectures, e.g. prompts and adapters, while the parameters of the pretrained model kept frozen. Existing task-specific adapters broadly fall into three categories: series adapter [12], parallel adapter [12] and LoRA [13]. To introduce the image-related inductive biases into the pretrained ViT model for pixel-wise dense prediction tasks, [14] proposed a specific parallel ViT-Adapter to further aggregate multi-scale context. Current adapters are devised to compensate the features or weights all from the spatial domain. However, crucial clues for the downstream COD task, such as subtle variations in textures and patterns, may not be easily observed in the spatial domain but can be revealed by the unique spectral characteristics in the frequency domain. Therefore, adapting the pretrained foundation model from the spatial domain alone can not take the full advantage of the merits brought by the frequency domain information which is especially required for the COD task.

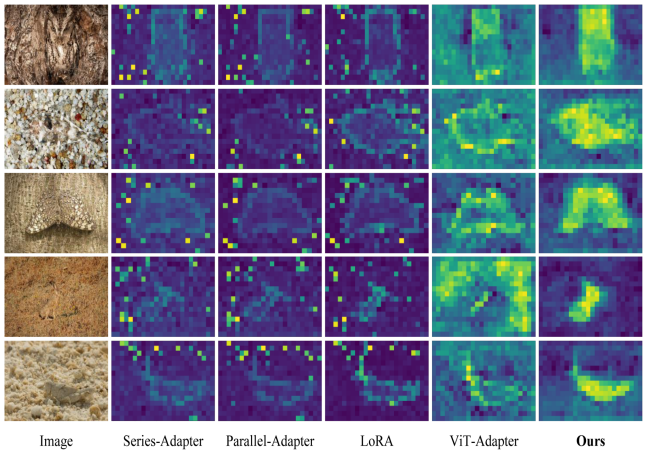

Fig. 1. Visualization and comparison of feature maps obtained after adapter tuning on COD task.

In this paper, we propose a novel adaptation model named as frequency-guided spatial adaptation network (FGSA-Net) for COD task. Firstly, we devise a frequency-guided spatial attention (FGSAttn) module by transforming the input features of the adapter into frequency domain. Then by grouping and interacting with frequency components located within non overlapping circles in the spectrogram, different frequency components are enhanced or weakened, making the intensity of image details and contour features adaptively adjusted.

Based on the FGSAttn module and the multi-scale context aggregation as in ViT-Adapter, we further propose a Frequency-Based Nuances Mining (FBNM) module which aims to mining subtle differences between foreground and background, and a Frequency-Based Feature Enhancement (FBFE) module which extracts and fuses multi-scale features containing general knowledge of the pretrained model and adaptation components learned from the new data of downstream COD task. As can be seen from Fig. 2, the FBNM module is inserted after the patch embedding layer and the FBFE module is inserted into the pretrained ViT backbone model after each K layers. During training, only the parameters of FBNM and FBFE modules are optimized while the parameters of pretrained ViT model are kept frozen. With only about 7% tunable parameters (over the total parameters of the pretrained model), our proposed FGSA-Net achieves state-of-the-art performances on four widely adopted benchmark datasets of COD and outperforms the spatial adaptation counterparts with a large margin. Fig. 1 illustrates the obtained feature maps after adaptation, it can be seen that our proposed novel adaptation mechanism clearly concentrate more on the concealed objects compared with other spatial adaptation methods.

To summarize, the contributions of this paper are threefolds,

- We propose a novel frequency-guided spatial adaptation network, which combines the advantage of general knowledge of vision foundation model and task-specific features learned from the new data of downstream COD task.
- A frequency-guided spatial attention module is devised to adapt the pretrained foundation model from spatial domain while guided by the adaptively adjusted frequency components to focus more on the camouflaged regions.
- The proposed method greatly outperforms the baseline methods and achieves state-of-the-art performances on four widely adopted COD benchmark datasets.

The rest of the paper is organized as follows, Section II reviews the relevant works of our paper. Section III elaborates each component of the proposed method. In Section IV, we conduct thorough experiments on four widely-used benchmark datasets to verify the superiority of our method and further analyze the effectiveness of each component. Finally, we draw the conclusion of the paper in Section V.

II. Related Work

A. Camouflaged Object Detection

Numerous efforts have been undertaken in the field of camouflaged object detection [4], [5], [6], [15], [16], [17], [18], [19], [20], [21], [22]. In order to obtain accurate boundary, [15] promote the model to generate features that highlight object structure for accurate boundary localization of camouflaged objects. Ref. [6] decouple an image into two feature maps and recurrently reason their high-order relations through graphs for roughly locating the target and accurately capturing its boundary details. Ref. [21] combine probabilistic-derived uncertainty and deterministic-derived edge information to accurately detect concealed objects. To capture rich features of camouflaged objects, [18] integrate and fuse multi-level image features to yield multi-scale representations for exploiting rich global context information. Ref. [5] iteratively refine low-resolution representations by high-resolution features to extract high-resolution texture details and avoid the detail degradation. Ref. [22] leverage the spatial organization of textons in the foreground and background regions as discriminative cues for camouflaged object detection. Additionally, some recent works [7], [8] investigate that clues in frequency domain can help the feature enhancement of concealed objects.

B. Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning aims to adapt pretrained models to downstream tasks by inserting a few learnable parameters [13], [23], [24], [25], [26], [27]. Ref. [27] propose lightweight adapters for Transformers [28] in the filed of NLP. [23] introduce learnable tokens (i.e. prompts) into Vision Transformers [29]. Ref. [13] inject trainable rank decomposition matrices into each layer of the Transformer architecture. All the methods only update the introduced small number of parameters (i.e. adapters, prompts, *etc.*) while keep the pretrained parameters fixed. Consequently, the training process requires much less memory and computation costs than fine-tuning the whole model. However, existing adapters deal with the feature adaptation problem from the spatial domain alone.

C. Frequency-Based Methods

Since the features of camouflaged objects and the background are more discriminative in the frequency domain, a line of approaches [7], [8], [9], [30], [31], [32], [33] dig frequency clues for camouflaged object detection or other tasks to enhance the

