High-Precision Dichotomous Image Segmentation With Frequency and Scale Awareness

Qiuping Jiang, *Senior Member, IEEE*, Jinguang Cheng, Zongwei Wu, Runmin Cong, *Senior Member, IEEE*, and Radu Timofte, *Member, IEEE*

***Abstract*— Dichotomous image segmentation (DIS) with rich fine-grained details within a single image is a challenging task. Despite the plausible results achieved by deep learning-based methods, most of them fail to segment generic objects when the boundary is cluttered with the background. In fact, the gradual decrease in feature map resolution during the encoding stage and the misleading texture clue may be the main issues. To handle these issues, we devise a novel frequency- and scale-aware deep neural network (FSANet) for high-precision DIS. The core of our proposed FSANet is twofold. First, a multimodality fusion (MF) module that integrates the information in spatial and frequency domains is adopted to enhance the representation capability of image features. Second, a collaborative scale fusion module (CSFM) which deviates from the traditional serial structures is introduced to maintain high resolution during the entire feature encoding stage. In the decoder side, we introduce hierarchical context fusion (HCF) and selective feature fusion (SFF) modules to infer the segmentation results from the output features of the CSFM module. We conduct extensive experiments on several benchmark datasets and compare our proposed method with existing state-of-the-art (SOTA) methods. The experimental results demonstrate that our FSANet achieves superior performance both qualitatively and quantitatively. The code will be made available at https://github.com/chasecjg/FSANet.**

***Index Terms*— Dichotomous image segmentation (DIS), frequency awareness, high precision, scale awareness.**

I. Introduction

DICHOTOMOUS image segmentation (DIS) aims to segment one or more generic objects with high accuracy from a single image [1]. High-precision DIS offers precise geometric object descriptions, with a wide range of applications, making it increasingly critical to meet the demands of refined human–machine interaction. For example, DIS can improve the performance of image editing by enabling refined synthesis and editing operations [2], [3], enhance virtual reality experiences [4], [5], facilitate medical image analysis for precise lesion localization [6], [7], and empower robots' ability to perceive their surroundings accurately [8], [9].

Driven by the deep learning techniques, existing models for salient object detection (SOD) [10], [11], [12] and camouflaged object detection (COD) [13], [14] have made significant progress. While these models have shown great effectiveness in specific tasks, directly applying them to high-precision DIS remains a big challenge. This is mainly due to the difference between SOD/COD and generic object segmentation tasks. Specifically, SOD aims to highlight the most visually salient objects within a single image, while COD targets to precisely localize and segment the object within a camouflaged scene context. Therefore, it is nontrivial to directly extend the existing SOD and COD models to address the more general DIS task [1], [3], [15].

To meet the demand for precise geometric object delineation, some researchers have explored image matting techniques as an alternative. Image matting tasks [16], [17], [18], [19], similar to high-precision DIS, have demonstrated promising results. However, the approaches for these tasks typically involve the combination of RGB images with trimap annotations, which requires labor-intensive and expensive manual acquisition [20], [21], [22]. Such a reliance limits the scalability and practicality of achieving high-precision segmentation. Recently, there have been several networks specifically designed for DIS [1], [23], [24], [25], which perform well in general settings. However, as shown in Fig. 1, we observe that even the state-of-the-art (SOTA) method [1] fails to reason about accurate segmentation when dealing with these challenging scenes. Since these methods are solely based on RGB visual clues, their performance in cluttered areas may deteriorate significantly. Hence, there is a pressing need to develop methods that can surpass the limitations of existing techniques.

In this work, we propose a novel approach that addresses these challenges by leveraging the frequency domain information derived directly from RGB images. Our motivation comes from the observation that, as shown in Fig. 1, frequency information contains high-frequency components that capture subtle details and fine textures of objects, thereby enhancing the preservation of details in segmentation. Meanwhile, frequency-based features are less sensitive to image variations, such as lighting changes, occlusions, and background interfer-

Manuscript received 11 November 2023; revised 16 May 2024; accepted 7 July 2024. Date of publication 16 August 2024; date of current version 5 May 2025. This work was supported in part by the Natural Science Foundation of Zhejiang under Grant LR22F020002, in part by the Natural Science Foundation of China under Grant 62271277, in part by the Key Research and Development Plan of Ningbo under Grant 2024Z292, in part by the Natural Science Foundation of Ningbo under Grant 2022J081, in part by Taishan Scholar Project of Shandong Province under Grant TSQN202306079, and in part by Xiaomi Young Talents Program. *(Corresponding authors: Jinguang Cheng; Zongwei Wu.)*

Qiuping Jiang and Jinguang Cheng are with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jiangqiuping@nbu.edu.cn; chase.jgcheng@gmail.com).

Zongwei Wu and Radu Timofte are with the Computer Vision Laboratory, IFI and CAIDAS, University of Würzburg, 97074 Würzburg, Germany (e-mail: zongwei.wu@uni-wuerzburg.de; radu.timofte@uni-wuerzburg.de).

Runmin Cong is with the School of Control Science and Engineering, Shandong University, Jinan 250100, China (e-mail: rmcong@sdu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2024.3426529

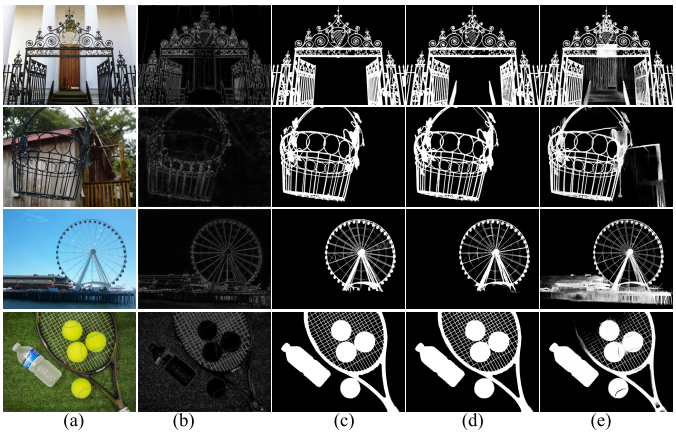

Fig. 1. Motivation of using frequency awareness. Given an input image (a), learnable high-pass filter in the frequency domain is applied to obtain the activation map (b). We observe that the high-pass filter can attenuate the misleading texture clues while preserving and enhancing the global structure of different objects. Inspired by this observation, we aim to make full use of the frequency information to improve feature modeling with strengthened discriminability in the context of DIS. (e) Compared to the recent DIS model [1], (d) FSANet can produce the segmentation result, and (c) closer to the ground truth.

ence, making them more reliable and robust under visually challenging scenarios. The rich boundary texture features and morphological structure information are closely associated with object boundaries, which are essential for achieving high-precision segmentation. Therefore, we aim to leverage the frequency-based features to enhance the feature modeling process. Our key idea is to mimic the advantages of multisensor fusion approaches, such as RGB-D [26], [27], [28], [29] or RGB-T [30], [31], [32], without requiring additional sensor inputs. Such a design could contribute to superior segmentation accuracy, surpassing the limitations of existing models and reducing the burden of manual processes, such as multimodal data collection or trimap processing.

Meanwhile, most existing segmentation networks [1], [14], [33], [34], [35] predominantly rely on the UNet [36] architecture, as shown in Fig. 2(a). While powerful in feature extraction and context awareness, they face limitations in DIS tasks, particularly in cases involving information loss across scales. Inspired by the recent success of high-resolution architectures such as HRNet [37] and its successors [38], [39], [40], as shown in Fig. 2(b), we propose a novel fusion strategy by combining HRNet with UNet to boost the accuracy of high-precision DIS during the feature modeling, as shown in Fig. 2(c). Our key idea is to directly correlate the multiscale features with the network hierarchy, getting rid of the constraint in HRNet that all resolutions need to be calculated parallel for a newly added stage. By integrating the multiscale granularity with fine-grained details, we can lead to more accurate DIS with sharpened edges. To conclude, our contributions are summarized as follows.

1) We propose a novel approach called frequency- and scale-aware deep neural network (FSANet) for high-precision DIS by leveraging frequency domain information, improving segmentation accuracy by capturing subtle details and fine textures while being robust to image variations.
2) We introduce a cross-resolution fusion strategy that combines HRNet with UNet to address information loss across scales, leading to more accurate DIS with sharpened edges.
3) Our method outperforms the SOTA methods and sets new records on both benchmark DIS datasets and other real-world applications, validating the effectiveness and robustness.

The rest of this article is organized as follows. Section II describes the related works. Section III illustrates the details of the proposed FSANet. Section IV presents the experimental results. Finally, conclusions are drawn in Section V.

II. Related Work

A. Binary Image Segmentation

1) Salient Object Detection: At the early stage, researchers concentrated on identifying the most salient object within an image. These works were primarily based on manually designed low-level features, such as contrast [41], boundary background [42], [43], center prior [44], [45], and others [46], [47]. In recent years, learning-based models have achieved impressive performance. For instance, Wang et al. [48] use two subnetworks to search for global and local information, establishing a relationship between local and global information. Long et al. [49] proposed the FCN to predict the semantic labels of each pixel. Deng et al. [50] used residual networks for network optimization, enhancing saliency details while suppressing nonsalient regions in intermediate saliency maps. Hu et al. [51] introduced a new FCN-based method called RADF, which cyclically aggregates deep multilevel features into each layer's features. While these methods generally perform well, they may struggle to produce accurate results in challenging scenarios, such as those involving occlusion and mixed foreground–background situations. Some researchers have introduced additional modality information such as infrared images and depth images in the input side [26], [27], [28], [30], [31], [32], [52]. However, such a multimodal setting with aligned input and accurate measurement is hard to achieve in practice.

