Dichotomous Image Segmentation with Frequency Priors

Yan Zhou[1,4], **Bo Dong**[2], **Yuanfeng Wu**[3], **Wentao Zhu**[3], **Geng Chen**[4*] and **Yanning Zhang**[4]
[1]State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, China
[2]College of Biomedical Engineering and Instrumental Science, Zhejiang University, China
[3]Zhejiang Lab, China
[4]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, China
geng.chen.cs@gmail.com

Abstract

Dichotomous image segmentation (DIS) has a wide range of real-world applications and gained increasing research attention in recent years. In this paper, we propose to tackle DIS with informative frequency priors. Our model, called FP-DIS, stems from the fact that prior knowledge in the frequency domain can provide valuable cues to identify fine-grained object boundaries. Specifically, we propose a frequency prior generator to jointly utilize a fixed filter and learnable filters to extract informative frequency priors. Before embedding the frequency priors into the network, we first harmonize the multi-scale side-out features to reduce their heterogeneity. This is achieved by our feature harmonization module, which is based on a gating mechanism to harmonize the grouped features. Finally, we propose a frequency prior embedding module to embed the frequency priors into multi-scale features through an adaptive modulation strategy. Extensive experiments on the benchmark dataset, DIS5K, demonstrate that our FP-DIS outperforms state-of-the-art methods by a large margin in terms of key evaluation metrics.

1 Introduction

Dichotomous image segmentation (DIS) [Qin *et al.*, 2022] aims to segment fine-grained objects from various natural scenes. As an emerging image segmentation task, DIS has great potential in a large number of applications, such as image editing [Kawar *et al.*, 2023], 3D reconstruction [Geiger *et al.*, 2011], remote sensing [Zhang *et al.*, 2022a], medical image analysis [Dong *et al.*, 2023b; Ji *et al.*, 2022], virtual reality [Singh *et al.*, 2020], and so on. Different from existing segmentation tasks, DIS focuses on challenging fine-grained object segmentation. Therefore, conventional segmentation models show unsatisfactory performance in DIS, raising significant demands for designing models dedicated to DIS.

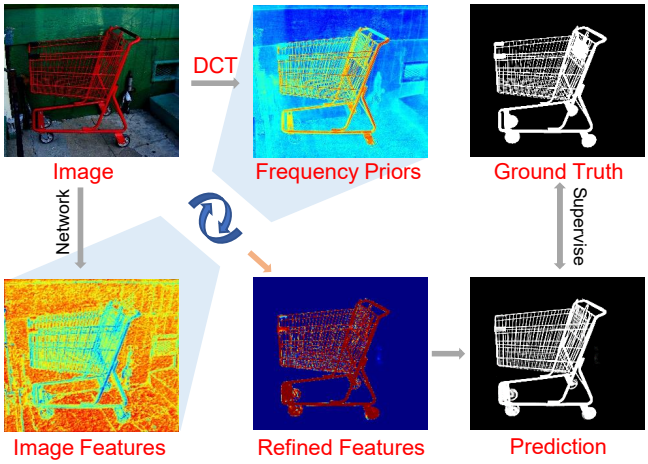

Figure 1: An overview of frequency prior embedding. Our FP-DIS embeds frequency priors into image features to obtain refined features for improving the accuracy of DIS. "DCT" is short for discrete cosine transformation.

Image segmentation is a long-standing topic in computer vision and has made significant progress with the development of deep learning techniques [Qin *et al.*, 2020]. Deep learning has also been employed for DIS. A typical method is IS-Net [Qin *et al.*, 2020], which employs an intermediate supervision strategy to improve network training for DIS. Despite the progress of existing works, many factors restrict the accuracy of DIS. Existing works usually rely on features extracted from images without the consideration of valuable frequency priors. In practice, a number of works have demonstrated that prior knowledge from the frequency domain can effectively improve the performance of various computer vision tasks. For instance, [Qian *et al.*, 2020] proposed a novel face forgery detection network, called F3Net, which utilizes frequency information to perceive forgery cues. [Zhong *et al.*, 2022] designed a frequency enhancement module to introduce the frequency domain as an additional clue to make up for the lack of a single RGB domain to better detect camouflaged objects. These studies demonstrate that knowledge in the frequency domain can assist in the identification of fine-grained object boundaries, which play a vital role in DIS.

*Corresponding author.

However, incorporating frequency priors into DIS faces a number of challenges, including (i) how to extract informative frequency priors from images, (ii) how to embed the frequency priors into multi-scale side-out features of the backbone, and (iii) how to address the heterogeneity of multi-scale features before frequency priors embedding.

To this end, we propose to improve DIS with frequency priors and design a novel deep learning model, called FP-DIS, to address the aforementioned challenges. Specifically, we propose a frequency prior generator to extract informative frequency priors with both fixed and learnable filters. The resulting frequency priors are then embedded into harmonized multi-scale features for accurate DIS, as shown in Figure 1. Here, we propose a feature harmonization module (FHM) to harmonize the multi-scale side-out features from a backbone based on our pyramid feature extractor. Our FHM harmonizes the grouped features to reduce the heterogeneity of multi-scale features. To embed these frequency priors efficiently, we propose a frequency prior embedding module (FPEM), where two feature embedding components are integrated in a cascaded manner. Our FP-DIS makes full use of frequency priors to improve DIS and addresses the challenges associated with frequency priors extraction and embedding with specially-designed modules. Extensive experiments on the benchmark dataset, DIS5K [Qin *et al.*, 2022], show that our FP-DIS achieves promising performance in comparison with cutting-edge methods with superior fine-grained segmentation capabilities. To put this in perspective, our FP-DIS outperforms the state-of-the-art model, IS-Net, by a large margin in terms of maximal F-measure on different sub-datasets, i.e., 4.4% on DIS-TE1, 2.8% on DIS-TE2, 3.8% on DIS-TE3, 1.9% on DIS-TE4, and 3.2% on DIS-VD. Our code is available at https://github.com/dongbo811/FP-DIS.