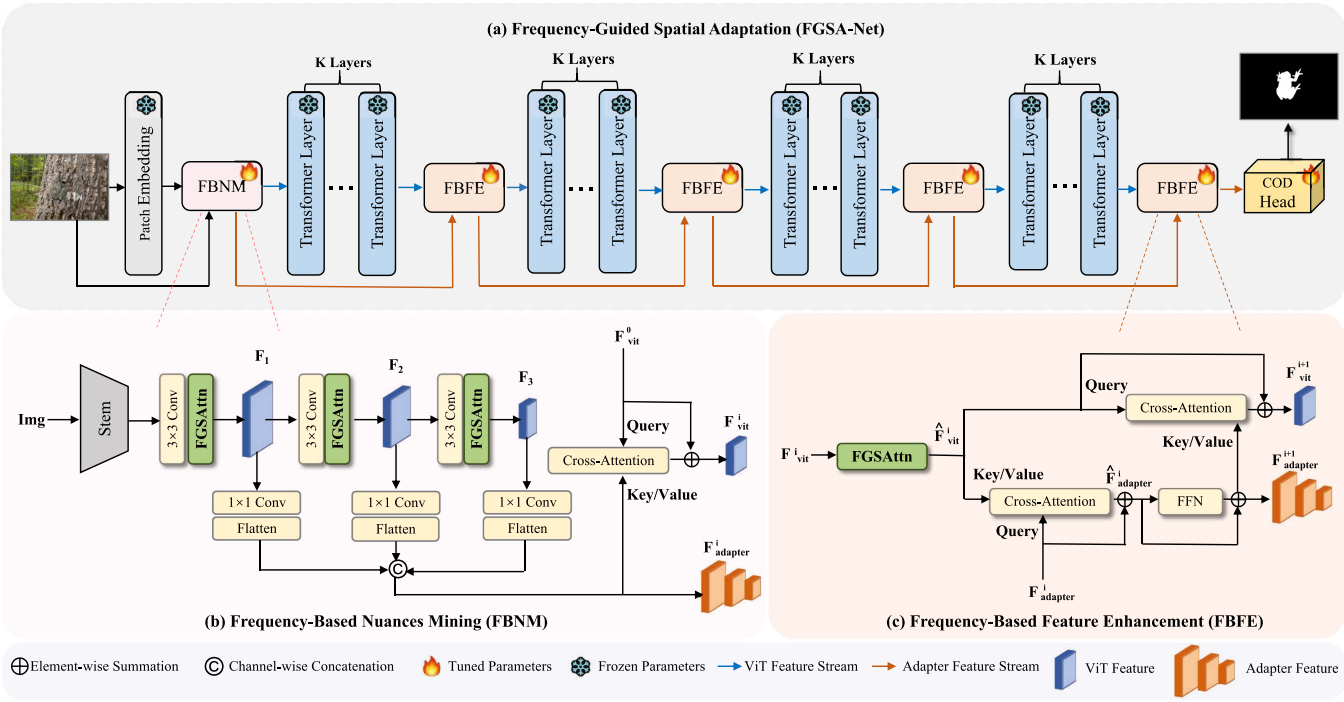

Fig. 2. Overall framework of our proposed FGSA-Net. (a) The main architecture. (b) Frequency-based nuances mining module. (c) Frequency-based feature enhancement module.

feature representation. Ref. [7] adopt the offline discrete cosine transform to extract frequency features, and then fuse the features from RGB domain and frequency domain. Ref. [9] aggregate multi-scale features from a frequency perspective and enhance the features of the learned important frequency components. Ref. [8] utilize the octave convolution [34] in the frequency perception module for coarse positioning, and combine high-level features with shallow features to achieve the detailed correction of the camouflaged objects. Different from the attention modules only performed in RGB domain, we exploit the spatial adapter while guided with frequency domain information, which is more helpful to distinguish between camouflaged objects and the background.

III. Methodology

A. Overview

As a typical low contrast structural segmentation task, COD methods require not only low-level structural details but also global context information. However, available adaptation models like ViT-adapter [14] and SAM-adapter [35], only consider global context information in spatial domain, limiting their ability to locate subtle differences between foreground and background. To accurately represent refined structure of camouflaged objects, we resort to extract and enhance detail information from the frequency perspective to design a frequency-guided spatial adapter (FGSA-Net). The overall architecture is shown in Fig. 2(a), including a large pretrained ViT model, a lightweight adapter module consist of frequency-based nuances mining (FBNM) and frequency-based feature enhancement (FBFE), as well as a detection head for COD. Specifically, as shown in Fig. 3, we devise the frequency-guided spatial attention (FGSAttn) module to concentrate more on the concealed objects by dynamically adjusting the frequency components. Based on the FGSAttn, as detailed in Fig. 2(b) and (c), two elaborate modules, i.e., FBNM and FBFE, responsible for subtle feature extraction and enhancement, are proposed to serve as the adapter. Among them, the FBNM module aims to receive original input images and serialized tokens to capture prior knowledge of the subtle differences between the foreground and background. Then, we evenly split the transformer layers of ViT model into M groups, each of which contains K layers. The FBFE module is inserted into the ViT model after each group and performs interaction operations on the general knowledge from the pretrained ViT branch and the task-specific knowledge from the adapter branch to recalibrate the feature distribution. Finally, the hierarchical features output by the last FBFE module are input into the detection head to generate more refined and accurate prediction map.

B. Frequency-Guided Spatial Attention Module

The detailed architecture of frequency-guided spatial attention (FGSAttn) module is illustrated in Fig. 3. The input feature $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ are processed by average-pooling and max-pooling operations along the channel dimension, and the pooled features are then combined through an element-wise summation to obtain a single channel most distinctive features $\mathbf{F}_g \in \mathbb{R}^{H\times W\times 1}$. It can be formulated as:

$$\mathbf{F}_g = AvgPool(\mathbf{F}) + MaxPool(\mathbf{F}). \tag{1}$$

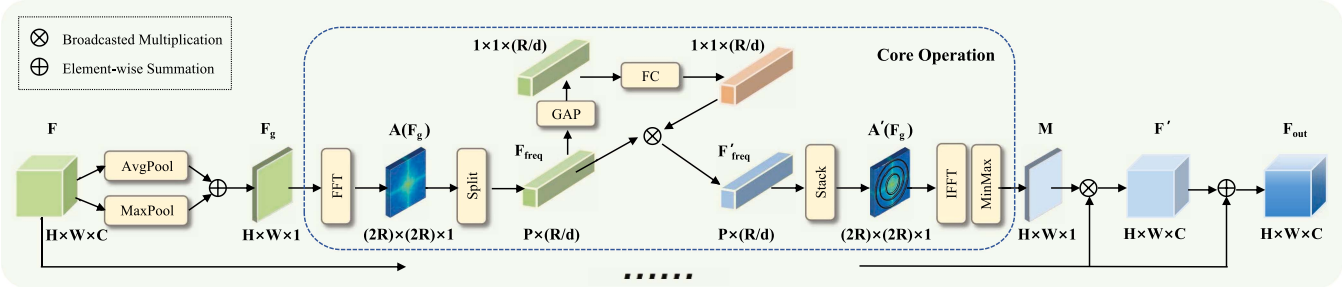

Fig. 3. The detailed architecture of our proposed frequency-guided spatial attention (FGSAttn) Module.

Then, $\mathbf{F}_g$ is transformed into the frequency domain by Fourier transform to obtain its amplitude spectrum $\mathcal{A}(\mathbf{F}_g) \in \mathbb{R}^{(2R)\times(2R)\times 1}$ and the phase spectrum $\mathcal{P}(\mathbf{F}_g) \in \mathbb{R}^{(2R)\times(2R)\times 1}$,

$$\begin{aligned}\mathcal{FT}(\mathbf{F}_g)(u,v) &= \frac{1}{\sqrt{HW}}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\mathbf{F}_g(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)},\\ \mathcal{A}(\mathbf{F}_g)(u,v) &= \sqrt{R^2(\mathbf{F}_g)(u,v)+I^2(\mathbf{F}_g)(u,v)},\\ \mathcal{P}(\mathbf{F}_g)(u,v) &= \arctan\left[\frac{I(\mathbf{F}_g)(u,v)}{R(\mathbf{F}_g)(u,v)}\right],\end{aligned} \tag{2}$$

where $\mathcal{FT}(.)$ represents the fast Fourier transform of the feature. $R(\mathbf{F}_g)$ and $I(\mathbf{F}_g)$ are the real and imaginary part of $\mathcal{FT}(\mathbf{F}_g)$, respectively.