2) Camouflage Object Detection: To tackle the challenge with mixed foreground–background, researchers have focused on a more challenging task called COD [14], [53]. Fan et al. [14] proposed a search and recognition approach inspired by predatory animal behavior. This method locates the object's position and then refines the segmentation results to improve performance. By going beyond the texture features, Ji et al. [54] proposed DGNet, which introduces gradient information to supervise the texture branch. Despite the plausible performance, due to the inherent network design, existing works can only tackle one semantic, i.e., the camouflaged objects, but fail to segment generic ones [55]. Therefore, it is nontrivial to directly extend such models to the DIS task.

3) Image Matting and DIS: Image matting is another task similar to DIS, aiming to accurately separate foreground objects in an image while preserving details and

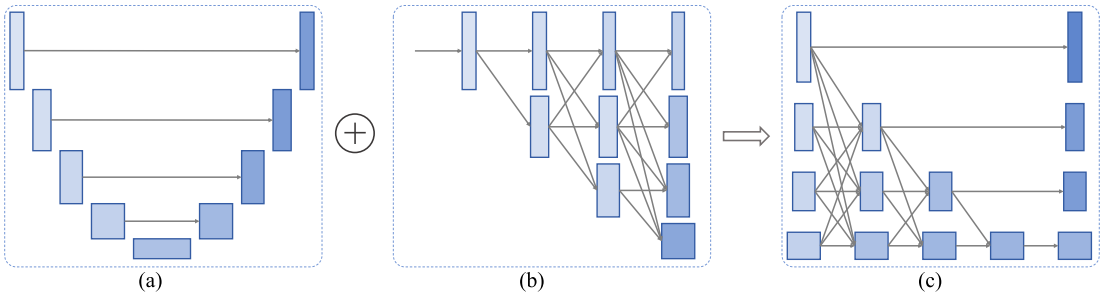

Fig. 2. Cross-scale interaction. (a) Well-acknowledged U-shaped architecture has demonstrated significant benefits in image segmentation. However, its limitation in cross-scale interaction hinders its effectiveness for high-precision segmentation tasks. (b) Recent HRNet has distinguished itself by preserving resolution and rich semantic information. Nevertheless, in the decoding stage, the smallest branch is directly treated as the output, failing to fully exploit the network hierarchy. (c) In this article, we propose a novel architecture by combining the retaining the advantages of both UNet and HRNet. In the encoding stage, we facilitate multiscale fusion to maximize the interaction among feature maps of varying resolutions. In the decoding part, we carefully analyze semantic clues to generate the final output in a gradual and attentive manner.

maintaining edge authenticity in diverse backgrounds. Recent learning-based models have also demonstrated great performance and accuracy [20], [21], [22]. However, most image matting methods require both RGB and trimap as input, with the latter heavily depending on labor-intensive labeling or preprocessing. Therefore, such a setting differs from DIS, which aims to achieve high-precision segmentation with RGB-only input and in an end-to-end manner. One pioneering work was realized by Qin et al. [1], who built a large-scale high-resolution dataset called DIS5K, thus greatly facilitating the development of high-precision image binary segmentation tasks. Inspired by this, Pei et al. [24] divide and conquer and reconstruct complementary features to achieve the effect of simultaneously enhancing the subject area and structural details, thereby improving high-precision segmentation of targets. Liu et al. [23] further exploit the DIS task on medical image segmentation.

However, most existing works only rely on visual clues for object segmentation, making it difficult to perform well when dealing with cluttered backgrounds. In this article, we contribute from a novel perspective by exploiting frequency domain information to enhance feature modeling. This design leads to improved performance and increased robustness when dealing with challenging scenes.

B. Serial and Parallel Networks

Resolution plays a crucial role in location-sensitive visual tasks, such as object detection [56], [57], [58], image segmentation [49], [59], and image editing [2]. In image segmentation tasks, existing methods are mostly based on a serially designed structure, such as UNet [36]. The encoding part of these methods employs a serial connection to downsample the input image from high resolution to low resolution, and the decoding part subsequently upscales the low-resolution image back to the original size. Although these methods [13], [14] have achieved good performance in image segmentation task, they suffer from the loss of low-level features caused by downsampling, which limits their performance in high-precision DIS. Recently, Sun et al. [37] proposed a parallel network structure named HRNet, which maintains high-resolution representations throughout the entire process to capture richer semantic and spatial information. Based on this, Wu et al. [40] proposed an improved version of HRNet by introducing mixed dilated convolution and multilevel data-dependent feature aggregation. Starting from the parameter quantity of HRNet, Yu et al. [38] introduced an efficient shuffle block in ShuffleNet and proposed a lightweight Lite-HRNet.

Despite the plausible performance achieved by HRNet, [60] suggests that the insufficient information exchange between different resolutions in HRNet makes it imperfectly suitable for dense prediction. As an alternative, they propose a U-HRNet network where each encoding and decoding stage is composed of cross-scale interaction formed by no more than two resolution branches. However, such a kind of architecture also has two drawbacks. First, the two-branch architecture limits its capability for information exchange. Second, the full HR-module architecture for both encoding and decoding may lead to significant computational costs.

We observe that the network hierarchy correlates well with the multiresolution design of HRNet. Therefore, in this work, we propose an improved UNet with hierarchical awareness to profit from both fine-grained details (higher resolution) and rich semantic information (lower resolution), leading to a more accurate DIS. Compared to U-HRNet [60], we leverage features from all hierarchies to maximize the cross-scale interaction during encoding. Moreover, our decoder is different since we dig into semantic clues and integrate them with fine-grained details. These designs make our proposed FSANet efficient and set new SOTA records in several benchmark datasets of DIS.

III. Proposed Method

Fig. 3 illustrates our proposed FSANet, which leverages both frequency domain assistance and cross-resolution awareness. FSANet comprises three main modules: the backbone module (BM), the collaborative scale fusion module (CSFM), and the high-precision perception fusion module (HPFM).

The BM plays a crucial role in extracting and fusing features from both the visual and frequency domains of the input image, thereby enhancing semantic features and compensating for the loss of fine details. CSFM adeptly combines multiscale granularity and network hierarchy to effectively model features with fine-grained details, enabling precise edge refinement. This module ensures efficient merge and control of features from different scales. Within HPFM, the hierarchical context fusion (HCF) compresses the output features from CSFM,

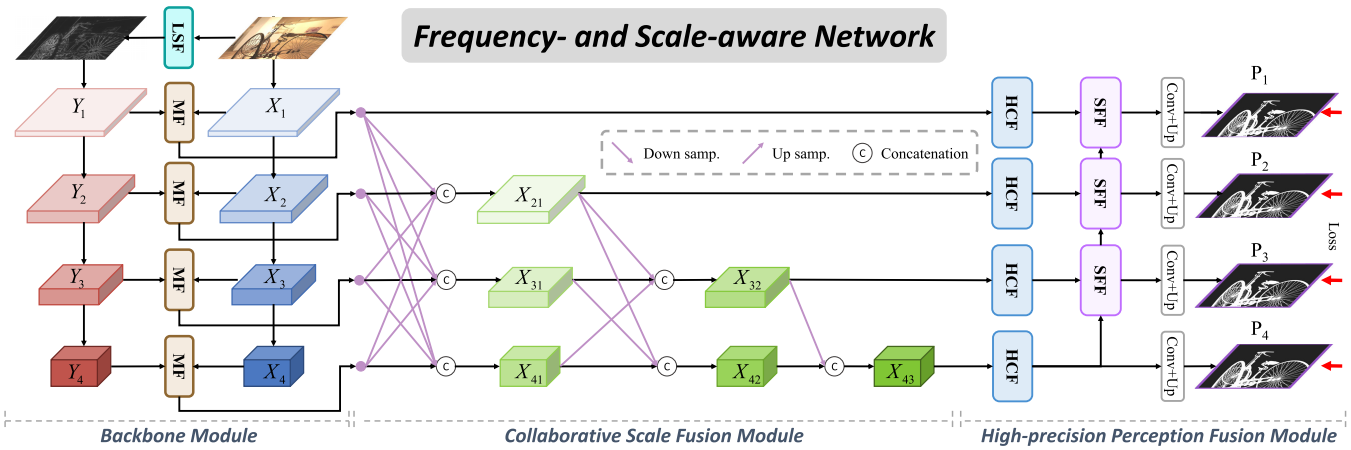

Fig. 3. Architecture overview: FSANet consists of three main components: encoder module BM (Section III-A), feature fusion module CSFM Section III-B, and decoder module HPFM Section III-C. In the encoding stage, the features received by the BM are divided into two processes: one for visual feature modeling and the other for frequency feature extraction with LSF Section III-A1, which enables RGB-Frequency transform. Then, the mimicked multimodal features are fused by the MF Section III-A3. To leverage multiscale information for high-precision detection, we introduce the CSFM to aggregate both high- and low-resolution clues for their rich fine-grained details and semantic clues, respectively. In the decoding stage, we first compress the features through HCF Section III-C1 while maintaining the most informative ones and then perform interparadigm feature fusion through SFF Section III-C2 module to achieve high-precision segmentation. FSANet is supervised by the ground truth to achieve end-to-end training.

reducing feature dimensionality while preserving essential information. This compression enhances efficiency without compromising key information. Finally, the selective feature fusion (SFF) in HPFM performs high-precision feature decoding, producing the binary segmentation maps P_1–P_4 from different scales. During training, these maps are supervised by ground truth. Our final prediction is P_1. By employing such a network architecture and training strategy, we optimize the segmentation performance and achieve high-precision segmentation results.