2 Related Work

2.1 Image Segmentation

Recent works have made great progress on image segmentation tasks [Pang *et al.*, 2022]. However, compared with the general segmentation models and the specific scene models, the core difficulty of the DIS task lies in the mining of object details. For general segmentation task models, the architecture design tends to be robust to each class in various scenes. Due to the relatively rough labeling of the dataset, it cannot be directly applied to the DIS task. There are also models designed for specific tasks, such as visually salient object detection [Dong *et al.*, 2021], camouflaged object detection [Fan *et al.*, 2021a], marine animal segmentation [Li *et al.*, 2021b], and medical image segmentation [Fan *et al.*, 2020; Liu *et al.*, 2021; Lin *et al.*, 2022]. However, these methods are unsuitable for the DIS task. Specifically, since objects in salient detection tasks always have prominent contours and distinct colors, current methods mainly use boundary enhancement [Qin *et al.*, 2019], semantic enhancement [Wu *et al.*, 2019] or design attention mechanism [Woo *et al.*, 2018] to improve segmentation performance. Camouflaged object detection aims to segment camouflaged objects from complex natural scenes. The difficulty is that the object to be segmented is indistinguishable from the background with very similar features such as color, outline, texture, etc. In recent studies, some construct multi-task learning frameworks [Zhai *et al.*, 2021], some propose uncertainty reasoning methods [Li *et al.*, 2021a] or multi-scale feature fusion [Chen *et al.*, 2022] for enhancing camouflaged object detection performance. According to IS-Net [Qin *et al.*, 2022], DIS is formulated as a *category-agnostic* task defined on *non-conflicting annotations* for accurately segmenting objects with *different structure complexities*, regardless of their characteristics. Following this principle, the DIS5K dataset is constructed, which uses three commonly used metrics, namely isoperimetric inequality quotient, number of object contours, and number of dominant points, to measure the fine-grained level of objects in various scenes. However, due to the elusive target objects with complex detailed structures in DIS, there is a bottleneck for feature representation, which limits the segmentation performance. Unlike current techniques that only focus on semantic enhancement, we add frequency priors to help the model capture more detailed information.

2.2 Frequency Priors in Computer Vision

Frequency domain signals have been widely used in computer vision tasks, such as image classification [Stuchi *et al.*, 2017], super-resolution [Huang *et al.*, 2017], and fake face detection [Li *et al.*, 2018]. Some current studies use high-pass filters [Pandey *et al.*, 2016] to extract useful detailed features, and some models use discrete Fourier transforms [Durall *et al.*, 2019] to convert images into the frequency domain to explore bottom-layer information. To improve the performance, many models propose to enhance the feature representation [Pang *et al.*, 2022]. Some methods artificially synthesize features to further enrich the contained information [De Carvalho *et al.*, 2013]. Some recent studies propose extracting important frequency components and filtering frequencies [Dong *et al.*, 2023a], which are beneficial to enhance semantic information. [Zhang *et al.*, 2022b] proposed to process different video bands separately to promote similarity within objects. In addition, some methods [Liu *et al.*, 2023] employ knowledge distillation to reduce the domain gap between the frequency domain and the image domain. Different from these methods, our model obtains the frequency priors of the image by DCT transform and inverse transform, combined with frequency filters. The frequency priors contain detailed information that is difficult to detect in the color space of RGB images. Furthermore, there exists a large gap between image and frequency domain features. Therefore, we design a frequency prior embedding module to eliminate the semantic gap between frequency priors and image features by modulating their distribution to obtain a finer segmentation effect.

3 Method

In this section, we explain details about the proposed FP-DIS. The overall architecture is shown in Figure 2.

3.1 Overall Architecture

The core sight of FP-DIS is that, with the guidance of the frequency priors, the multi-scale image features can capture