Previous studies [30], [36] have proved that the amplitude component obtained by the Fourier transform contains more critical information for the object. Hence, in our work, we mainly explore the influence of different frequency components in the amplitude spectrum while keeping the phase spectrum unchanged.

After frequency centralization, the origin point in the amplitude spectrum represents center frequency. The distances between a certain point and the origin denotes its frequency component. Therefore, circles with different diameters in the amplitude spectrum correspond to different frequency components. When they are transformed back to the spatial domain, they represent different type of features, such as the approximate shape or edge details of objects. In our method, we decompose the amplitude spectrum into many non-overlapping circular rings with a width of d along the radius dimension, and the hyperparameter d defines the range of frequency components.

We group the features located in the same circular ring into one channel and obtain $\mathbf{F}_{freq} \in \mathbb{R}^{P\times(R/d))}$. P denotes the number of frequency components on each channel. Then the $GAP(.)$ and $FC(.)$ are performed to generate weights for adaptively recalibrating the responses of different frequency components. Thus, mapping the adjusted frequency components back into the spatial domain will change the spatial feature values on the feature map, i.e. frequency-guided spatial adaptation, which can be expressed as:

$$\mathbf{F}'_{freq} = \mathbf{F}_{freq} \otimes FC(GAP(\mathbf{F}_{freq})) \tag{3}$$

where "$\otimes$" denotes element-wise broadcasted multiplication. $GAP(.)$ and $FC(.)$ represent global average pooling and sequences of 1×1 convolutions followed by a LeakyReLU activation function. Finally, the processed features $\mathbf{F}'_{freq}$ are rearranged and stacked sequentially to get a new amplitude spectrum $\mathcal{A}'(\mathbf{F}_g)$. Combined with original phase spectrum $\mathcal{P}(\mathbf{F}_g)$, the spatial attention map $\mathbf{M} \in \mathbb{R}^{H\times W\times 1}$ is then obtained by inverse Fourier transform,

$$\mathbf{M} = MinMax(\mathcal{FT}^{-1}(\mathcal{A}'(\mathbf{F}_g), \mathcal{P}(\mathbf{F}_g))). \tag{4}$$

The final output $\mathbf{F}_{out}$ is

$$\mathbf{F}_{out} = \mathbf{F} + \mathbf{M} \otimes \mathbf{F}, \tag{5}$$

where $\mathcal{FT}^{-1}$ represents the inverse Fourier transform. "$\otimes$" denotes the element-wise broadcasted multiplication along the channel dimension.

C. Frequency-Based Nuances Mining Module

Since camouflaged objects always exhibit very similar appearance features with nearby noisy objects and background, the slight differences are difficult to be distinguished by the spatial domain features of the foundation model alone. We design a Frequency-Based Nuances Mining (FBNM) module aiming at mining nuances between foreground and background, and the detailed architecture is shown in Fig. 2(b).

Specifically, a standard convolution stem borrowed from ResNet is employed to model the local spatial contexts of the input image, which consists of three convolutions and a max-pooling layer. After that, three consecutive sequences are applied to gradually aggregate multi-scale features with three resolutions of 1/8, 1/16, and 1/32, obtaining a feature pyramid of similar resolutions to FPN [37], which is widely used in dense prediction tasks. Each sequence contains a 3×3 convolution kernel to reduce the scale of the feature map, followed by a FGSAttn module which leverages the frequency components to adjust feature layers representing different visual attributes from a global perspective. This can effectively highlight the nuance parts in texture-rich regions to distinguish the foreground and background.

Next, we project the feature maps to the same dimension D using several 1x1 convolution layers. After a flatten layer and a concatenate layer, a feature pyramid $\mathbf{F}^i_{adapter} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$ can be then obtained. On one hand, it serves as the input for the next adapter module. On the other

hand, it is injected into serialized tokens F_{vit}^{0} via cross-attention mechanism to obtain F_{vit}^{i} that absorb task related knowledge, which will be used as the input for the successive layers of the pretrained ViT backbone.

D. Frequency-Based Feature Enhancement Module

ViT can encode the relationships between all input tokens. However, the feature differences among different tokens in spatial domain are very slight in COD task, making it difficult for the model to discriminate candidate targets. Thus it is non-trivial to use the more discriminative features learned from the adapter stream to enhance the ViT stream. We design the frequency-based feature enhancement (FBFE) module to enhance the features of ViT stream, and take full advantages of both the general knowledge and task-related knowledge.

As shown in Fig. 2(c), FGSAttn is first applied to the output of pretrained ViT model $\mathbf{F}_{vit}^{i}$, which aims to enhance the target-relevant regions and at the same time suppress background interference with the guidance of frequency domain information. Then, we take $\mathbf{F}_{adapter}^{i}$ as query to extract the most related information from the adjusted general knowledge $\hat{\mathbf{F}}_{vit}^{i}$, and obtain the updated adapter feature $\mathbf{F}_{adapter}^{i+1}$,

$$\begin{aligned}\hat{\mathbf{F}}_{adapter}^{i} &= \mathbf{F}_{adapter}^{i} + Attention(\mathbf{F}_{adapter}^{i}, \hat{\mathbf{F}}_{vit}^{i}) \\ \mathbf{F}_{adapter}^{i+1} &= \hat{\mathbf{F}}_{adapter}^{i} + FFN(\hat{\mathbf{F}}_{adapter}^{i}),\end{aligned} \tag{6}$$

where $Attention(\cdot,\cdot)$ denotes cross-attention mechanism. $FFN(\cdot)$ denotes the convolutional feed-forward network to remedy the defect of fixed-size position embeddings [38]. After that, the updated adapter feature $\mathbf{F}_{adapter}^{i+1}$ acts as key and value, and $\hat{\mathbf{F}}_{vit}^{i}$ as query to inject task-related knowledge into ViT feature $\mathbf{F}_{vit}^{i+1}$, which will be fed back into the backbone. This process can be expressed as follows:

$$\mathbf{F}_{vit}^{i+1} = \hat{\mathbf{F}}_{vit}^{i} + Attention(\hat{\mathbf{F}}_{vit}^{i}, \mathbf{F}_{adapter}^{i+1}). \tag{7}$$

Note that the last FBFE module only outputs the adapter features, which are used for detection.

E. Loss Function

During training, camouflaged images are fed into both the backbone and adapter simultaneously. We only optimize the parameters of the adapter module and detection head, while keeping the parameters of the original pretrained model frozen, so that the power of the ViT foundation model can be efficiently transferred to downstream COD task with little computational cost. Our entire training process is supervised by the combination of weighted binary cross-entropy loss (L_{BCE}^{w}) [39] and weighted intersection-over-union loss (L_{IOU}^{w}) [39], which can be formulated as $L = L_{BCE}^{w} + L_{IOU}^{w}$, forcing the model to pay more attention to hard pixels.