A. Backbone Module

This section consists of three main components: an encoder, a learnable spectral filtering (LSF), and a multimodal fusion (MF) module.

1) Learnable Spectral Filtering: Accurately describing and capturing object boundaries and morphological structures are crucial in the task of DIS [61], [62], [63], [64]. To achieve high-precision segmentation, we observe that the magnitude of image frequency can naturally serve as an indicator of boundary variations within an object. As suggested in previous works [64], [65], [66], [67], high-frequency information can effectively capture fine boundary details, texture features, and morphological structures present in the image. Furthermore, frequency-based features exhibit robustness against environmental changes, such as illumination variations, occlusion, and background interference, making them reliable in challenging scenarios. Based on these characteristics, we introduce a LSF to exploit high-frequency information in addition to the conventional visual clue.

Fig. 4 illustrates the details of the LSF. We begin by applying the fast Fourier transform (FFT) to the input image, followed by high-pass filtering with a learnable cut-off frequency parameter to remove low-frequency components. The resulting image, with low-frequency information eliminated, undergoes inverse Fourier Transform, yielding a

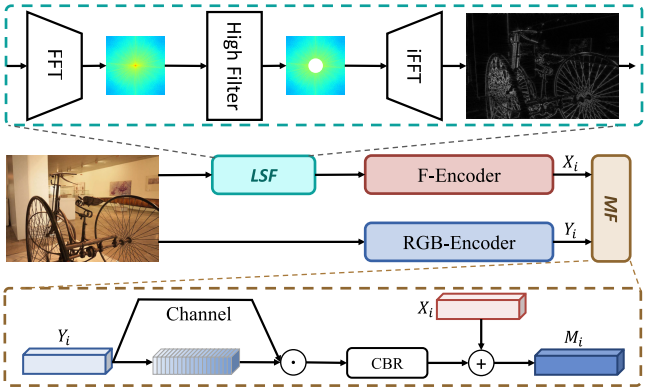

Fig. 4. Overview of the BM. From the RGB input, we first apply the high-pass filter in the frequency domain to generate the activation map from the input image. This activation map, rich in structural clues, provides an additional awareness that can contribute to attenuating perceptual uncertainty. To achieve such a goal, we adopt a dual encoder design to first extract features from different resources separately and then attentively fuse them to generate the encoded features with strengthened discriminability.

high-frequency feature map enriched with structural information. Mathematically, the specific implementation can be expressed as follows:

$$Y = i\mathcal{FFT}(h(\mathcal{FFT}(X), \beta)) \tag{1}$$

where X represents the input image, $\mathcal{FFT}$ denotes the FFT, $i\mathcal{FFT}$ stands for the inverse FFT (iFFT), and $h(I, \beta)$ represents a high-pass filter with the learnable parameter β.

2) Feature Extraction: Following the approach of Fan et al. [28], this article also employs a dual-path encoding design for feature extraction. However, what distinguishes our approach is the utilization of two distinct feature extraction backbone networks. Specifically, we employ a ResNet [68] backbone for frequency feature extraction, while a PVTv2 [69] backbone is employed for visual feature extraction. Such a choice is made building upon the observation that the high-frequency map is relatively simple compared to its corresponding RGB

counterpart. Therefore, a relatively simpler backbone is used for learning feature representations from the high-frequency map. As for the RGB input, which is richer in terms of perceptual clue, a deeper backbone is naturally required. Such a choice can contribute to reducing the number of parameters compared to using duplicated deep backbones while maintaining the effectiveness.

3) Multimodal Fusion: With the LSF module, we can obtain high-frequency information from the RGB input, making it possible to transfer our problem setting into the mimicked multimodal fusion design. The fusion of RGB and high-frequency information encounters two main challenges: 1) the compatibility between these modalities and 2) the redundancy and noise in low-quality frequency information. Fan et al. [28], we introduce a MF module to address these challenges, as shown in Fig. 4. The MF module aims to enhance the compatibility of multi-modal features and extract valuable information such as details, textures, and morphological structures from high-frequency information. Let X_i and Y_i represent the encoded output features from the RGB-encoder and the F-encoder, respectively. To enhance compatibility and promote cross-modal understanding, we first apply channel attention calculation, inspired by [70], followed by a convolutional block to process Y_i. Then, we perform feature fusion with X_i to facilitate information exchange between different modalities. Mathematically, the fusion process is expressed as follows:

$$M_i = \mathrm{CBR}(\varphi(Y_{i+1})) + X_i, \quad i \in (1, 2, 3, 4) \tag{2}$$

where CBR is the convolutional layer with batch normalization and ReLU activation, φ represents the channel attention from [70], and M_i represents the fused result.

B. Collaborative Scale Fusion Module

Traditional segmentation networks often adopt a serial structure based on UNet [36], which exhibits powerful feature extraction and context awareness capabilities. However, the pure serial structure suffers from feature information loss and blurring, limiting the accuracy and detail preservation ability in high-precision segmentation tasks. To address this limitation, we draw inspiration from HRNet [37] and introduce the CSFM.

The CSFM directly couples multiscale granularity with the network hierarchy to model features with fine-grained details, enabling more accurate edge-refined DIS, as shown in Fig. 3. Specifically, CSFM constructs a high-resolution feature flow by progressively downsampling or aggregating output features at each stage. During feature aggregation, the resolution of the output features in each layer is preserved, and low-level feature information is integrated into the high-level features. As a result, the output features in subsequent stages incorporate high-resolution information from all preceding layers. To facilitate understanding, the details of upsampling and downsampling operations in CSFM are provided in Table I. Through the collaborative scale fusion mechanism, we can effectively integrate local detailed information with global semantic clues, facilitating high-precision DIS with improved accuracy.

TABLE I

TECHNICAL DESIGNS FOR CSFM. Down_i REPRESENTS THE DOWNSAMPLING OPERATION WITH A SCALE OF i AND Up_i REPRESENTS THE UPSAMPLING OPERATION WITH A SCALE OF i. THE CSFM IS BUILT UPON CONVOLUTIONAL OPERATION. THE DETAILS REGARDING THESE OPERATIONS ARE PROVIDED BELOW

Sampling Module	Input Channels	Output Channels	Kernel Size	Stride	Padding
$Down_8$	in_chan	out_chan1	3	2	1
	out_chan1	out_chan2	3	2	1
	out_chan2	out_chan3	3	2	1
$Down_4$	in_chan	out_chan1	3	2	1
	out_chan1	out_chan2	3	2	1
$Down_2$	in_chan	out_chan	3	2	1
Up_n	in_chan	out_chan	1	1	0
	Upsample				

C. High-Precision Perception Fusion Module

To obtain an accurate binary segmentation map, the output features from the CSFM need to be decoded. However, considering the computational complexity, we apply a compression step to reduce the dimensionality of the features. Additionally, to address differences between various features, we adopt an interparadigm feature fusion method that aggregates features across different scales. This module consists of two components: HCF and SFF.

1) Hierarchical Context Fusion: The HCF addresses the dimensionality reduction of the feature map H_i. Since directly compressing the original features through convolution may result in information loss, we employ a staged strategy that combines multiple branches and convolutional layers for dimensionality reduction. This approach is depicted in Fig. 5. By effectively processing high-dimensional features, the HCF enhances the performance and accuracy of the segmentation model while mitigating the impact of information loss. The computational complexity is reduced while preserving the richness of feature information, thereby meeting the requirements of high-precision segmentation tasks.

Specifically, the HCF consists of five branches. In each branch, the first branch is a 1×1 convolution, and the last two layers of the first three branches are 1×3, and 3×1 convolutions, 1×5, 5×1, and 1×7, 7×1 convolutions, and the remaining branch ($k = 4$) performs a 1×1 convolution operation. These convolutional operations concatenate or add features and subsequently compress the output channels of the H_i features to 64 dimensions, resulting in the feature $S_i, i \in (1, 2, 3, 4)$. Please refer to Fig. 5 for a more detailed depiction of the HCF.

2) Selective Feature Fusion: Effectively merging features from different branches is a key topic for most UNet-based networks. Existing works often use concatenation–convolution for decoding fusion. However, such a design may lead to feature redundancy and cannot fully explore the differences between different decoding branches. To tackle this challenge, in this article, we propose a SFF design, as shown in Fig. 6. Our key idea is to mine the cross-branch features to attenuate the irrelevant feature response. Specifically, given two feature maps S_i and S_{i+1} from different feature layers as inputs to SFF, we first fuse them by mining their affinities. We can obtain

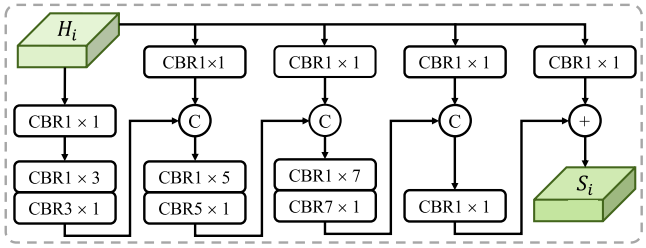

Fig. 5. HCF. Contextual clues play a vital role in high-precision DIS. Local context contributes to fine-grained details for sharp edge generation, while larger context provides semantic clues for object localization and segmentation. Therefore, we employ a multibranch fusion design, where each branch tackles a specific level of context clues. These branches are also inner-correlated to maximize the hierarchical information exchanges.