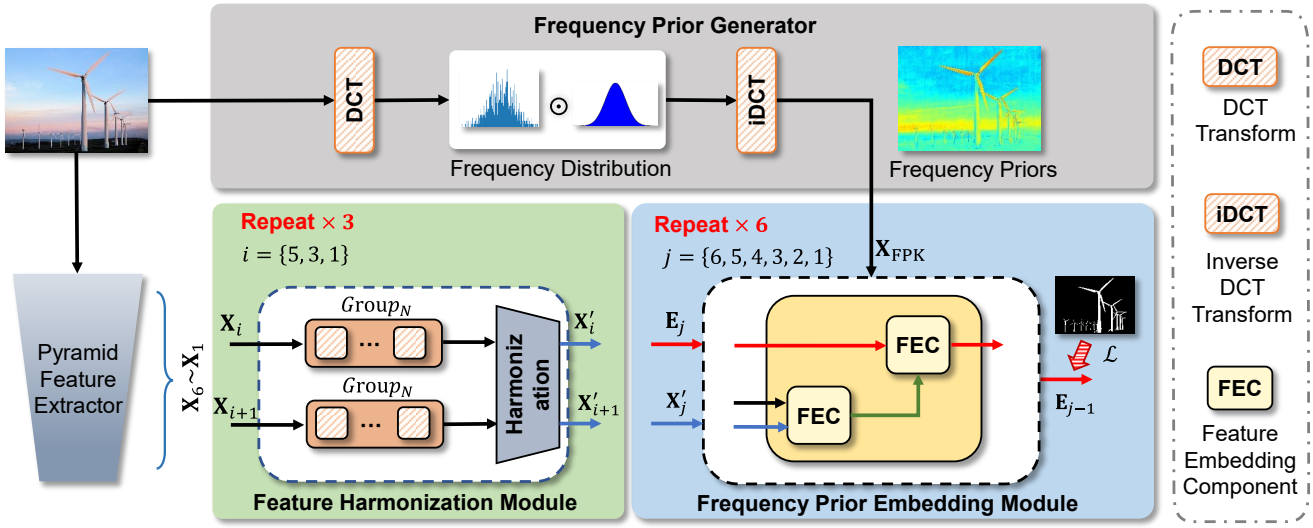

Figure 2: Overall architecture of FP-DIS. DCT and iDCT are discrete cosine transform and inverse discrete cosine transform, respectively. Two adjacent layers of features ($\mathbf{X}_i$, $\mathbf{X}_{i+1}$) go through the feature harmonization module to obtain harmonized feature ($\mathbf{X}'_i$, $\mathbf{X}'_{i+1}$). The frequency priors are embedded into the harmonized feature $\mathbf{X}'_j$ and the output $\mathbf{E}_j$ of upper layer FPEM. Multi-scale frequency embedding feature $\mathbf{E}_j$ propagates from top to down layer. Finally, the final prediction $\mathbf{E}_0$ is upsampled to the original image size.

more details of the input image. Motivated by this, FP-DIS mainly consists of four parts: a pyramid feature extractor, a frequency prior generator, a feature harmonization module, and a frequency prior embedding module. As shown in Figure 2, we first capture multi-scale features of the input image with a pyramid feature extractor consisting of a CNN-based backbone and a transformer-based component. The feature harmonization module is adopted to harmonize features at adjacent semantics in different scales. Meanwhile, the input image is fed into the frequency prior generator to calculate frequency priors. Finally, the frequency priors are embedded into the harmonized multi-scale features with the frequency prior embedding module. The details of each module are described in the following subsections.

3.2 Pyramid Feature Extractor

Convolution neural networks are widely adopted for common vision tasks and get a satisfactory performance. However, in the DIS task, the size of the input image is always large and the target objects own abundant details. It is difficult for shallow networks to learn rich semantics and refine the features fed with large input since they concentrate more on local information. To obtain more semantic information on multi-scale, we use the vision transformer with long-distance modeling capability to deepen the network in addition to the convolution layers in the pyramid feature extractor.

Specifically, the pyramid feature extractor first uses ResNet-50 as the CNN-based backbone to extract multi-scale features $\{\mathbf{X}_i\}_{i=1}^4 \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i}$ from the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the height and width, and $C_i \in \{256, 512, 1024, 2048\}$ is the number of channels. Then the feature $\mathbf{X}_4$ is downsampled by a 3×3 convolution layer with a stride of 2, fed into a transformer block [Ren *et al.*, 2022] to get $\mathbf{X}_5$. Finally, we use the same operations on $\mathbf{X}_5$ to obtain $\mathbf{X}_6$. Notably, the channel of $\mathbf{X}_5$ and $\mathbf{X}_6$ are 256. To facilitate subsequent processing, we convert the channel of all these features $\{\mathbf{X}_i\}_{i=1}^6$ to 96.

3.3 Frequency Prior Generator

Inspired by [Qian *et al.*, 2020], to generate frequency priors, the discrete cosine transform (DCT) is first used to transform the image $\mathbf{I}$ into the frequency domain to generate the frequency distribution map $\mathbf{M}$, i.e.,

$$\mathbf{M} = \text{DCT}(\mathbf{I}). \tag{1}$$

Next, the fixed filter and the learnable filters extract different and valid frequency components. Especially, the fixed filter divides the frequency components into different bands (low frequency, medium frequency, high frequency, and all frequency), while the learnable filters provide more abundant information, and $\sigma = \frac{1-exp(-x)}{1+exp(-x)}$ used to normalize x to the range between -1 and $+1$. Finally, the frequency priors $\mathbf{X}_{\text{FP}}$ are generated using the inverse discrete cosine transform (iDCT):

$$\mathbf{X}_{\text{FP}} = \text{iDCT}(\mathbf{M} \otimes (\mathbf{F}_{\text{f}} + \sigma(\mathbf{F}_{\text{l}}))), \tag{2}$$

where $\text{iDCT}(\cdot)$ denotes inverse discrete cosine transform, $\otimes$ is Hadamard product, $\mathbf{F}_{\text{f}}$ and $\mathbf{F}_{\text{l}}$ are the fixed filter and the learnable filters, respectively.

3.4 Feature Harmonization Module

The multi-scale features generated by the pyramid structure contain different structural and semantic information, inducing considerable heterogeneity. The shallow layers capture abundant details, while the deeper layers extract features with more semantics. The fusion of these multi-scale features would drive the model to focus on both detailed and abstract information. Therefore, we design a feature harmonization

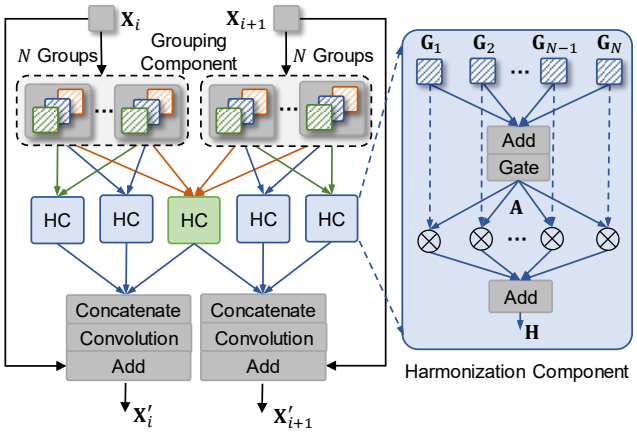

Figure 3: The structure of feature harmonization module. We use feature grouping to cluster similar semantics and different semantics between different scales. The obtained grouping features were harmonized separately in inter-group and intra-group.

module to harmonize features at different scales. As shown in Figure 3, the FHM consists of two major parts, namely the grouping component and the harmonization component, each of which is detailed as follows.