IV. Experiments

A. Experimental Setup

1) Datasets: The experiments are conducted on four benchmark datasets: CHAMELEON [40], CAMO [41], COD10 K [42], and NC4K [43]. CHAMELEON contains 76 images for test only. While CAMO has 1,000 images for training, and 250 images for testing, consisting of eight categories which fall into both natural and artificial camouflage types. COD10 K [42] is the largest COD dataset till now, consisting of COD10K-Train (3,040 images) and COD10K-Test (2,026 images). NC4K [43] served as the largest testing dataset which includes 4,121 samples and are typically used to evaluate the generalization ability of models. Following experimental protocols in [42], our method is trained on the training sets of CAMO and COD10 K, and the detection performance on the whole CHAMELEON and NC4K datasets, together with the test sets of CAMO and COD10 K are reported.

2) Evaluation metrics: Four commonly used metrics are adopted for evaluation: Structure measure (S_α) [44], Mean enhanced-alignment measure (E_ϕ) [45], weighted F-measure (F_β^w) [46], and mean absolute error (M) [47].

3) Training details: In the training phase, we use Vision Transformer [29] as the foundation model and UperNet [48] as the COD head. The Vision Transformer is pretrained with large-scale multi-modal data as in Uni-Perceiver [49] and kept frozen once pretrained. The parameters of adapter and the COD head are both randomly initialized. We employ an AdamW optimizer with initial learning rate of 6×10^{-5} and a weight decay of 0.05. They are trained 200 epochs with a batch size of 2. For testing, the images are resized to 512 $\times$512 to input into the model, and the outputs are resized back to the original size.

4) Competitors: We compare our method with 26 state-of-the-art COD methods, including: SINet [42], PraNet [1], TINet [17], PFNet [50], UGTR [19], C^2FNet [18], S-MGL [6], R-MGL [6], LSR [43], JCSOD [20], ERRNet [51], BASNet [52], SINetV2 [4], ZoomNet [53], PENet [54], MFFN [55], FSP-Net [56], HitNet [5], DINet [57], DCT-Net [22], UEDG [21], FDNet [7], FBNet [9], FPNet [8], FEDER-MS-4 [58], SAM-Adapter [35]. Among these SOTA methods, it is worth noting that FDNet [7], FBNet [9], FPNet [8], FEDER-MS-4 [58] all introduce frequency clue from various aspects in their methods. And SAM-Adapter [35] proposed to adapt the Segment Anything foundation model from a spatial perspective without any guidance. For a fair comparison, all results are either provided by the published paper or reproduced by an open-source model re-trained on the same training set with recommended settings.

B. Comparison With the State-of-the-Arts Methods

1) Quantitative evaluation: Table I reports the detailed comparision results of our FGSA-Net against other 26 state-of-the-art methods on four benchmark datasets. It can be seen that our proposed method outperforms all the comparison SOTA methods with a large margin on all the benchmark datasets. For example, our method achieves 0.893 S_α and 0.849 F_β^w on

TABLE I
QUANTITATIVE COMPARISON OF OUR FGSA-NET AND 26 SOTA METHODS FOR COD ON FOUR BENCHMARK DATASETS

Methods	Pub/Year	CHAMELEON (76)				CAMO-Test (250)				COD10K-Test (2,026)				NC4K-Test (4,121)			
		$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
SINet	$CVPR_{20}$	.869	.891	.740	.044	.751	.771	.606	.100	.771	.806	.551	.051	.808	.871	.723	.058
PraNet	$MICCAI_{20}$	.860	.898	.763	.044	.769	.824	.663	.094	.789	.861	.629	.045	.822	.876	.724	.059
TINet	$AAAI_{21}$	.874	.916	.783	.038	.781	.847	.678	.087	.793	.848	.635	.043	.829	.879	.734	.055
PFNet	$CVPR_{21}$	.882	.942	.810	.033	.782	.852	.695	.085	.800	.868	.660	.040	.829	.888	.745	.053
UGTR	$ICCV_{21}$	.888	.918	.796	.031	.785	.859	.686	.086	.818	.850	.667	.035	.839	.874	.747	.052
C^2FNet	$IJCAI_{21}$	.888	.935	.828	.032	.796	.854	.719	.080	.813	.890	.686	.036	.838	.897	.762	.049
S-MGL	$CVPR_{21}$	.892	.921	.803	.032	.772	.850	.664	.089	.811	.851	.655	.037	.829	.863	.731	.055
R-MGL	$CVPR_{21}$	.893	.923	.813	.030	.775	.847	.673	.088	.814	.865	.666	.035	.833	.867	.740	.052
LSR	$CVPR_{21}$	.893	.938	.839	.033	.793	.826	.725	.085	.793	.868	.685	.041	.839	.883	.779	.053
JCSOD	$CVPR_{21}$	.894	.943	.848	.030	.803	.853	.759	.076	.817	.892	.726	.035	.842	.898	.771	.047
ERRNet	PR_{22}	.877	.927	.805	.036	.761	.817	.660	.088	.780	.867	.629	.044	.787	.848	.638	.070
BASNet	$AAAI_{22}$	.914	.954	.866	.022	.749	.796	.646	.096	.802	.855	.677	.038	.817	.859	.732	.058
SINetV2	$TPAMI_{22}$	.888	.942	.816	.030	.820	.882	.743	.070	.815	.887	.680	.037	.847	.903	.770	.048
ZoomNet	$CVPR_{22}$	.902	.958	.845	.023	.820	.892	.752	.066	.838	.911	.729	.029	.853	.912	.784	.043
PENet	$IJCAI_{23}$	.902	.960	.851	.024	.828	.890	.771	.063	.831	.908	.723	.031	.855	.912	.795	.042
MFFN	$WACV_{23}$	.905	.963	.852	.021	-	-	-	-	.846	.917	.745	.028	.856	.915	.791	.042
FSPNet	$CVPR_{23}$	-	-	-	-	.856	.899	.799	.050	.851	.895	.735	.026	.879	.915	.816	.035
DINet	TMM_{24}	-	-	-	-	.821	.874	.790	.068	.832	.903	.761	.031	.856	.909	.824	.043
DTC-Net	TMM_{22}	.876	.897	.773	.039	.778	.804	.667	.084	.790	.821	.616	.041	-	-	-	-
HitNet	$AAAI_{23}$	**.922**	.970	**.903**	.018	.844	.902	.801	.057	.868	.932	.798	.024	.870	.921	.825	.039
UEDG	TMM_{23}	.911	.960	.866	.022	.868	.922	.819	.048	.858	.924	.766	.025	.881	.928	.829	.035
FDNet	$CVPR_{22}$	.898	.949	.837	.027	.844	.898	.778	.062	.837	.918	.731	.030	-	-	-	-
FBNet	$MCCA_{23}$	.888	.939	.828	.032	.783	.839	.702	.081	.809	.889	.684	.035	-	-	-	-
FPNet	MM_{23}	.914	.961	.85	.022	.852	.905	.806	.056	.850	.913	.748	.029	-	-	-	-
FEDER-MS-4	$CVPR_{23}$	.907	.964	.874	.025	.822	.886	.809	.067	.851	.917	.752	.028	.863	.917	.827	.042
SAM-Adapter	$ICCVW_{23}$	.896	.919	.824	.033	.847	.873	.765	.070	.883	.918	.801	.025	-	-	-	-
Ours(FGSA-Net)	-	.916	**.975**	**.903**	**.016**	**.889**	**.944**	**.870**	**.036**	**.893**	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**

↑ / ↓ indicates that larger/smaller is better. The best and second best are bolded and underlined for highlighting, respectively. "—": not available.