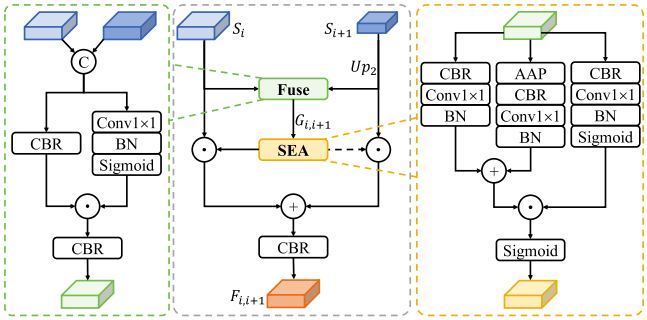

Fig. 6. SFF. To gradually and attentively fuse multilevel features during the decoding stage, we propose a selective fusion design based on semantic clues. We first merge the features by mining their affinities. Then, from the fused feature, we dig into the semantics clues and compute semantically enhanced attention. The generated attention map is applied to one input feature, whereas the opposite is applied to the other. Such a design enables a full exploration of semantics during the decoding stage.

the fused feature $G_{i,i+1}$ by the following equation:

$$G_{i,i+1} = \mathcal{F}\text{use}\big(S_i, \text{Up}_2(S_{i+1})\big). \tag{3}$$

The obtained $G_{i,i+1}$ is then fed into the semantic enhanced attention (SEA) to dig into semantic clues and compute the attention map, resulting in feature map $A_{i,i+1}$

$$A_{i,i+1} = \text{SEA}\big(G_{i,i+1}\big). \tag{4}$$

The technical details of $\mathcal{F}$use and SEA can be found in Fig. 6. Then, the attention map is applied to one input, whereas the opposite form is applied to the other. In such a manner, we can attentively improve the feature modeling for the cross-branch inputs. Finally, we merge the enhanced features to form the output. Mathematically, we can obtain the fused output $F_{i,i+1}$ by the following equation:

$$F_{i,i+1} = \text{CBR}\big(A_{i,i+1} \cdot S_i + \big(1 - A_{i,i+1}\big) \cdot (S_{i+1})\big). \tag{5}$$

D. Loss Function

Despite the mimicked multimodal fusion design, FSANet can be trained end-to-end. We leverage the GT mask for multiscale supervision. Therefore, our loss function consists of four main components

$$\mathcal{L} = \mathcal{L}_1 + \lambda\mathcal{L}_2 + \lambda\mathcal{L}_3 + \lambda\mathcal{L}_4 \tag{6}$$

where $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$ correspond to the loss functions for P_1, P_2, P_3, and P_4, respectively. P_i are the predicted segmentation masks. We introduce a weight coefficient, λ, to balance the contributions of each loss function. In our experiments, we set λ to $1/7$ to ensure appropriate weight distribution.

To compute these loss functions, we employ weighted binary cross-entropy loss [71] (wbce) and intersection over union (IoU) loss [72]. The IoU loss ($\mathcal{L}_{\text{iou}}$) captures global structural information and provides global constraints to guide the network, while the wbce ($\mathcal{L}_{\text{wbce}}$) computes the pixel-wise loss and enforces pixel-level constraints. The expressions for the four loss functions are as follows:

$$\mathcal{L}_i = (\mathcal{L}_{\text{wbce}}(P_i, \text{GT}) + \mathcal{L}_{\text{iou}}(P_i, \text{GT})), \quad i = 1, 2, 3, 4 \tag{7}$$

where GT represents the ground truth.

IV. Experiments

A. Implementation Details

We followed [1] to train FSANet on DIS5K which was implemented using the PyTorch framework. We employed the Adam optimizer with an initial learning rate of 2×10^{-4} and adjusted the weight decay to 1×10^{-4} to update the network parameters. Additionally, we set the input image size to 736×736 during training, conducted 80 epochs, and used a batch size of 8. We resized the input image to 1024×1024 during testing.

B. Datasets

To evaluate the performance of FSANet, we conducted training on the DIS5K training dataset and carried out validation and testing on the DIS5K test and validation datasets. The DIS5K test datasets, namely DIS-TE1 to DIS-TE4, consist of objects with increasingly complex morphological structures. The DIS-VD dataset serves as the validation dataset, ensuring consistency in categories across all datasets.

To assess the generalizability of FSANet, we also performed retraining using a combined training dataset comprising MAS3K [81] (a marine organism segmentation dataset), and COD10K [14] (a camouflage object detection dataset). Subsequently, we validated the retrained model on the CHAMELEON [82], COD10K test datasets, and MAS3K test datasets.

C. Evaluation Metrics

We employed six evaluation metrics, including structural similarity measure (S_α) [83], weighted F-measure (F_β^ω) [84], mean absolute error (M) [85], mean enhanced alignment measure (mE_ϕ) [86], maximal F-measure (max F) [87], and human correction efforts (HCE) [1], to assess the model performance. Among them, S_α reflects the structural similarity between predicted and ground truth images, F_β^ω is the harmonic mean of average precision and average recall, eliminating the impact of equal consideration for each pixel, M measures the average difference between predicted and ground truth images, and mE_ϕ reflects the pixel-wise similarity between the predicted map and the true map. The HCE metric is a novel evaluation standard specifically introduced

TABLE II

WE PERFORMED A QUANTITATIVE EVALUATION ON THE DIS5K VALIDATION AND TEST SETS, AND THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY. ALL METHODS WERE TRAINED USING THE SAME DATASET. WE ASSESSED THE PERFORMANCE USING SIX METRICS: S_α, F_β^ω, M, mE_ϕ, max F, AND HCE. IN THE EVALUATION, ↑ INDICATES HIGHER SCORES ARE BETTER, WHILE ↓ INDICATES LOWER SCORES ARE BETTER