Grouping Component. In this component, given features $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ from the pyramid feature extractor, we expand the channel by N times and then split them into N groups. In this way, the intra-layer groups from the same input contain similar features, while the inter-layer groups from different inputs contain features with a large variation. Finally, we obtain $2N$ groups of features. Each grouped feature is further split into three sub-groups along the channel dimension.

Harmonization Component. After the grouping of the given inputs $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$, the semantics within each group of layers are tight, and the semantic difference between each group of different scales is obvious. Therefore, we need to achieve feature harmonization, and the core of the harmonization mechanism is to use a gate unit for filtering. There are two harmonization mechanisms namely the inter-group harmonization and intra-group harmonization. For inter-group harmonization, as shown in Figure 3, it (green block) collects a splitter of all groups. So it has $2N$ inputs with different scales. For intra-group harmonization, it collects splitters with the same scale as shown in Figure 3 (blue block). So, for each scale, it has two harmonization components.

Using inter-group harmonization as an example, we denote the grouped features as $\{\mathbf{G}_n\}_{n=1}^{N}$, set $N = 4$. As shown in the far right of Figure 3, we first obtain the aggregated feature by adding all input features. Note that, for inter-group harmonization, we interpolate the splitters from $\mathbf{X}_i$ to the same scale as $\mathbf{X}_{i+1}$ since these inputs are with different scales. These splitters are then fed into a gate unit, which computes the modulation weight matrix $\mathbf{A}$. Mathematically, the gate unit is defined as follows:

$$\mathbf{A} = \text{Softmax}(\text{MLP}(\text{ReLU}(\text{MLP}(\text{Avg}(\sum_{n=1}^{N} \mathbf{G}_n))))), \quad (3)$$

where $\text{Avg}(\cdot)$ is an adaptive average pooling operation acting on the spatial dimensions of features, $\text{MLP}(\cdot)$ stands for a multilayer perception, $\text{ReLU}(\cdot)$ denotes a ReLU activation function, and $\text{Softmax}(\cdot)$ represents the softmax function.

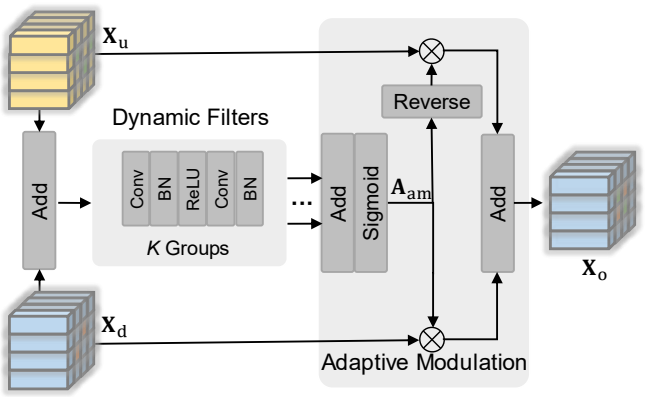

Figure 4: The structure of feature embedding component. We add $\mathbf{X}_u$ and $\mathbf{X}_d$, feed the output to K groups of dynamic filters. Then, we adopt adaptive modulation to get final output feature $\mathbf{X}_o$.

The modulation weights are multiplied with the original features to obtain a harmonized feature map, i.e.,

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{G}_n \otimes \mathbf{A}. \quad (4)$$

Finally, the output of inter-group harmonization is interpolated into the same size of each branch. We concatenate the harmonized features from the same branch, pass them through a convolution layer for smoothness, and add them with the input features to obtain the output features $(\mathbf{X}'_i, \mathbf{X}'_{i+1})$ of the feature harmonization module.

3.5 Frequency Prior Embedding Module

Incorporating frequency priors into image features makes the output contain more detailed information. However, direct fusion leads to inferior performance due to the semantic gap between the frequency and image domains. To address this problem, in the frequency prior embedding module (FPEM), we design the feature embedding component which consists of K groups of dynamic filters followed by an adaptive modulation. To make full use of the frequency priors, we adopt the cascaded feature propagation mechanism.

Feature Embedding Component. For clarity, we denote the inputs of the feature embedding component as paired features $(\mathbf{X}_u, \mathbf{X}_d)$ at upper and lower branches, and the output as feature $\mathbf{X}_o$. As shown in Figure 4, the features at two branches are added up to obtain the aggregated feature first. Then, the aggregated feature is passed through K branches of dynamic filters to learn different information from different representation domains. Each branch consists of a convolution operation to reduce the channel dimension to a quarter of the original for information compression, a BatchNorm layer and a ReLU layer to select the important part and another convolution layer and BatchNorm layer to restore the channel dimension. The output $\mathbf{X}_{df}(k)$ of each branch is calculated with:

$$\mathbf{X}_{df}(k) = \text{DynFilter}_k(\mathbf{X}_u + \mathbf{X}_d), \quad (5)$$

where $\text{DynFilter}_k(\cdot)$ denotes the kth dynamic filter. $\{\mathbf{X}_{\text{df}}(k)\}_{k=1}^{K}$ is then fed into the adaptive modulation unit to compute the attention coefficient matrix $\mathbf{A}_{\text{am}}$:

$$\mathbf{A}_{\text{am}} = \omega \text{Sigmoid}(\sum_{k=1}^{K} \mathbf{X}_{\text{df}}(k)), \tag{6}$$

where $\text{Sigmoid}(\cdot)$ denotes the sigmoid operation, and ω is 2. Finally, the output feature $\mathbf{X}_{\text{o}}$ is defined as follows:

$$\mathbf{X}_{\text{o}} = \mathbf{X}_{\text{u}} \otimes \text{Reverse}(\mathbf{A}_{\text{am}}) + \mathbf{X}_{\text{d}} \otimes \mathbf{A}_{\text{am}}, \tag{7}$$

where $\text{Reverse}(\cdot)$ is a reverse operation [Chen *et al.*, 2018b] .