COD10K dataset, greatly outperforms the second best SAM-adapter method. And on NC4K dataset, our method sets a remarkable record to increase S_α by 2.50%, E_ϕ by 2.48%, F_β^w by 6.51% and lowers the MAE error by 34.3%, compared with the second best UEDG method. It is worth noting that, our method greatly outperforms SAM-adapter method which is also based on a vision foundation model (SAM) and tuned with a spatial adapter. As SAM-Adapter mainly learns task specific knowledge and injects novel knowledge of downstream task into the model through the adapter from the perspective of spatial domain alone. Due to the high similarity between the camouflaged object and the surrounding environment, spatial adaptation with the guidance from spatial domain directly is easily confused and can not effectively extract subtle features. By contrast, introducing task specific knowledge into the model under the guidance of frequency domain can enable the network to pay more attention to concealed targets. Furthermore, our method also outperforms all the existing frequency-based methods, namely FDNet, FBNet, FPNet and FEDER-MS-4, on all four standard metrics. Compared to other frequency-based methods, our advantage lies in fully utilizing the general knowledge of vision foundation model, proving that the designed adapter can effectively transfer vision foundation model into downstream tasks such as COD.

2) Qualitative evaluation: In Fig. 4, we show the qualitative comparison of our method with several representative SOTA methods on some challenging situations. Benefiting from the discriminative frequency information, our FGSA-Net achieves more competitive visual performance mainly in the following aspects: More accurate localization and complete prediction of targets in low contrast scenes (Row 1), stronger interference suppression when there are confusing objects in the surrounding environment (Row 2) and more precise recognition of complex and fine structure, such as slender details of the object (Row 3). Moreover, our method is also effective in detecting other challenging situations such as indefinable boundary, small object, multiple objects and occlusion (Row 4 to Row 7). The impressive prediction results further verified the effectiveness of the frequency-guided spatial adaptation network.

C. Further Analysis

1) Effectiveness of frequency-guided adapter: To show the effectiveness of our FGSA-Net, we implement other four types of adapters, namely Series-Adapter, Parallel-Adapter, LoRA and ViT-Adapter, while keeping the same pretrained foundation model and pretrained weights with our method. The results in Table II show that the proposed FGSA-Net can greatly outperform all the spatial adaptation variants on four benchmark datasets, indicating that through frequency-guided spatial adaptation on COD task, the general knowledge can be better transferred to deal with the COD problem.

2) Core operation in FGSAttn: To show the advantage of our frequency-based attention over other spatial attention, we compare our method with two variants, which are replacing core operation in FGSAttn with a regular convolution module and a deformable convolution module, respectively. The results are shown in Table III. It can be seen that our method performs better on all datasets compared with other variants. We analyze the reason is that the brightness difference between camouflaged

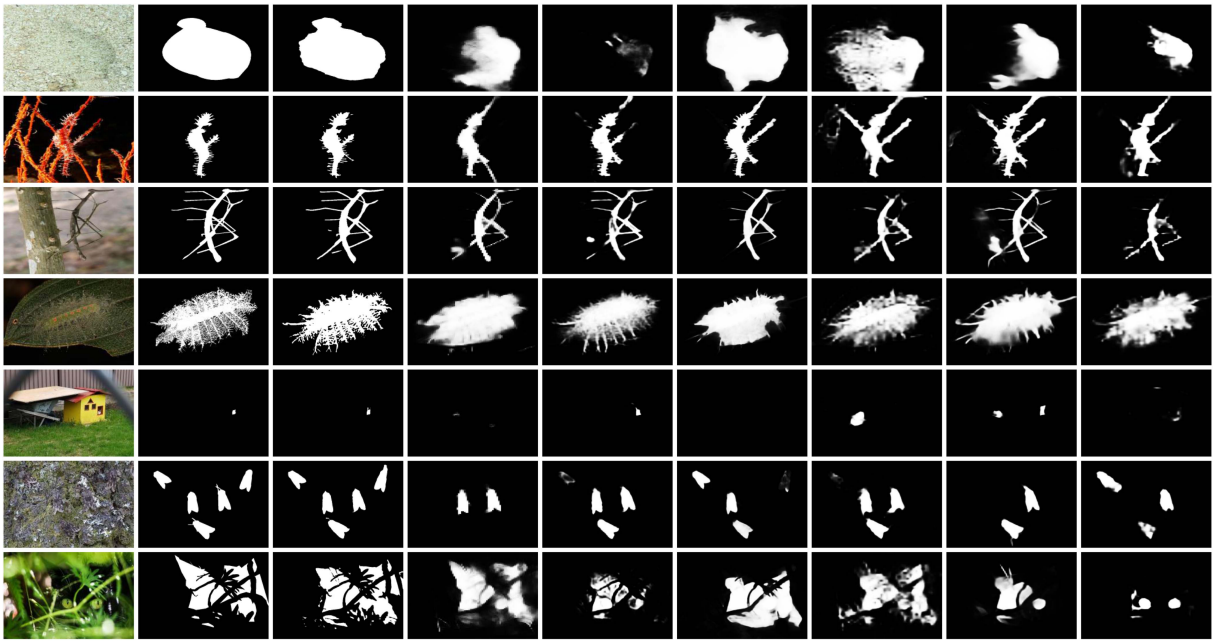

Fig. 4. Visual comparison of our method with six representative state-of-the-art methods. Our method is capable of tackling challenging cases (e.g., low contrast, confusing objects, complex and fine structure, indefinable boundary, small object, multiple objects, and occlusion).

TABLE II
COMPARISON OF DIFFERENT PARAMETER-EFFICIENT TUNING METHODS WITH THE SAME PRETRAINED MODEL ON FOUR BENCHMARK DATASETS

Methods	CHAMELEON (76)				CAMO-Test (250)				COD10K-Test (2,026)				NC4K-Test (4,121)			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Full-Tuning	**.923**	.973	**.913**	.017	**.890**	**.946**	**.877**	**.035**	.891	.942	.845	.016	.898	.940	.874	.024
Backbone-Frozen	.870	.924	.819	.028	.845	.918	.815	.052	.849	.916	.778	.023	.870	.927	.833	.032
Series-Adapter	.895	.952	.866	.020	.873	.936	.853	.041	.877	.943	.824	.018	.891	.941	.863	.026
Parallel-Adapter	.895	.950	.867	.019	.874	.936	.853	.040	.874	.935	.818	.018	.890	.939	.863	.026
LoRA	.897	.949	.875	.021	.879	.938	.859	.040	.881	.944	.832	.017	.894	.945	.871	.025
ViT-Adapter	.909	.959	.891	.018	.888	.942	.868	.037	.883	.943	.836	.016	.896	.945	.874	.024
Ours(Frequency-Adapter)	.916	**.975**	.903	**.016**	.889	.944	.870	.036	**.893**	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**

The best and second best are bolded and underlined for highlighting, respectively.