Datasets	Models	CPD	PraNet	F3Net	SINet-v2	MSNet	PFNet	PSGLoss	RankNet	Polyp-PVT	C2FNet	BSANet	BGNet	ISNet	BCMNet	HitNet	UDUN	FPDIS	Ours
		[73]	[74]	[75]	[14]	[76]	[13]	[77]	[78]	[7]	[33]	[79]	[35]	[1]	[80]	[34]	[24]	[25]	
	Publications	CVPR	MICCAI	AAAI	TPAMI	MICCAI	CVPR	TIP	CVPR	TMI	TCSVT	AAAI	IJCAI	ECCV	TCSVT	AAAI	ACM MM	IJCAI	
	Years	2019	2020	2020	2021	2021	2021	2021	2021	2021	2022	2022	2022	2022	2023	2023	2023	2023	
DIS-TE1	Sα ↑	0.623	0.761	0.764	0.778	0.741	0.778	0.742	0.774	0.798	0.775	0.792	0.801	0.774	0.780	0.816	0.816	0.821	0.821
	F_β^ω ↑	0.398	0.584	0.607	0.626	0.557	0.634	0.612	0.624	0.681	0.639	0.663	0.673	0.624	0.642	0.715	0.717	0.712	0.718
	M ↓	0.116	0.079	0.079	0.070	0.094	0.071	0.076	0.074	0.061	0.077	0.067	0.066	0.079	0.072	0.060	0.059	0.060	0.059
	mE_φ ↑	0.650	0.808	0.821	0.836	0.774	0.836	0.820	0.829	0.858	0.822	0.844	0.848	0.805	0.836	0.858	0.864	0.860	0.860
	$maxF$ ↑	0.504	0.688	0.708	0.717	0.647	0.720	0.686	0.720	0.737	0.713	0.736	0.749	0.721	0.714	0.768	0.784	0.783	0.777
	HCE ↓	209	259	242	272	236	259	250	260	261	199	251	231	157	255	193	141	160	135
DIS-TE2	Sα ↑	0.648	0.788	0.803	0.809	0.784	0.805	0.773	0.806	0.822	0.807	0.814	0.824	0.802	0.808	0.841	0.842	0.844	0.860
	F_β^ω ↑	0.468	0.643	0.682	0.690	0.650	0.694	0.685	0.691	0.740	0.703	0.711	0.725	0.690	0.704	0.774	0.768	0.767	0.800
	M ↓	0.129	0.089	0.077	0.076	0.090	0.074	0.076	0.076	0.063	0.075	0.070	0.068	0.081	0.073	0.057	0.058	0.060	0.052
	mE_φ ↑	0.678	0.832	0.859	0.869	0.830	0.868	0.858	0.866	0.889	0.857	0.871	0.875	0.835	0.868	0.894	0.885	0.892	0.905
	$maxF$ ↑	0.738	0.746	0.767	0.773	0.731	0.774	0.753	0.774	0.793	0.773	0.782	0.797	0.775	0.772	0.822	0.828	0.826	0.845
	HCE ↓	476	568	550	591	535	578	554	578	583	472	564	533	356	575	476	328	373	310
DIS-TE3	Sα ↑	0.648	0.799	0.814	0.815	0.802	0.820	0.781	0.824	0.828	0.828	0.830	0.843	0.824	0.823	0.857	0.865	0.871	0.873
	F_β^ω ↑	0.479	0.662	0.702	0.702	0.680	0.719	0.711	0.721	0.753	0.743	0.741	0.755	0.730	0.734	0.808	0.808	0.811	0.827
	M ↓	0.127	0.088	0.077	0.076	0.082	0.072	0.073	0.076	0.062	0.068	0.066	0.062	0.073	0.069	0.052	0.050	0.049	0.047
	mE_φ ↑	0.691	0.850	0.875	0.882	0.856	0.886	0.875	0.887	0.908	0.891	0.897	0.904	0.865	0.897	0.920	0.917	0.922	0.923
	$maxF$ ↑	0.591	0.765	0.788	0.788	0.765	0.798	0.780	0.804	0.809	0.813	0.811	0.830	0.814	0.802	0.852	0.864	0.867	0.870
	HCE ↓	914	1066	1058	1097	1052	1097	1051	1106	1100	948	1079	1051	686	1100	1006	659	780	640
DIS-TE4	Sα ↑	0.610	0.777	0.787	0.783	0.779	0.791	0.750	0.796	0.794	0.807	0.802	0.815	0.827	0.803	0.831	0.849	0.851	0.874
	F_β^ω ↑	0.446	0.635	0.668	0.659	0.656	0.680	0.680	0.687	0.708	0.717	0.707	0.716	0.735	0.706	0.772	0.791	0.788	0.831
	M ↓	0.155	0.106	0.096	0.097	0.099	0.092	0.089	0.090	0.083	0.082	0.083	0.081	0.079	0.084	0.065	0.059	0.062	0.050
	mE_φ ↑	0.658	0.826	0.848	0.848	0.833	0.858	0.850	0.860	0.882	0.867	0.869	0.874	0.859	0.876	0.905	0.900	0.906	0.926
	$maxF$ ↑	0.550	0.738	0.752	0.746	0.743	0.760	0.754	0.766	0.767	0.785	0.779	0.794	0.819	0.776	0.818	0.845	0.845	0.872
	HCE ↓	3468	3691	3730	3717	3795	3820	3765	3862	3718	3564	3794	3792	2762	3814	3788	2786	3347	2808
DIS-VD	Sα ↑	0.628	0.769	0.779	0.791	0.768	0.792	0.752	0.796	0.792	0.797	0.800	0.814	0.801	0.790	0.828	0.838	0.842	0.845
	F_β^ω ↑	0.440	0.615	0.646	0.658	0.627	0.672	0.657	0.687	0.692	0.689	0.688	0.709	0.685	0.678	0.757	0.763	0.763	0.776
	M ↓	0.141	0.098	0.090	0.084	0.099	0.082	0.085	0.090	0.078	0.081	0.077	0.072	0.080	0.082	0.061	0.060	0.063	0.057
	mE_φ ↑	0.663	0.825	0.834	0.850	0.814	0.856	0.847	0.860	0.867	0.849	0.861	0.873	0.837	0.859	0.890	0.891	0.891	0.894
	$maxF$ ↑	0.546	0.713	0.733	0.743	0.710	0.754	0.727	0.766	0.749	0.759	0.760	0.782	0.775	0.747	0.805	0.823	0.823	0.825
	HCE ↓	1454	1559	1569	1587	1574	1608	1548	1614	1581	1467	1566	1585	1072	1605	1550	1095	1310	1094
DIS-TE 1-4	Sα ↑	0.632	0.781	0.792	0.796	0.777	0.799	0.762	0.792	0.811	0.804	0.810	0.821	0.807	0.804	0.836	0.842	0.846	0.857
	F_β^ω ↑	0.448	0.631	0.665	0.669	0.636	0.682	0.672	0.666	0.721	0.701	0.706	0.717	0.695	0.697	0.767	0.770	0.768	0.794
	M ↓	0.132	0.091	0.082	0.080	0.091	0.077	0.079	0.083	0.067	0.076	0.072	0.069	0.078	0.075	0.059	0.057	0.059	0.052
	mE_φ ↑	0.669	0.829	0.851	0.859	0.823	0.862	0.851	0.849	0.884	0.859	0.870	0.875	0.841	0.869	0.894	0.891	0.894	0.903
	$maxF$ ↑	0.596	0.734	0.754	0.756	0.722	0.763	0.743	0.751	0.777	0.771	0.777	0.793	0.782	0.766	0.815	0.829	0.829	0.841
	HCE ↓	1267	1396	1395	1419	1405	1439	1405	1452	1416	1296	1422	1402	990	1436	1366	1001	1194	973

for high-precision DIS tasks. It measures the human effort required to correct misprediction results to meet specific accuracy requirements in real applications and is quantified by the number of mouse clicks, which, at the same time, is the metric we focus on the most.

D. Compared With SOTA Models

In this section, we present a comprehensive comparison between our proposed FSANet model and 17 SOTA models designed for various segmentation tasks, including medical image segmentation, SOD, camouflage object detection, and marine organism segmentation. To ensure a fair comparison, we retrained the comparison methods on the DIS5K training dataset using the default parameter settings recommended to evaluate the model in their respective original papers or use the results by these comparison methods directly to evaluate the model.

1) Quantitative Comparison: Using the evaluation criteria outlined in Section IV-C, we quantitatively evaluate and compare our proposed method with other segmentation methods using six widely used metrics. The results are presented in Table II. It can be observed that FSANet consistently outperforms the SOTA methods across all evaluation metrics on the compared datasets. Specifically, on the DIS-TE1 to DIS-TE4 test datasets, FSANet method achieves absolute improvements of 2.5% in S_α, 3.5% in F_β^ω, 11.8% in M, 1% in mE_ϕ, 3.2% in max F, and 1.7% in HCE in terms of average values compared to the current SOTA HitNet [34] or ISNet [1]. And it can be seen that as the difficulty of datasets DIS-TE1 to DIS-TE4 increases, FSANet method shows better performance than other models.

2) Qualitative Comparison: In Fig. 7, we showcase qualitative results to visually compare our proposed FSANet with 12 other recent SOTA segmentation methods. The visualizations demonstrate the consistent superiority of our FSANet in producing high-quality segmentation results. This can be attributed to the effectiveness of our high-frequency information extraction module and collaborative scale feature fusion module. These modules enable us to accurately capture the complete morphological structure

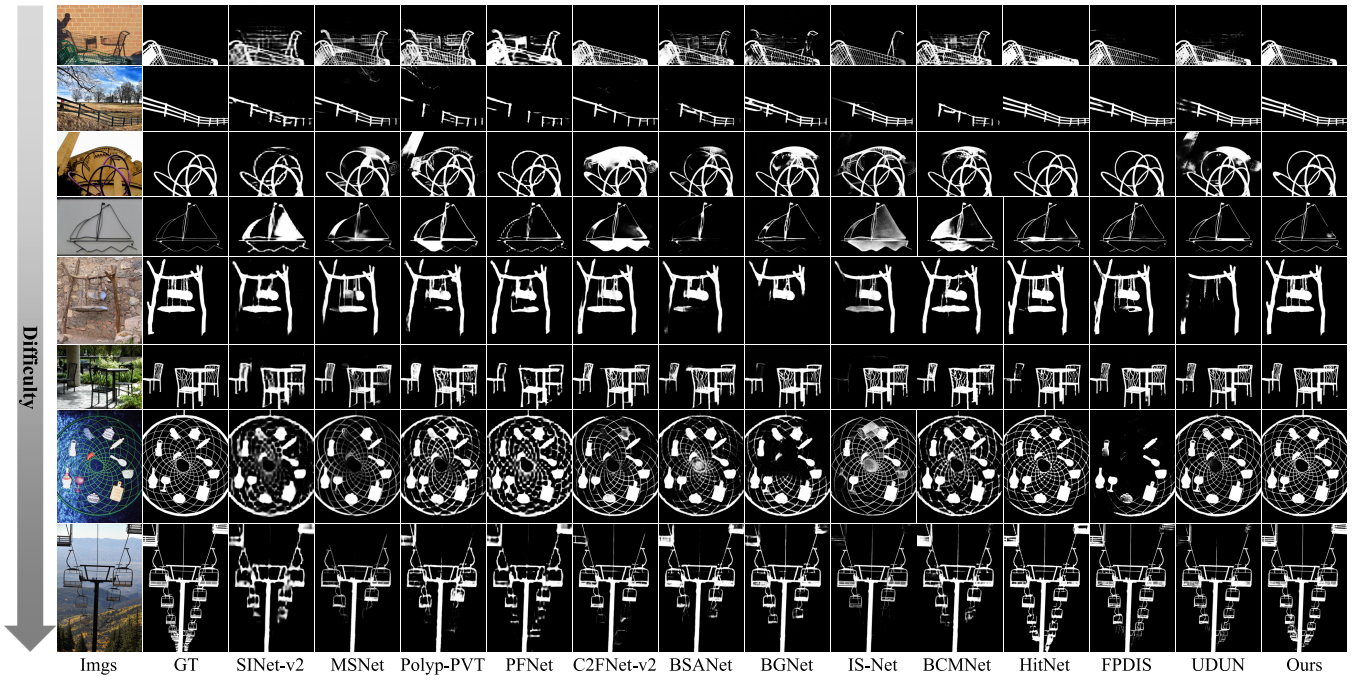

Fig. 7. Qualitative Comparison. The proposed model is qualitatively compared with 12 recent SOTA methods. Our FSANet method can consistently generate highly accurate binary segmentation maps that are very close to the true labels. Note that the presented examples are comprehensive, the structure of the object from top to bottom is increasingly complex, and the segmentation difficulty is gradually increasing. It can be found that in the case of relatively simple scenes, even if other models have segmented relatively complete targets, the segmentation results of FSANet are more delicate and clear. In challenging scenarios, our model still performs well in segmentation, such as the last row. Here, GT represents the ground truth. Please zoomed-in view for more details.