Cascaded Feature Propagation. Benefiting from the advantages of the frequency prior feature, we first embed frequency priors into the reconciled features. Since the feature information embedded in the frequency priors has a strong representation ability, in order to further make full use of this information, we adopt a cascade method to propagate the information from the low-resolution deep semantics to the shallow high-resolution feature space. Therefore, in the FPEM, there are two cascaded frequency embedding components. The first frequency embedding component uses the frequency priors embedded into the coordinated feature as input to obtain the frequency-embedded coordinated feature. Then the second frequency embedding component takes the output of the first frequency embedding component and the output of the upper layer FPEM as input to realize the transfer from deep semantics to shallow semantics. Considering that the deepest features are missing the output of the previous level, the $\mathbf{E}_6$ is obtained from $\mathbf{X}_6$ by using a transformer block to enrich the global semantics.

3.6 Loss Function

Ground truth supervises the predictions of six FPEMs and $\mathbf{E}_6$ with the same loss function $\mathcal{L}$, which consists of a weighted Intersection over Union (IoU) loss $\mathcal{L}_{\text{IoU}}^{\omega}$ [Wei *et al.*, 2020] and a weighted Binary Cross Entropy (BCE) loss $\mathcal{L}_{\text{BCE}}^{\omega}$ [Wei *et al.*, 2020] defined as follows:

$$\mathcal{L}(\mathbf{P}, \mathbf{G}) = \mathcal{L}_{\text{IoU}}^{\omega}(\mathbf{P}, \mathbf{G}) + \mathcal{L}_{\text{BCE}}^{\omega}(\mathbf{P}, \mathbf{G}), \tag{8}$$

where $\mathbf{P}$ and $\mathbf{G}$ denote the prediction and the ground truth.

4 Experiments

4.1 Experimental Settings

Dataset. We performed extensive experiments on a large-scale benchmark dataset, DIS5K [Qin *et al.*, 2022], which contains a total of 5,470 images from 225 categories. The entire dataset is divided into three subsets: DIS-TR, DIS-VD and DIS-TE. DIS-TR and DIS-VD contain 3,000 training images and 470 validation images, respectively. DIS-TE is further split into four subsets (DIS-TE1, 2, 3, 4) with ascending shape complexities, each containing 500 images. The DIS5K dataset covers diverse objects with different geometric structures and appearances. Meanwhile, it provides images with higher resolution, more complex structural details, and higher annotation accuracy than existing object segmentation datasets. Therefore, segmentation on DIS5K is challenging and demands models with solid capabilities in identifying structural details.

Implementation Details. The model is implemented with the Pytorch framework on an A100 GPU. In the training stage, the ResNet-50 [He *et al.*, 2016] is pre-trained on ImageNet-1K [Deng *et al.*, 2009], the rest of the modules are initialized randomly. And we use the Adam optimizer with an initial learning rate of 1e-4, decaying by 10 every 50 epochs. The number of training epochs is set to 200. The resampled images with a size of 1024×1024 are fed into the proposed FP-DIS to get segmentation results in an end-to-end manner. No post-processing is needed throughout the inference process.

Evaluation Metrics. To make a comprehensive and fair comparison, we adopt five widely-used metrics to evaluate the performance, including maximal F-measure F_β^m [Achanta *et al.*, 2009], weighted F-measure F_β^w [Margolin *et al.*, 2014], Mean Absolute Error M, Structure Measure S_α [Fan *et al.*, 2017] and Mean Enhanced Alignment Measure E_ϕ^m [Fan *et al.*, 2018]. F_β^m and F_β^w consider both precision and recall in the binary classification of all the pixels. M is a measure of element-wise difference between the prediction and the paired ground truth mask. S_α is an effective indicator for evaluating structural similarity at region and object level. E_ϕ^m is widely used for evaluating pixel-level and image-level matching between the prediction and ground truth.

4.2 Comparison with State-of-the-arts

Quantitative Evaluation. Tables 1 and 2 show the comparison results of FP-DIS and 10 other methods, including UNet [Ronneberger *et al.*, 2015], BASNet [Qin *et al.*, 2019], GCPANet [Chen *et al.*, 2020], U^2Net [Qin *et al.*, 2020], SINetV2 [Fan *et al.*, 2021a], PSPNet [Zhao *et al.*, 2017], DLV3+ [Chen *et al.*, 2018a], HRNet [Wang *et al.*, 2020], STDC [Fan *et al.*, 2021b], and IS-Net [Qin *et al.*, 2022]. The experiment results demonstrate that our model significantly outperforms other models across all five evaluation metrics. Specifically, taking all four test subsets into consideration (as shown in Table 2, using dataset DIS-TE(1-4)), the proposed method obtains remarkable advancements even compared with the second-best model, improving 3.2%, 4.4%, 1.3%, 2.8% and 3.7% of the F_β^m, F_β^w, M, S_α and E_ϕ^m, respectively. These comparison results demonstrate the effectiveness of our model.

Qualitative evaluation. Figure 5 shows the visual results of ours and typical methods. Compared with other methods, our model can capture more fine details from complex backgrounds, and accurately segment objects with complex structures. Furthermore, it can be seen from Figure 6 that the feature embedding frequency priors are more delicate, which can significantly improve the segmentation performance.