TABLE III
ABLATION STUDIES OF THE CORE OPERATION IN FGSATTN

Core operation	CHAMELEON (76)				CAMO-Test (250)				COD10K-Test (2,026)				NC4K-Test (4,121)			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Regular convolution	.910	.964	.893	.018	.874	.938	.862	.038	.888	.946	.839	.016	.899	.948	.875	.025
Deformable convolution	.912	.964	.895	.017	.873	.939	.862	.037	.889	.945	.840	.016	.897	.947	.878	.025
Ours(Frequency-Based)	**.916**	**.975**	**.903**	**.016**	**.889**	**.944**	**.870**	**.036**	**.893**	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**

Best results are marked in bold fonts.

object and surrounding environment is very small in the spatial domain, and attention maps generated by using convolution operations based on spatial domain may be confused, making it difficult for the model to effectively focus on the camouflaged object and detailed clues.

3) Effect of FGSAttn on FBNM and FBFE: To show how much our proposed FGSAttn takes effect on FBNM and FBFE module, we evaluate our method while removing FGSAttn in FBNM and FBFE respectively. As can be seen from Table IV that removing FGSAttn in either module would get a significant degradation of our FGSA-Net.

4) Feature map visualization after adaptation: Fig. 1 illustrates some representative cases of the obtained feature maps after adapter tuning on COD task. It is noticeable that series-adapter, parallel-adapter and LoRA can only focus on the boundaries of the target roughly, while ViT-adapter highlights the target

TABLE IV
ABLATION STUDIES OF FGSATTN IN DIFFERENT MODULES

Settings	Components		CHAMELEON (76)				CAMO-Test (250)				COD10K-Test (2,026)				NC4K-Test (4,121)			
	FBNM	FBFE	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Baseline			.909	.959	.891	.018	.878	.935	.868	.037	.883	.943	.836	.016	.896	.945	.874	.024
Variant 1		✓	.910	.969	.892	.019	.884	.942	.866	.037	.887	.949	.839	.016	.898	.946	.873	.024
Variant 2	✓		.910	.969	.894	.018	.887	.943	.866	.037	.890	.951	.845	.016	.899	.947	.875	.025
Ours(FGSA-Net)	✓	✓	**.916**	**.975**	**.903**	**.016**	**.889**	**.944**	**.870**	**.036**	**.893**	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**

Best results are marked in bold.

TABLE V
EFFECTIVENESS OF VARIOUS INPUT SIZES

Input size	CHAMELEON (76)				CAMO-Test (250)				COD10K-Test (2,026)				NC4K-Test (4,121)			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
352×352	.901	.963	.884	.019	.880	.939	.857	.040	.873	.939	.813	.019	.893	.944	.865	.025
384×384	.901	.969	.888	.020	.881	.938	.858	.038	.879	.946	.824	.018	.897	.947	.870	.025
416×416	.908	.971	.894	.018	.882	.940	.861	.038	.883	.947	.831	.017	.898	.948	.873	.024
512×512	**.916**	**.975**	**.903**	**.016**	**.889**	**.944**	**.870**	**.036**	**.893**	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**

Best results are marked in bold.

TABLE VI
EFFECTIVENESS OF DIFFERENT PRETRAINED WEIGHTS

Backbone	CHAMELEON (76)				CAMO-Test (250)				COD10K-Test (2,026)				NC4K-Test (4,121)			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
AugReg	.910	.964	.895	.018	.875	.939	.853	.040	.883	.934	.833	.017	.891	.933	.863	.024
BEiT	.911	.966	.898	.017	.871	.930	.847	.046	.881	.932	.825	.019	.889	.936	.862	.026
SAM	.915	.969	.901	.017	.873	.938	.853	.043	.889	.936	.839	.018	.893	.934	.865	.027
Uni-Perceiver	**.916**	**.975**	**.903**	**.016**	**.889**	**.944**	**.870**	**.036**	**.893**	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**

Best results are marked in bold.

but also generates more background noise, which is detrimental to the model's localization and recognition capabilities. Only our proposed frequency-guided spatial adaptation method clearly tunes the pretrained model to focus more on the concealed foreground objects compared with other four spatial adaptation counterparts.

5) Various input image sizes: To explore the impact of various input image sizes on model performance, we present the results of our model at different input sizes, including 352 × 352, 384 × 384, 416 × 416 and 512 × 512, which are illustrated in Table V. As can be seen, the performance of the model gradually improves with the increase of input image resolution, and our model performs the best at a setting of 512 × 512. It is worth noting that when decreasing input size into 416 × 416, 384 × 384, 352 × 352, our method also achieve SOTA results on three datasets(CAMO, COD10K and NC4K) and competitive performance on CHAMELEON dataset, even though other models use larger input sizes, such as FPNet use 512 × 512, HitNet use 704 × 704, and ZoomNet use multiple inputs (maximum 576 × 576).

6) Different pretrained weights: In order to explore the impact of different pretrained weights, we experiment with ViT as the backbone and initialize it with different pretrained weights, including AugReg [59] which trained on ImageNet-22 K, BEiT [60] which trained on ImageNet-1 K, SAM [11] which trained on 11 million images and 1.1 billion masks, and Uni-Perceiver [49] which is trained with large scale multi-modal data. As summarized in Table VI, we find that using various pretrained weights both achieve competitive performance, which verifies the effectiveness of our designed adapter for different pretrained backbone. Among them, the backbone pretrained with multi-modal data show the best performance. It is worth noting that our method significantly outperforms the SAM-Adapter method by utilizing the SAM initialized backbone, demonstrating the superiority of our proposed FGSA-Net.

7) Different K and M: For the pretrained ViT model with L=24, we study the effect of K and M, and the results are shown in Table VII. It can be seen that the model achieves optimal performance on most datasets and metrics when K=6 and M=4, and dividing into more groups cannot bring significant gains. Therefore, we empirically set K=6 and M=4.

8) Different d in the amplitude decomposition: The specific value for width d represents the range of frequency components contained in each channel. To explore the influence of d in the amplitude decomposition, we present the results of our model at different widths, as shown in Table VIII. It can be observed that the performance is the best when d=1, achieving the highest score on multiple metrics, and the performance gradually

TABLE VII
EFFECTIVENESS OF DIFFERENT K AND M

K	M	CHAMELEON(76)				CAMO-Test(250)				COD10K-Test(2,026)				NC4K-Test(4,121)			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
3	8	.910	.971	.900	.018	.884	.939	.866	.038	.884	.949	.840	.019	.899	.941	.872	.025
4	6	.915	.973	**.903**	.017	**.890**	.942	.868	**.036**	.891	.950	.846	.016	.902	**.952**	.882	**.023**
6	4	**.916**	**.975**	**.903**	**.016**	.889	**.944**	**.870**	**.036**	**.893**	**.953**	**.849**	**.015**	**.903**	.951	**.883**	**.023**
8	3	.906	.968	.898	.019	.881	.936	.865	.038	.885	.947	.842	.018	.899	.942	.875	.026

Best results are marked in bold.