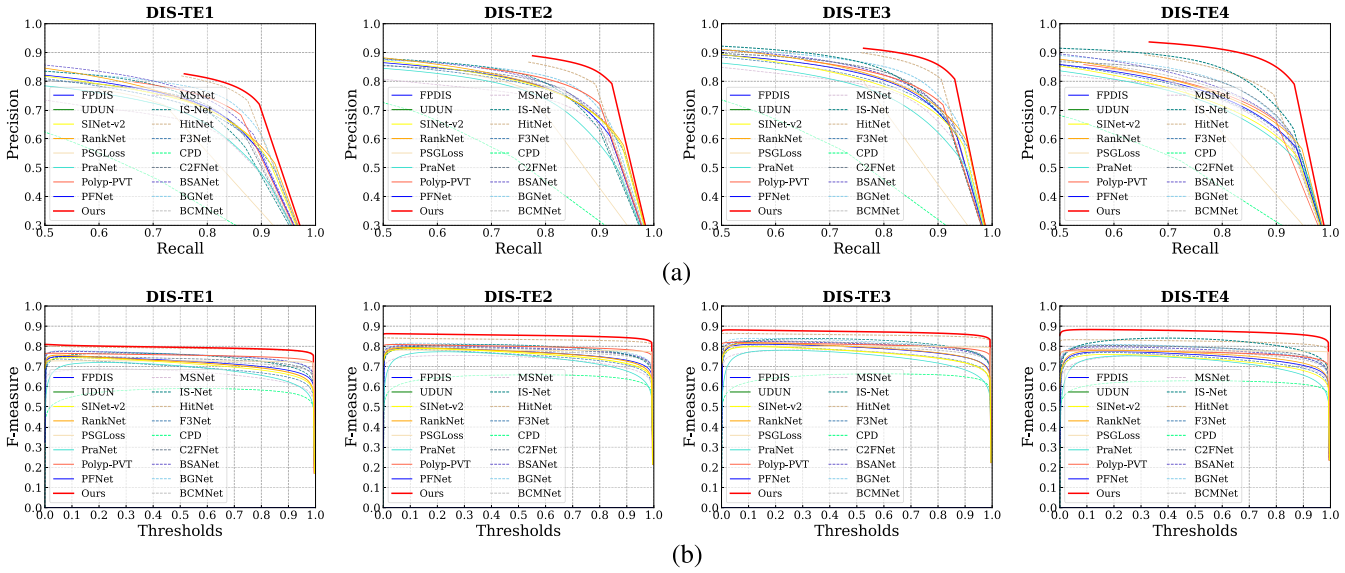

Fig. 8. (a) PR curve and (b) F-measure curve of the proposed model and the recent SOTA models are tested on four datasets with different difficulty levels, i.e., from DIS-TE1 to DIS-TE4. Our method (red) outperforms other SOTA counterparts by a large margin. It can be seen that as the difficulty of the dataset increases, our FSANet model shows better performance than other models.

of objects and generate refined and smooth boundary textures.

Specifically, taking the last two rows as examples, which are also the scenes from the most challenging subdataset, it is evident that other methods struggle to accurately locate and segment the target's content and morphological structure in complex scenes. In contrast, our FSANet performs exceptionally well, providing a closer approximation to the true label for both the primary (such as the foreground objects and the near cable cars) and secondary targets (such as the fine grid and farther cable cars).

3) PR/F-Measure Curve: In addition to the qualitative and quantitative comparisons, we provide precision–recall curves and F-measure curves to further evaluate the performance of all segmentation methods on different test datasets of DIS5K, as depicted in Fig. 8. Notably, the red solid line, representing FSANet, consistently outperforms the other segmentation models across most thresholds. This superior performance can

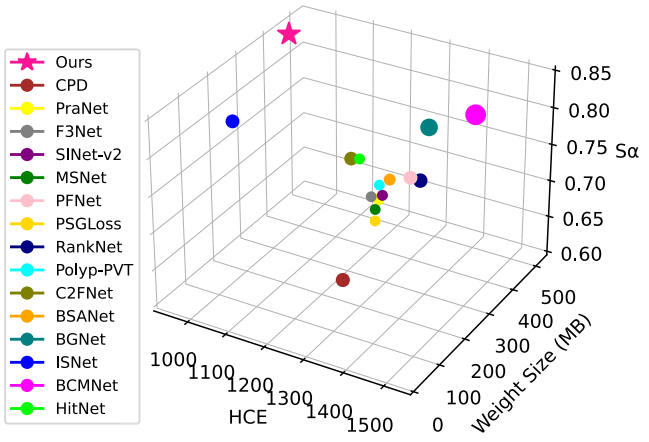

Fig. 9. Trade-off between performance and complexity. We compare FSANet model with 17 SOTA models in three metrics: model weight size, structural similarity S_α, and HCE. Among them, a smaller model weight and a smaller HCE, and a larger S_α indicate a better performance of the model. Our model sets new SOTA records with great reasonable complexity.

TABLE III

ABLATION STUDY OF KEY COMPONENTS. IN THIS ARTICLE, WE PERFORM DISSOCIATION EXPERIMENTS BY PROGRESSIVELY DECOUPLING OR REPLACING THE PROPOSED MODULES OR REPLACING THEM WITH SIMPLIFIED ALTERNATIVES. IN THE EVALUATION, ↓ INDICATES LOWER SCORES ARE BETTER

#	Ablation Models						TE1-TE4
	Frequency			CSFM	HPFM		$HCE\downarrow$
	Base	Sobel	LSF		HCF	SFF	
1	✔	-	✔	-	-	-	1405
2	✔	-	✔	✔	-	-	1301
3	✔	-	✔	✔	-	✔	1091
4	✔	-	✔	✔	✔	-	1092
5	✔	-	-	✔	✔	✔	1041
6	✔	✔	-	✔	✔	✔	1055
7	✔	-	✔	-	✔	✔	998
8	✔	-	✔	✔	✔	✔	**973**

be attributed to the incorporation of high-frequency features and the CSFM in our design, which contribute to generating clearer structure and content information. Consequently, our method achieves better precision–recall curves and F-measure curves. Furthermore, it is evident that the performance of our proposed method surpasses that of the other models as the dataset difficulty increases. This demonstrates the robustness and effectiveness of our approach in handling more challenging segmentation tasks.

4) Model Analysis: To provide a comprehensive evaluation of FSANet's performance, we conduct a detailed comparison with 17 SOTA models in terms of three metrics: model weight size, structural similarity (S_α), and HCE. The results are illustrated in Fig. 9. We can observe that our FSANet model outperforms the comparison models in terms of S_α and HCE metrics, even when the model weight size increase is not substantial. This indicates a significant improvement in performance achieved by our model.

V. ABLATION EXPERIMENTS AND ANALYSIS

To evaluate the effectiveness of the proposed key components, we conducted experiments in this section by decoupling different components of the FSANet model and examining their impact on performance.

A. Effectiveness of RGB-F Encoding

In this section, we conducted ablation studies to investigate the significance of high-frequency information in images, which aligns with one of our fundamental ideas. We remove the frequency awareness by directly feeding RGB features to the CSFM module after feature extraction. The experimental results are summarized in Table III. Comparing results #5 and #8, we observe a significant decrease in the HCE metric (approximately 7%) after removing the high-frequency information. This demonstrates the importance of incorporating high-frequency information for achieving higher segmentation performance. We also conducted experiments by replacing FSANet's high-pass filter with the Sobel filter, which also led to deteriorated performance #6.

We also present a visual comparison Fig. 10(a). It can be observed that the content integrity and morphological structure of the segmentation results show a noticeable decline after removing the auxiliary frequency domain information. These ablation studies highlight the crucial role played by the high-frequency information and the effectiveness of our proposed LSF in improving the performance of FSANet for high-precision DIS.

B. Effectiveness of Cross-Scale Interaction

In this section, we explore the effect of CSFM. Specifically, we remove the CSFM module and directly feed the BM features into the decoder. The quantitative results can be found in #7 of Table III. It can be observed that our full network achieves a notable reduction in manual correction costs compared to the single serial structure as in #7. To better understand the functionality of CSFM, we also provide a visual comparison shown in Fig. 10(b). After removing CSFM, the detection ability of the model for the target has significantly decreased, especially the high-precision perception ability for details, which verifies the effectiveness of CSFM in retaining high-precision local features and improving segmentation accuracy.

C. Effectiveness of Feature Decoding

We also conduct ablation studies to validate FSANet's decoding design, i.e., HPFM, which comprises two components: the HCF and the SFF. We verify their effectiveness by removing or replacing each module with basic operations.

We first investigate the impact of the whole decoder HPFM by replacing it with simple addition. The experimental results are shown in Table III. By comparing #2 and #8, it can be seen that removing the HPFM module leads to a significant performance decline compared to FSANet, with noticeable losses (about 34%) in content information and edge features. Additionally, the visual comparison in Fig. 11(a) supports these findings. These observations validate that our proposed decoding method can effectively perceive the content and details of the object, outperforming simple feature addition.