4.3 Ablation Study

In this section, we perform comprehensive ablation experiments on DIS-TE1. First, we investigate the frequency priors generated using different filters. Then we make a comparison of models with the feature harmonization module and frequency prior embedding module, respectively.

Frequency Prior Generator. The frequency prior generator contains discrete cosine transform and inverse discrete cosine

Dataset	DIS-TE1					DIS-TE2					DIS-TE3				
Metric	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
UNet	0.625	0.514	0.106	0.716	0.750	0.703	0.597	0.107	0.755	0.796	0.748	0.644	0.098	0.780	0.827
BASNet	0.688	0.595	0.084	0.754	0.801	0.755	0.668	0.084	0.786	0.836	0.785	0.696	0.083	0.798	0.856
GCPANet	0.598	0.495	0.103	0.705	0.750	0.673	0.570	0.109	0.735	0.786	0.699	0.590	0.109	0.748	0.801
U^2Net	0.694	0.601	0.083	0.760	0.801	0.756	0.668	0.085	0.788	0.833	0.798	0.707	0.079	0.809	0.858
SINetV2	0.644	0.558	0.094	0.727	0.791	0.700	0.618	0.099	0.753	0.823	0.730	0.641	0.096	0.766	0.849
PSPNet	0.645	0.557	0.089	0.725	0.791	0.724	0.636	0.092	0.763	0.828	0.747	0.657	0.092	0.774	0.843
DLV3+	0.601	0.506	0.102	0.694	0.772	0.681	0.587	0.105	0.729	0.813	0.717	0.623	0.102	0.749	0.833
HRNet	0.668	0.579	0.088	0.742	0.797	0.747	0.664	0.087	0.784	0.840	0.784	0.700	0.080	0.805	0.869
STDC	0.648	0.562	0.090	0.723	0.798	0.720	0.636	0.092	0.759	0.834	0.745	0.662	0.090	0.771	0.855
IS-Net	0.740	0.662	0.074	0.787	0.820	0.799	0.728	0.070	0.823	0.858	0.830	0.758	0.064	0.836	0.883
FP-DIS	**0.784**	**0.713**	**0.060**	**0.821**	**0.860**	**0.827**	**0.767**	**0.059**	**0.845**	**0.893**	**0.868**	**0.811**	**0.049**	**0.871**	**0.922**

Table 1: Comparisons of different methods on different subsets of DIS5K, including DIS-TE1, DIS-TE2, and DIS-TE3 in terms of F_β^m, F_β^w, M, S_α, and E_ϕ^m. $\uparrow$ denotes larger is better, while $\downarrow$ represents smaller is better.

Dataset	DIS-TE4					DIS-TE(1-4)					DIS-VD				
Metric	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
UNet	0.759	0.659	0.102	0.784	0.821	0.708	0.603	0.103	0.759	0.798	0.692	0.586	0.113	0.745	0.785
BASNet	0.780	0.693	0.091	0.794	0.848	0.752	0.663	0.086	0.783	0.835	0.731	0.641	0.094	0.768	0.816
GCPANet	0.670	0.559	0.127	0.723	0.767	0.660	0.554	0.112	0.728	0.776	0.648	0.542	0.118	0.718	0.765
U^2Net	0.795	0.705	0.087	0.807	0.847	0.761	0.670	0.083	0.791	0.835	0.748	0.656	0.090	0.781	0.823
SINetV2	0.699	0.616	0.113	0.744	0.824	0.693	0.608	0.101	0.747	0.822	0.665	0.584	0.110	0.727	0.798
PSPNet	0.725	0.630	0.107	0.758	0.815	0.710	0.620	0.095	0.755	0.819	0.691	0.603	0.102	0.744	0.802
DLV3+	0.715	0.621	0.111	0.744	0.820	0.678	0.584	0.105	0.729	0.810	0.660	0.568	0.114	0.716	0.796
HRNet	0.772	0.687	0.092	0.792	0.854	0.743	0.658	0.087	0.781	0.840	0.726	0.641	0.095	0.767	0.824
STDC	0.731	0.652	0.102	0.762	0.841	0.710	0.628	0.094	0.754	0.832	0.696	0.613	0.103	0.740	0.817
IS-Net	0.827	0.753	0.072	0.830	0.870	0.799	0.726	0.070	0.819	0.858	0.791	0.717	0.074	0.813	0.856
FP-DIS	**0.846**	**0.788**	**0.061**	**0.852**	**0.906**	**0.831**	**0.770**	**0.057**	**0.847**	**0.895**	**0.823**	**0.763**	**0.062**	**0.843**	**0.891**

Table 2: Comparisons of different methods on DIS5K datasets, including DIS-TE4, DIS-TE(1-4), and DIS-VD.

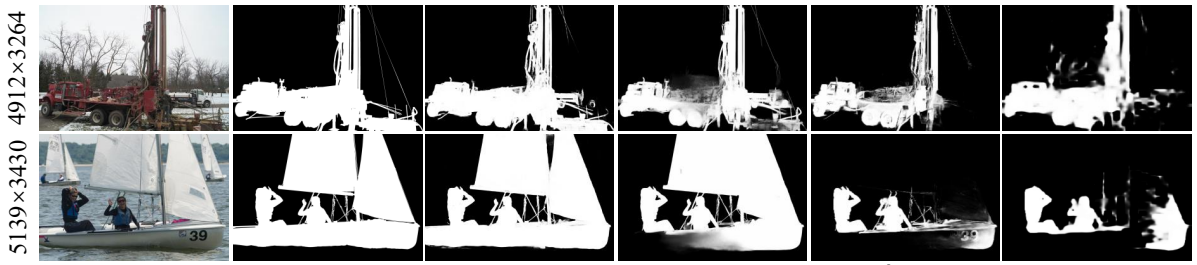

Figure 5: Visual comparison of the proposed method and cutting-edge methods.

transform using different filters. Therefore, we conduct three sets of experiments to verify the effectiveness of these filters with various settings. Actually, in the "w/o FPG" setting, we replace the FPG module with a convolutional layer, which means that the input image is processed by a convolutional layer before being fed into the FPEM module. As shown in Table 3, there is no significant difference in the performance using fixed or learnable filters in the FPG module, but the combination of these two kinds of filters contributed to better performance, indicating that the two filters are mutually beneficial and can effectively improve model's ability to detect detailed information. We also investigate the impact of changing the number of dynamic filter groups (1, 2, 3, 4, default is 2) in the feature embedding component. The results are shown in Table 4. Figure 6 shows the heatmaps of the backbone are relatively rough, and the heatmaps generated by the feature harmonization module and feature prior embedding module have more refined details.