TABLE VIII
EFFECTIVENESS OF DIFFERENT WIDTH IN THE AMPLITUDE DECOMPOSITION

d	CHAMELEON(76)				CAMO-Test(250)				COD10K-Test(2,026)				NC4K-Test(4,121)			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
1	**.916**	**.975**	**.903**	**.016**	**.889**	**.944**	**.870**	.036	.893	**.953**	**.849**	**.015**	**.903**	**.951**	**.883**	**.023**
2	.915	.971	.901	.017	.887	.942	.869	**.035**	**.895**	.952	.847	.016	.902	.949	.882	**.023**
4	.912	.970	.899	.018	.888	.943	.868	.036	.892	.952	.847	.016	.902	.950	.881	.024
8	.910	.969	.899	.018	.882	.942	.865	.037	.890	.949	.846	.017	.901	.948	.879	.024

Best results are marked in bold.

TABLE IX
QUANTITATIVE COMPARISON OF OUR FGSA-NET AND 15 SOTA METHODS FOR SOD ON FOUR BENCHMARK DATASETS

Methods	ECSSD				DUTS-TE				HKU-IS				DUT-OMRON			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
BMPM	.911	.914	.871	.045	.862	.860	.761	.049	.907	.937	.859	.039	.809	.837	.681	.064
RAS	.893	.914	.857	.056	.839	.861	.740	.059	.887	.929	.843	.045	.814	.846	.695	.062
PiCA-R	.917	.913	.867	.046	.869	.862	.754	.043	.904	.936	.840	.043	.832	.841	.695	.065
DGRL	.906	.917	.903	.043	.846	.863	.764	.051	.896	.941	.881	.037	.810	.843	.709	.063
CPD-R	.918	.925	.898	.037	.869	.886	.795	.043	.905	.944	.875	.034	.825	.866	.719	.056
PoolNet	.926	.925	.904	.035	.886	.896	.817	.037	.919	.953	.888	.030	.831	.868	.725	.054
SIBA	.924	.928	.908	.035	.879	.892	.811	.040	.913	.950	.886	.032	.832	.860	.736	.059
EGNet	.925	.927	.903	.037	.887	.891	.815	.039	.918	.950	.887	.031	.841	.867	.738	.053
F3Net	.924	.925	.912	.034	.888	.902	.835	.035	.917	.953	.900	.028	.838	.870	.747	.053
ICON	.931	.924	.920	.031	.892	.900	.839	.037	.925	.956	.908	.027	.845	.866	.762	.058
TSNet	.936	.917	.915	.036	.883	.854	.804	.038	.921	.912	.902	.031	.858	.809	.761	.044
PRNet	.917	.946	.895	.039	.879	.910	.811	.039	.913	.956	.885	.032	.829	.866	.723	.056
TINet	.926	.953	.914	.033	.891	.925	.842	.035	.922	.960	.906	.027	.842	.876	.754	.051
JCSOD	.933	.960	.935	.030	.899	.937	.866	.032	.931	.867	.924	.026	.850	.884	.782	.051
BIPGNet	.938	.962	.938	.025	.905	.940	.877	.029	.924	.964	.922	.023	.854	.886	.787	.047
Ours(FGSA-Net)	**.949**	**.981**	**.952**	**.019**	**.926**	**.956**	**.899**	**.024**	**.939**	**.981**	**.939**	**.016**	**.861**	**.892**	**.796**	**.045**

↑ / ↓ indicates that larger/smaller is better. The best and second best are bolded and underlined for highlighting, respectively.

decreases with the increase of d. We speculate that finer decomposition of amplitude is beneficial for the model to better adaptively adjust various attributes of objects, such as the approximate shape or edge details, etc.

9) Generalization performance on salient object detection (SOD) Task: To validate the generalization of our method on SOD task, we train our FGSA-Net on the DUTS-TR [61] dataset, and directly evaluate on other four testing datasets, including ECSSD [62], DUTS-TE [61], HKU-IS [63] and DUT-OMRON [64]. We compare our method with 15 representative methods, including BMPM [65], RAS [66], PiCA-R [67], DGRL [68], CPD-R [69], PoolNet [70], SIBA [71], EGNet [72], F3Net [39], ICON [73], TSNet [74], PRNeT [75], TINet [17] and JCSOD [20], BIPGNet [76]. Table IX reports the quantitative results on four SOD benchmark datasets. It can be seen that our model performs favorably against the existing methods in terms of nearly all evaluation metrics. For example, compared with the second-best model BIPGNet on ECSSD dataset, our model increases S_α, E_ϕ, and F_β^w by 1.17%, 1.98%, and 1.50% respectively, and lowers the MAE error by 24%. This demonstrates the strong capability and effectiveness of our network to deal with other binary segmentation task.

10) Parameter comparison: In Table X, we compare the number of tunable parameters of our method and some representative SOTA methods, including SINetV2, ZoomNet, PFNet, SINet, LSR, S-MGL, R-MGL, ERRNet, BASNet, JSCOD and SAM-Adapter. For fair comparison, all parameter results are either provided in the published paper or calculated based on

TABLE X
PARAMETER COMPARISON OF OUR FGSA-NET WITH 11 REPRESENTATIVE STATE-OF-THE-ART METHODS

	Ours	SINetV2	ZoomNet	PFNet	SINet	LSR	S-MGL	R-MGL	ERRNet	BASNet	JSCOD	SAM-Adapter
#Param(M)	**59.90**	26.98	32.38	46.50	48.95	57.90	63.60	67.64	69.76	87.06	121.63	206.3

the implementation details in the paper and open-source model code. These statistics highlight that our proposed FGSA-Net is lightweight, requiring less or comparable parameters to achieve promising performance.

V. CONCLUSION

In this paper, we propose a frequency-guided spatial adaptation network for COD. Specifically, a frequency-guided spatial attention module is devised to adapt the pretrained foundation model from spatial domain to focus more on the camouflaged regions, while guided by the frequency components dynamically adjusted in the frequency domain. Based on the attention module, the FBNM and FBFE module are further proposed to extract and fuse multi-scale features which contain both the general knowledge of the pretrained model and specialized knowledge learned from the downstream COD dataset. Extensive experiments verify that our proposed method outperforms the baseline counterparts with large margins and achieves state-of-the-art performances on four benchmark datasets.

REFERENCES

[1] D.-P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, Springer, 2020, pp. 263–273.

[2] Y. Huang et al., "MCMT-GAN: Multi-task coherent modality transferable GAN for 3D brain image synthesis," *IEEE Trans. Image Process.*, vol. 29, pp. 8187–8198, 2020.

[3] M. Dean, R. Harwood, and C. Kasari, "The art of camouflage: Gender differences in the social behaviors of girls and boys with autism spectrum disorder," *Autism*, vol. 21, pp. 678–689, 2017.

[4] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, Oct. 2022.

[5] X. Hu et al., "High-resolution iterative feedback network for camouflaged object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 881–889.

[6] Q. Zhai et al., "Mutual graph learning for camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12992–13002.

[7] Y. Zhong et al., "Detecting camouflaged object in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4494–4503.

[8] R. Cong et al., "Frequency perception network for camouflaged object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 1179–1189.

[9] J. Lin, X. Tan, K. Xu, L. Ma, and R. W. H. Lau, "Frequency-aware camouflaged object detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, pp. 1–16, 2023.