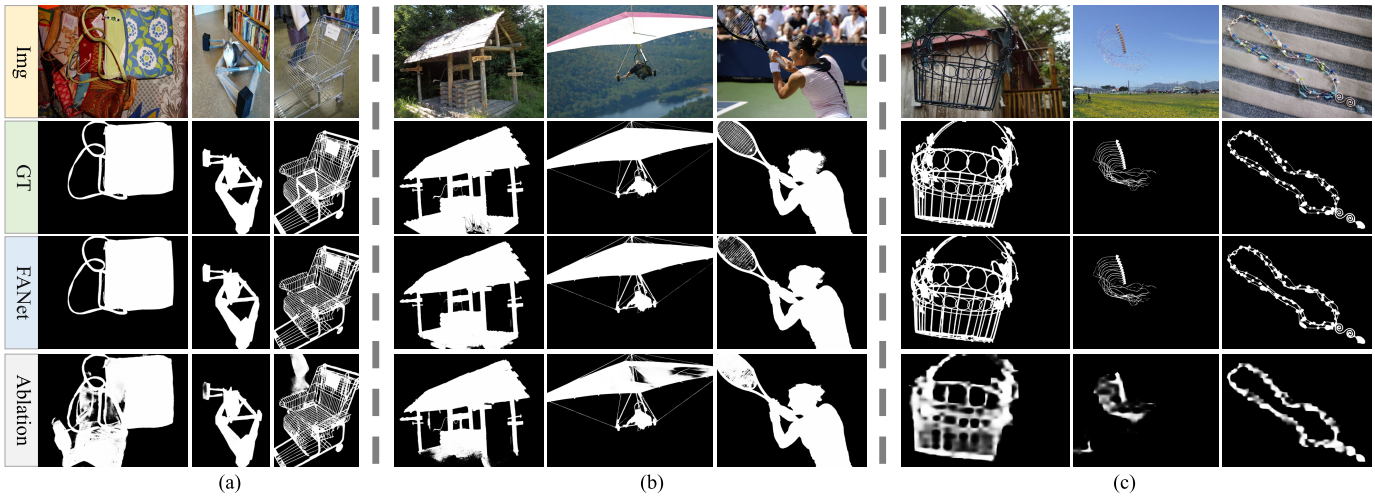

Fig. 10. Visualization of ablation experiment features. We showcase the output results obtained by removing the (a) LSF, (b) CSFM, and (c) both CSFM and HPFM modules. In this context, GT represents the ground truth. The results demonstrate that removing these modules leads to errors in segmenting target localization, structural deficiencies, and blurry targets, thereby confirming the effectiveness of each individual module. Here, w/o. stands for without.

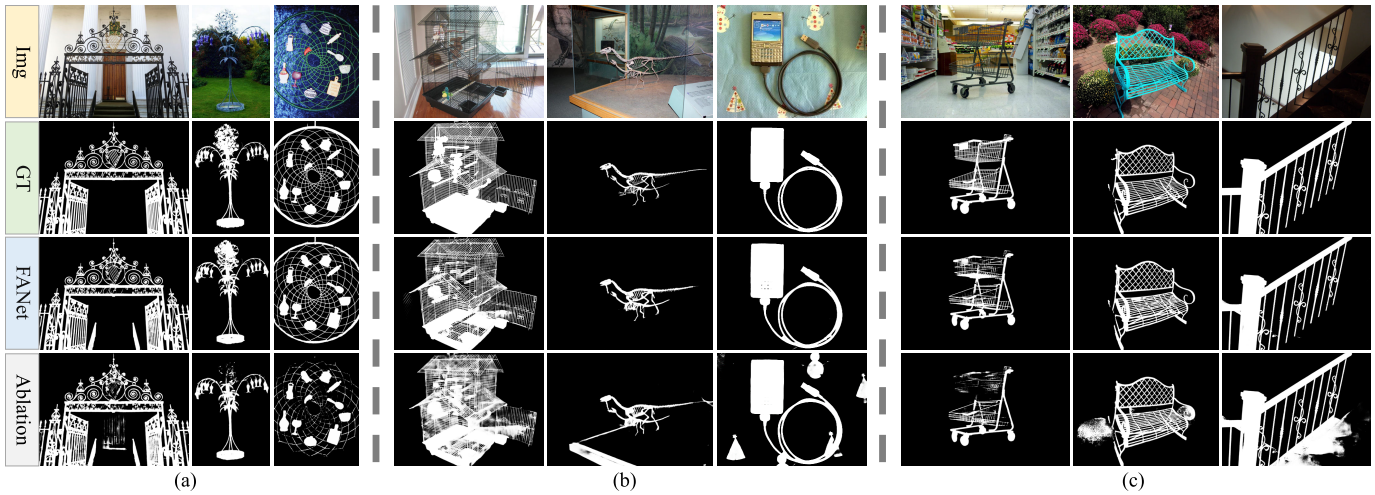

Fig. 11. Visualization of ablation experiment features. We show the output results of removing (a) HPFM, (b) HCF, and (c) SFF in the HPFM module. In this context, GT represents ground truth. It can be seen that removing these modules will lead to problems such as missing segmentation targets and redundant information, which verifies the effectiveness of each module. Here, w/o. stands for without.

To investigate the effectiveness of the HCF component, we replace it with 1×1 convolutions while keeping the other structures unchanged. The performance comparison is presented in Table III (#3 and #8). The degraded results suggest that the HCF plays a crucial role in reducing dimensionality and reserving important information during the segmentation process. Additionally, a visual comparison between the modified and original models is provided in Fig. 11(b). It is observed that directly using convolutions for dimensionality reduction results in content missing or noise in the network's segmentation outputs. This further supports the effectiveness of the proposed HCF.

Finally, we aim to validate the effectiveness of our SFF by removing all the semantic fusion and attention. The results are presented in Table III (#4 and #8). We can observe a significant drop in the HCE metric. Qualitatively, the removal of SFF results in blurry prediction outputs with noticeable noise, as depicted in Fig. 11(c). These results confirm the effectiveness of the proposed SFF, which plays a crucial role in achieving accurate and detailed segmentation results in the FSANet model.

VI. Other Applications

To evaluate the generalization ability of FSANet, we conducted further experiments by retraining it on a combined dataset comprising the COD10K camouflage object detection dataset and the MAS3K marine organism detection dataset. We also trained 14 other models on the same dataset and evaluated their performance on the CHAMELEON, COD10K, and MAS3K test datasets.

The quantitative comparisons of the models are presented in Table IV. The results demonstrate that FSANet achieves strong performance across different scenarios. It exhibits excellent adaptability and generalization capability when applied to diverse datasets, as evidenced by its competitive performance in comparison to the other models. These findings validate the effectiveness and versatility of our proposed model, indicating its potential for real-world applications in various domains,

TABLE IV

COMPARISON OF FSANET WITH SOTA MODELS ON THE MAS3K, CHAMELEON, AND COD10K TEST DATASETS. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE. THE PERFORMANCE WAS EVALUATED BY FOUR METRICS INCLUDING S_α, F_β^ω, M, AND mE_ϕ

Model	Years	Publication	MAS3K				CHAMELEON				COD10K			
			$S\alpha\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$mE_\varphi\uparrow$	$S\alpha\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$mE_\varphi\uparrow$	$S\alpha\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$mE_\varphi\uparrow$
CPD	2019	CVPR	0.869	0.776	0.032	0.897	0.870	0.758	0.034	0.886	0.774	0.588	0.041	0.801
SCRN	2019	ICCV	0.870	0.750	0.033	0.890	0.865	0.722	0.043	0.871	0.789	0.572	0.046	0.801
F3Net	2020	AAAI	0.872	0.801	0.028	0.927	0.868	0.764	0.038	0.917	0.805	0.650	0.039	0.872
PraNet	2020	MICCAI	0.883	0.817	0.026	0.929	0.873	0.785	0.034	0.922	0.812	0.671	0.036	0.877
Polyp-PVT	2021	TMI	0.889	0.840	0.027	0.934	0.881	0.830	0.026	0.943	0.814	0.705	0.035	0.887
PFSNet	2021	AAAI	0.880	0.816	0.027	0.929	0.866	0.781	0.033	0.924	0.806	0.669	0.038	0.876
PSGLoss	2021	TIP	0.848	0.779	0.031	0.883	0.828	0.753	0.031	0.872	0.732	0.566	0.040	0.732
SINet-v1	2020	CVPR	0.870	0.766	0.031	0.902	0.867	0.741	0.039	0.891	0.789	0.595	0.042	0.819
SINet-v2	2021	TPAMI	0.894	0.843	0.021	0.942	0.897	0.828	0.026	0.950	0.829	0.707	0.032	0.899
RankNet	2021	CVPR	0.858	0.764	0.034	0.909	0.838	0.707	0.048	0.884	0.782	0.599	0.049	0.841
BSANet	2022	AAAI	0.900	0.856	0.021	0.943	0.888	0.830	0.027	0.941	0.833	0.722	0.028	0.897
C2FNet	2022	TCSVT	0.898	0.852	0.022	0.939	0.891	0.839	0.026	0.942	0.827	0.715	0.031	0.896
ECDNet	2022	TCSVT	0.850	0.766	0.036	0.901	0.843	0.749	0.038	0.893	0.683	0.446	0.049	0.781
BCMNet	2023	TCSVT	0.906	0.865	0.019	0.945	0.900	0.863	0.022	0.954	0.829	0.723	0.027	0.899
FSANet	Ours		0.909	0.875	0.020	0.949	0.904	0.864	0.022	0.945	0.859	0.785	0.024	0.920

including camouflage object detection and marine organism detection.