Feature Harmonization Module. The feature harmonization module aims to reduce the variability between multi-

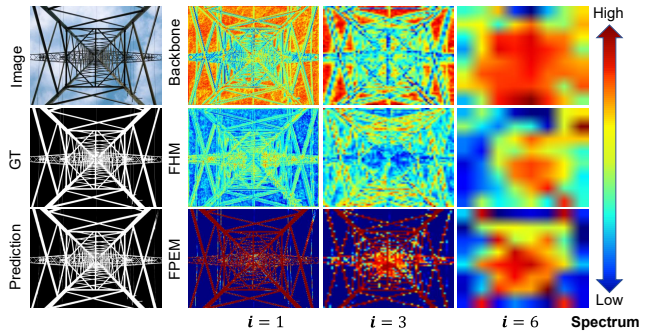

Figure 6: Visualization of the heatmaps at different stages of the decoder.

Settings	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
w/o FPG	0.765	0.679	0.069	0.805	0.842
Fixed	0.770	0.701	0.063	0.814	0.853
Learnable	0.782	0.712	0.060	0.819	0.858
Fixed+Learnable	**0.784**	**0.713**	**0.060**	**0.821**	**0.860**

Table 3: Effectiveness of frequency prior generator on DIS-TE1. Frequency prior generator adopt different types of filter combinations. w/o FPG: without frequency prior generator module and replaced it with a convolutional layer, Fixed: only fixed filter, learnable: only learnable filters.

Model	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
filter $\times 1$	0.768	0.686	0.066	0.809	0.848
filter $\times 3$	0.772	0.695	0.063	0.814	0.852
filter $\times 4$	0.768	0.686	0.066	0.809	0.845
Ours	**0.784**	**0.713**	**0.060**	**0.821**	**0.860**

Table 4: Ablation study of the number of dynamic filter groups on DIS-TE1.

scale information and help the model to extract rich information from different scale features. From the experiment results shown in Table 5, we can see that the feature harmonization module contributes significantly to the performance of the model. Considering that the structure of the feature harmonization module may have an impact on the model, we list seven different grouping and separation schemes, and the experiments confirm that our proposed method performs better and has a more rational design. We also conduct experiments on removing the intra-layer and inter-layer harmonization components, as well as the gating mechanisms in the feature harmonization component, to study their contributions. The results shown in Table 6 demonstrate that these changes always lead to performance degradation compared to the proposed method, which verifies the benefits of each component. Comparing the output features from the pyramid feature extractor and harmonization features in Figure 6, we can see that the FHM effectively reduces ambiguity, and the harmonization features have higher confidence.

Frequency Prior Embedding Module. To verify the effectiveness of the frequency prior embedding module, we replace the FEC in different positions with addition operations. The results, shown in Table 7, reveal that FPEM is beneficial to improve the segmentation performance. The results of

Settings	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
w/o FHM	0.765	0.689	0.066	0.809	0.847
FHM ($2g$+$3s$)	0.779	0.710	0.060	0.819	0.858
FHM ($3g$+$3s$)	0.776	0.703	0.062	0.816	0.854
FHM ($4g$+$3s$)	**0.784**	**0.713**	**0.060**	**0.821**	**0.860**
FHM ($5g$+$3s$)	0.776	0.709	0.062	0.818	0.857
FHM ($6g$+$3s$)	0.783	0.708	0.062	0.818	0.857
FHM ($4g$+$2s$)	0.774	0.699	0.064	0.813	0.853
FHM ($4g$+$4s$)	0.778	0.706	0.062	0.816	0.856
FHM ($4g$+$5s$)	0.778	0.703	0.062	0.816	0.854

Table 5: Comparisons of feature harmonization module structure on DIS-TE1. w/o FHM: without FHM. FHM $(ig + js), i \in \{2, 3, 4, 5, 6\}, j \in \{2, 3, 4, 5\}$: with i groups and j sub-groups.

Model	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
w/o all-HC	0.760	0.676	0.068	0.802	0.838
w/o intral-HC	0.764	0.683	0.068	0.805	0.843
w/o inter-HC	0.760	0.678	0.070	0.802	0.842
w/o gate	0.770	0.687	0.066	0.809	0.848
Ours	**0.784**	**0.713**	**0.060**	**0.821**	**0.860**

Table 6: Ablation study of the harmonization components in FHM on DIS-TE1.

Settings	$F_\beta^m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$
w/o adapt	0.759	0.678	0.069	0.805	0.840
w/o FPEM	0.765	0.687	0.067	0.807	0.850
FEC2	0.781	0.709	0.060	0.818	0.859
FEC1	0.768	0.693	0.065	0.810	0.851
FEC1+FEC2	**0.784**	**0.713**	**0.060**	**0.821**	**0.860**

Table 7: Comparisons of frequency prior embedding module on DIS-TE1. w/o FPEM: Without frequency prior embedding module; Other is feature embedding component of different stages in FPEM.

w/o FPEM prove that there is indeed a semantic gap between frequency prior feature and image features. Additionally, we replace the adaptive modulation with an addition operation to evaluate its effectiveness (i.e., w/o adapt). The adaptive modulation strategy can reduce this gap, improving the segmentation accuracy of the model. According to the heatmap from the frequency prior embedding module in Figure 6, the detected objects have better refine textures.