[10] W. Wang et al., "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14408–14419.

[11] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.

[12] Z. Hu et al., "LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models," in *Proc. 2023 Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 5254–5276.

[13] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–20.

[14] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–20.

[15] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 1335–1341.

[16] Q. Jia et al., "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4703–4712.

[17] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflaged objects by texture-aware interactive guidance network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3599–3607.

[18] Y. Sun et al., "Context-aware cross-level fusion network for camouflaged object detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1025–1031.

[19] F. Yang et al., "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4126–4135.

[20] A. Li et al., "Uncertainty-aware joint salient object and camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10066–10076.

[21] Y. Lyu et al., "UEDG: Uncertainty-edge dual guided camouflage object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 4050–4060, 2024.

[22] W. Zhai, Y. Cao, H. Xie, and Z.-J. Zha, "Deep texton-coherence network for camouflaged object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 5155–5165, 2023.

[23] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 709–727.

[24] R. Zhang et al., "Tip-Adapter: Training-free CLIP-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.

[25] J. He et al., "Towards a unified view of parameter-efficient transfer learning," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–15.

[26] F. Yuan, X. He, A. Karatzoglou, and L. Zhang, "Parameter-efficient transfer from sequential behaviors for user modeling and recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1469–1478.

[27] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[29] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–21.

[30] M. Zhou, J. Huang, C.-L. Guo, and C. Li, "Fourmer: An efficient global modeling paradigm for image restoration," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 42589–42601.

[31] C. Li et al., "Embedding Fourier for ultra-high-definition low-light image enhancement," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–27.

[32] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 763–772.

[33] K. Xu et al., "Learning in the frequency domain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1737–1746.

[34] Y. Chen et al., "Drop an Octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3434–3443.

[35] T. Chen et al., "SAM-adapter: Adapting segment anything in underperformed scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2023, pp. 3359–3367.

[36] Y. Zhang, Y. Lu, Y. Yan, H. Wang, and X. Li, "Frequency domain nuances mining for visible-infrared person re-identification," 2024, *arXiv:2401.02162*.

[37] T.-Y. Lin et al., "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[38] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, pp. 415–424, 2022.

[39] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12321–12328.

[40] P. Skurowski et al., "Animal camouflage analysis: Chameleon database," Unpublished manuscript, vol. 2, no. 6 , p. 7, 2018.
[41] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Understanding*, vol. 184, pp. 45–56, 2019.
[42] D.-P. Fan et al., "Camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2774–2784.
[43] Y. Lv et al., "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11586–11596.
[44] D.-P. Fan et al., "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 4558–4567.
[45] D.-P. Fan et al., "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.
[46] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.
[47] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.
[48] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 432–448.
[49] X. Zhu et al., "Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16783–16794.
[50] H. Mei et al., "Camouflaged object segmentation with distraction mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8768–8777.
[51] G.-P. Ji, L. Zhu, M. Zhuge, and K. Fu, "Fast camouflaged object detection via edge-based reversible re-calibration network," *Pattern Recognit.*, vol. 123, 2022, Art. no. 108414.
[52] X. Qin et al., "BASNet: Boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7471–7481.
[53] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2150–2160.
[54] X. Li et al., "Locate, refine and restore: A progressive enhancement network for camouflaged object detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2023, pp. 1116–1124.
[55] D. Zheng et al., "MFFN: Multi-view feature fusion network for camouflaged object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6221–6231.
[56] Z. Huang et al., "Feature shrinkage pyramid for camouflaged object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5557–5566.
[57] X. Zhou, Z. Wu, and R. Cong, "Decoupling and integration network for camouflaged object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 7114–7129, 2024.
[58] C. He et al., "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22046–22055.
[59] A. Steiner et al., "How to train your ViT? Data, augmentation, and regularization in vision transformers," *Trans. Mach . Learn. Res.*, vol. 2022, 2022.
[60] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2021.
[61] L. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3796–3805.
[62] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1155–1162.
[63] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5455–5463.
[64] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.
[65] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1741–1750.
[66] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 236–252.
[67] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3089–3098.
[68] T. Wang et al., "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3127–3135.
[69] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3902–3911.
[70] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3912–3921.
[71] J. Su et al., "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3798–3807.
[72] J.-X. Zhao et al., "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8778–8787.
[73] M. Zhuge et al., "Salient object detection via integrity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3738–3752, Mar. 2023.
[74] Z. Wu, S. Li, C. Chen, A. Hao, and H. Qin, "Deeper look at image salient object detection: Bi-stream network with a small training dataset," *IEEE Trans. Multimedia*, vol. 24, pp. 73–86, 2022.
[75] J. Zhu et al., "Perception-and-regulation network for salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 6525–6537, 2023.
[76] Z. Yao and L. Wang, "Boundary information progressive guidance network for salient object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 4236–4249, 2022.

Shizhou Zhang received the B.E. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2010 and 2017, respectively. He is currently a Tenured Associate Professor with Northwestern Polytechnical University, Xi'an, China. His research interests include content-based image analysis, pattern recognition and machine learning, specifically in deep learning-based vision tasks such as image classification, object detection, and re-identification.

Dexuan Kong received the B.S. degree from the School of Computer Science, Hebei Normal University, Shijiazhuang, China, in 2022. She is currently working toward the M.S. degree with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. Her research interests include camouflaged object detection and remote sensing image processing.

Yinghui Xing (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Artificial Intelligence, Xidian University, Xi'an, China, in 2014 and 2020, respectively. She is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an. Her research interests include remote sensing, image processing, image fusion, and image super resolution.

Yue Lu received the B.E. degree in automation from the Beijing Information Science and Technology University, Beijing, China, in 2020, and the M.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2023. He is currently working toward the Ph.D. degree in computer science with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University. His research interests include computer vision and continual learning.

Lingyan Ran received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2018, respectively. From 2013 to 2015, he was a Visiting Scholar with the Stevens Institute of Technology, Hoboken, NJ, USA. He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University. His research interests include image classification and semantic segmentation. Dr. Ran is also a Member of CSIG.

Guoqiang Liang received the B.S. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from Xi′an Jiaotong University, Xi′an, China, in 2012 and 2018, respectively. From March 2017 to September. 2017, he was a visiting Ph.D. student with the University of South Carolina, Columbia, SC, USA. He is currently a Postdoctoral Research with the School of Computer Science and Engineering, Northwestern Polytechnical University, Xi′an. His research interests include human pose estimation and human action classification.

Hexu Wang is currently an Associate Professor with Xijing University, Xi'an, China. She is also with the School of Information and Technology, Xi'an Key Laboratory of Human-Machine Integration, Control Technology for Intelligent Rehabilitation, Northwest University, Xi'an. Her research interests mainly include AI supply chain and machine learning.

Yanning Zhang (Senior Member, IEEE) received the B.S. degree from the Dalian University of Science and Engineering, Dalian, China, in 1988, the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1993 and 1996, respectively. She is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. She is also the organization Chair of the Ninth Asian Conference on Computer Vision. Her research interests include signal and image processing, computer vision, and pattern recognition. She has authored or coauthored more than 200 papers in international journals, conferences, and Chinese key journals.