VII. CONCLUSION

In this article, we address the challenge of achieving precise object segmentation by leveraging the rich structural details in high-frequency information. Departing from conventional approaches that rely solely on RGB visual features, we propose a fusion approach that integrates RGB spatial features with mimicked high-frequency information. This integration allows us to capture fine structural details and enhance the overall segmentation accuracy. Furthermore, we propose an improved U-shaped connection design that leverages, preserves, and merges the multiresolution clues for feature modeling and decoding, leading to more accurate reasoning on the object contour. Experimental results on the DIS benchmarks and other relative real-world applications demonstrate the effectiveness, robustness, and versatility in achieving precise and accurate segmentation results.

REFERENCES

[1] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. Van Gool, "Highly accurate dichotomous image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 38–56.

[2] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[3] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler, "EditGAN: High-precision semantic image editing," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 16331–16345.

[4] C. Anthes, R. J. García-Hernández, M. Wiedemann, and D. Kranzlmüller, "State of the art of virtual reality technology," in *Proc. IEEE Aerosp. Conf.*, Mar. 2016, pp. 1–19.

[5] X. Qin et al., "Boundary-aware segmentation network for mobile and web applications," 2021, *arXiv:2101.04704*.

[6] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[7] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," 2021, *arXiv:2108.06932*.

[8] S. Chen, X. Ma, Y. Lu, and D. Hsu, "Ab initio particle-based object manipulation," in *Proc. Robot., Sci. Syst.*, 2021.

[9] M. Jorda, E. G. Herrero, and O. Khatib, "Contact-driven posture behavior for safe and interactive robot operation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9243–9249.

[10] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2019.

[11] W. Zhang, L. Zheng, H. Wang, X. Wu, and X. Li, "Saliency hierarchy modeling via generative kernels for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 570–587.

[12] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7471–7481.

[13] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8768–8777.

[14] 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.

[15] X. Xie, G. Xie, and X. Xu, "High precision image segmentation algorithm using SLIC and neighborhood rough set," *Multimedia Tools Appl.*, vol. 77, no. 24, pp. 31525–31543, Dec. 2018.

[16] E. Shahrian, D. Rajan, B. Price, and S. Cohen, "Improving image matting using comprehensive sampling sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 636–643.

[17] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175–2188, Sep. 2013.

[18] K. He, J. Sun, and X. Tang, "Fast matting using large kernel matting Laplacian matrices," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2165–2172.

[19] J. Liu et al., "Boosting semantic human matting with coarse annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8560–8569.

[20] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 311–320.

[21] S. Cai et al., "Disentangled image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8818–8827.

[22] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Semantic image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11115–11124.

[23] L. Liu et al., "Instructive feature enhancement for dichotomous medical image segmentation," 2023, *arXiv:2306.03497*.

[24] J. Pei, Z. Zhou, Y. Jin, H. Tang, and P.-A. Heng, "Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 2139–2147.

[25] Y. Zhou, B. Dong, Y. Wu, W. Zhu, G. Chen, and Y. Zhang, "Dichotomous image segmentation with frequency priors," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, p. 3.

[26] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.

[27] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 225–241.

[28] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 275–292.

[29] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "HiDAnet: RGB-D salient object detection via hierarchical depth awareness," *IEEE Trans. Image Process.*, vol. 32, pp. 2160–2173, 2023.

[30] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, Dec. 2019.

[31] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-guided fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2949–2961, May 2022.

[32] Q. Guo, W. Zhou, J. Lei, and L. Yu, "TSFNet: Two-stage fusion network for RGB-T salient object detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1655–1659, 2021.

[33] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, and T. Zhou, "Camouflaged object detection via context-aware cross-level fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6981–6993, Oct. 2022.

[34] X. Hu et al., "High-resolution iterative feedback network for camouflaged object detection," 2022, *arXiv:2203.11624*.

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

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[37] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[38] C. Yu et al., "Lite-HRNet: A lightweight high-resolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10435–10445.

[39] S. Seong and J. Choi, "Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates," *Remote Sens.*, vol. 13, no. 16, p. 3087, Aug. 2021.

[40] H. Wu, C. Liang, M. Liu, and Z. Wen, "Optimized HRNet for image semantic segmentation," *Expert Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114532.

[41] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[42] 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.*, Jun. 2013, pp. 3166–3173.

[43] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.

[44] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2214–2219.

[45] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.

[46] Y. Zhang and L. Wu, "Optimal multi-level thresholding based on maximum Tsallis entropy via an artificial bee colony approach," *Entropy*, vol. 13, no. 4, pp. 841–859, Apr. 2011.

[47] L. Yuan and X. Xu, "Adaptive image edge detection algorithm based on Canny operator," in *Proc. 4th Int. Conf. Adv. Inf. Technol. Sensor Appl. (AITS)*, Aug. 2015, pp. 28–31.

[48] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

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

[50] Z. Deng et al., "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[51] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 6943–6950.

[52] Z. Wu et al., "Object segmentation by mining cross-modal semantics," 2023, *arXiv:2305.10469*.

[53] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1025–1031.

[54] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, "Deep gradient learning for efficient camouflaged object detection," *Mach. Intell. Res.*, vol. 20, no. 1, pp. 92–108, Jan. 2023.

[55] X.-J. Luo et al., "CamDiff: Camouflage image augmentation via diffusion model," 2023, *arXiv:2304.05469*.

[56] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[57] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[58] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[59] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[60] J. Wang, X. Long, G. Chen, Z. Wu, Z. Chen, and E. Ding, "U-HRNet: Delving into improving semantic representation of high resolution network for dense prediction," 2022, *arXiv:2210.07140*.

[61] L. Zhang and K. Yang, "Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 916–920, May 2014.

[62] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[63] C. He, Z. Chen, and C. Liu, "Salient object detection via images frequency domain analyzing," *Signal, Image Video Process.*, vol. 10, no. 7, pp. 1295–1302, Oct. 2016.

[64] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4494–4503.

[65] X. Jin, C. Guo, Z. He, J. Xu, Y. Wang, and Y. Su, "FCMNet: Frequency-aware cross-modality attention networks for RGB-D salient object detection," *Neurocomputing*, vol. 491, pp. 414–425, Jun. 2022.

[66] S. Zheng, Z. Wu, Y. Xu, Z. Wei, and A. Plaza, "Learning orientation information from frequency-domain for oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628512.

[67] 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, no. 2, pp. 1–16, Mar. 2023.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[69] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[70] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[71] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[72] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520.
[73] 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. (CVPR)*, Jun. 2019, pp. 3902–3911.
[74] D.-P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. MICCAI*, 2020, pp. 263–273.
[75] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 12321–12328.
[76] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," in *Proc. MICCAI*, 2021, pp. 120–130.
[77] S. Yang, W. Lin, G. Lin, Q. Jiang, and Z. Liu, "Progressive self-guided loss for salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 8426–8438, 2021.
[78] Y. Lv et al., "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11586–11596.
[79] H. Zhu et al., "I can find you! Boundary-guided separated attention network for camouflaged object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 3608–3613.
[80] J. Cheng, Z. Wu, S. Wang, C. Demonceaux, and Q. Jiang, "Bidirectional collaborative mentoring network for marine organism detection and beyond," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6595–6608, Nov. 2023.
[81] L. Li, B. Dong, E. Rigall, T. Zhou, J. Dong, and G. Chen, "Marine animal segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2303–2314, Apr. 2022.
[82] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, "Animal camouflage analysis: Chameleon database," *Unpublished Manuscript*, vol. 2, p. 7, Jan. 2018.
[83] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.
[84] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.
[85] 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.*, Jun. 2012, pp. 733–740.
[86] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.
[87] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

Qiuping Jiang (Senior Member, IEEE) is currently a Professor with the School of Information Science and Engineering, Ningbo University, Ningbo, China. His research interests include image quality assessment, visual perception modeling, and underwater visual information processing.

Dr. Jiang is an Associate Editor of *Displays*, *Journal of Visual Communication and Image Representation*, and *Journal of Electronic Imaging*.

Jinguang Cheng received the master's degree from Ningbo University, Ningbo, China, in 2024. He is currently pursuing the Ph.D. degree with Beijing University of Posts and Telecommunications, Beijing, China.

His research interests include image processing, especially in salient object detection and computational photography.

Zongwei Wu received the French Engineering degree (master's degree) in mechanical engineering from the University of Technologies of Compiègne (Grande Ecole)—Sorbonne University Alliance, Compiègne, France, in 2019 and the Ph.D. degree in computer vision from the University of Burgundy, Dijon, France, in 2022.

He is a Post-Doctoral Researcher at the University of Würzburg, Würzburg, Germany. His research interests include multimodal models and image processing.

Runmin Cong (Senior Member, IEEE) is currently a Professor with the School of Control Science and Engineering, Shandong University, Jinan, China. He has published more than 80 articles in prestigious international journals and conferences, including two ESI hot papers and 12 ESI highly cited articles. His research interests include computer vision, multimedia understanding, and content enhancement.

Mr. Cong serves as an Associate Editor for several high-level journals and as the area Chair/SPC/PC for multiple leading conferences.

Radu Timofte (Member, IEEE) received the Ph.D. degree in electrical engineering from KU Leuven, Leuven, Belgium, in 2013.

He was a Lecturer and the Group Leader at ETH Zürich, Zürich, Switzerland. His current research interests include deep learning, mobile AI, visual tracking, computational photography, image/video compression, restoration, enhancement, and manipulation.

Dr. Timofte is the 2022 Awardee of an Alexander von Humboldt Professorship for Artificial Intelligence. He holds the Chair for Computer Science IV (computer vision) at the University of Würzburg, Germany