5 Conclusions

In this paper, we have proposed a novel DIS model, FP-DIS, which can generate frequency priors to guide fine-grained segmentation. We innovatively embed frequency priors into the image features to achieve accurate DIS. For this purpose, we adapt the frequency prior feature through dynamic filters to extract accurate frequency information. Meanwhile, we reduce the heterogeneity in multi-scale image features by harmonizing the features. Finally, we propose a frequency prior embedding module, which provides efficient frequency priors embedding for predicting DIS maps. Extensive experiments demonstrate the advantages of the proposed model over the existing models. Further ablation experiments sufficiently verify the effectiveness of the proposed modules.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62201465 and 62001425, and the Fundamental Research Funds for the Central Universities under Grant D5000220213.

Contribution Statement

Yan Zhou and Bo Dong contribute equally to this work.

References

[Achanta *et al.*, 2009] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE, 2009.

[Chen *et al.*, 2018a] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[Chen *et al.*, 2018b] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 234–250, 2018.

[Chen *et al.*, 2020] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, volume 34, pages 10599–10606, 2020.

[Chen *et al.*, 2022] Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *TCSVT*, 32(10):6981–6993, 2022.

[De Carvalho *et al.*, 2013] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *TIFC*, 8(7):1182–1194, 2013.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[Dong *et al.*, 2021] Bo Dong, Yan Zhou, Chuanfei Hu, Keren Fu, and Geng Chen. Bcnet: Bidirectional collaboration network for edge-guided salient object detection. *Neurocomputing*, 437:58–71, 2021.

[Dong *et al.*, 2023a] Bo Dong, Pichao Wang, and Fan Wang. Head-free lightweight semantic segmentation with linear transformer. *AAAI*, 2023.

[Dong *et al.*, 2023b] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI*, 2023.

[Durall *et al.*, 2019] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 7 2018.

[Fan *et al.*, 2020] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273. Springer, 2020.

[Fan *et al.*, 2021a] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *TPMAI*, 2021.

[Fan *et al.*, 2021b] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *CVPR*, pages 9716–9725, 2021.

[Geiger *et al.*, 2011] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 963–968. Ieee, 2011.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2017] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, pages 1689–1697, 2017.

[Ji *et al.*, 2022] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, pages 1–19, 2022.

[Kawar *et al.*, 2023] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

[Li *et al.*, 2018] Junxuan Li, Shaodi You, and Antonio Robles-Kelly. A frequency domain neural network for fast image super-resolution. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[Li *et al.*, 2021a] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *CVPR*, pages 10071–10081, 2021.

[Li *et al.*, 2021b] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *TCSVT*, 32(4):2303–2314, 2021.

[Lin *et al.*, 2022] Yi Lin, Jichun Wu, Guobao Xiao, Junwen Guo, Geng Chen, and Jiayi Ma. BSCA-Net: Bit slicing context attention network for polyp segmentation. *Pattern Recognition*, 132:108917, 2022.

[Liu *et al.*, 2021] Jiannan Liu, Bo Dong, Shuai Wang, Hui Cui, Deng-Ping Fan, Jiquan Ma, and Geng Chen. COVID-19 lung infection segmentation with a novel two-stage cross-domain transfer learning framework. *Medical image analysis*, 74:102205, 2021.

[Liu *et al.*, 2023] Shaolei Liu, Siqi Yin, Linhao Qu, and Manning Wang. Reducing domain gap in frequency and spatial domain for cross-modality domain adaptation on medical image segmentation. *AAAI*, 2023.

[Margolin *et al.*, 2014] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014.

[Pandey *et al.*, 2016] Ramesh C Pandey, Sanjay K Singh, and Kaushal K Shukla. Passive forensics in image and video using noise features: A review. *Digital Investigation*, 19:1–28, 2016.

[Pang *et al.*, 2022] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.

[Qin *et al.*, 2020] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[Qin *et al.*, 2022] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.

[Ren *et al.*, 2022] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *CVPR*, pages 10853–10862, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Singh *et al.*, 2020] Ravi Pratap Singh, Mohd Javaid, Ravinder Kataria, Mohit Tyagi, Abid Haleem, and Rajiv Suman. Significant applications of virtual reality for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):661–664, 2020.

[Stuchi *et al.*, 2017] José A Stuchi, Marcus A Angeloni, Rodrigo F Pereira, Levy Boccato, Guilherme Folego, Paulo VS Prado, and Romis RF Attux. Improving image classification with frequency domain layers for feature extraction. In *Machine Learning for Signal Processing*, pages 1–6. IEEE, 2017.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPMAI*, 43(10):3349–3364, 2020.

[Wei *et al.*, 2020] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, volume 34, pages 12321–12328, 2020.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[Wu *et al.*, 2019] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.

[Zhai *et al.*, 2021] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *CVPR*, pages 12997–13007, 2021.

[Zhang *et al.*, 2022a] Bing Zhang, Yuanfeng Wu, Boya Zhao, Jocelyn Chanussot, Danfeng Hong, Jing Yao, and Lianru Gao. Progress and challenges in intelligent remote sensing satellite systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[Zhang *et al.*, 2022b] Fengyu Zhang, Ashkan Panahi, and Guangjun Gao. Fsanet: Frequency self-attention for semantic segmentation. *arXiv preprint arXiv:2211.15595*, 2022.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[Zhong *et al.*, 2022] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *CVPR*, pages 4504–4513, 2022.