Frequency-Spatial Entanglement Learning for Camouflaged Object Detection

Yanguang Sun[1], Chunyan Xu[1], Jian Yang[1], Hanyu Xuan[2] ✉, and Lei Luo[1] ✉

[1] PCA Lab, Nanjing University of Science and Technology, Nanjing, China
[2] School of Big Data and Statistics, Anhui University, Heifei, China
Sunyg@njust.edu.cn

Abstract. Camouflaged object detection has attracted a lot of attention in computer vision. The main challenge lies in the high degree of similarity between camouflaged objects and their surroundings in the spatial domain, making identification difficult. Existing methods attempt to reduce the impact of pixel similarity by maximizing the distinguishing ability of spatial features with complicated design, but often ignore the sensitivity and locality of features in the spatial domain, leading to suboptimal results. In this paper, we propose a new approach to address this issue by jointly exploring the representation in the frequency and spatial domains, introducing the Frequency-Spatial Entanglement Learning (FSEL) method. This method consists of a series of well-designed Entanglement Transformer Blocks (ETB) for representation learning, a Joint Domain Perception Module for semantic enhancement, and a Dual-domain Reverse Parser for feature integration in the frequency and spatial domains. Specifically, the ETB utilizes frequency self-attention to effectively characterize the relationship between different frequency bands, while the entanglement feed-forward network facilitates information interaction between features of different domains through entanglement learning. Our extensive experiments demonstrate the superiority of our FSEL over 21 state-of-the-art methods, through comprehensive quantitative and qualitative comparisons in three widely-used datasets. The source code is available at: https://github.com/CSYSI/FSEL.

Keywords: Camouflaged object detection · Computer vision · Frequency-Spatial entanglement learning

1 Introduction

"Camouflage" is a natural defense mechanism used by certain animals, such as chameleons, grasshoppers, and caterpillars, to blend into their surroundings and protect themselves. The study of camouflaged object detection (COD) focuses on identifying concealed targets in real-world situations. This research is crucial in developing robust visual perception models in computer vision. COD has a wide range of applications, including medical image analysis [8, 9], species conservation [33], and industrial defect detection [10].

✉ Corresponding Author

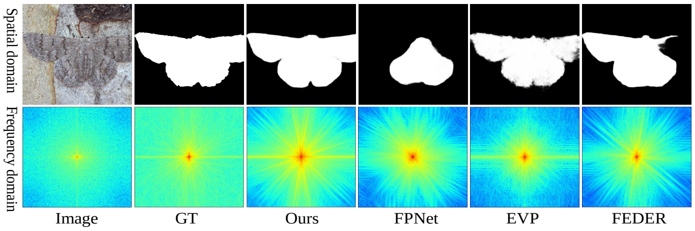

Fig. 1: The visual comparison results of the proposed FSEL and current COD methods (*i.e.*, FPNet [4], EVP [27], and FEDER [13]) in the spatial and frequency domain.

In the early stages, some COD methods [11,32,55] relied on manually crafted features to detect camouflaged objects. However, due to the extremely challenging appearance of these objects, the results obtained were often unsatisfactory. Later, with the advancement of deep learning and the availability of large-scale datasets [7, 20], many COD methods [18, 21, 30, 34, 52] based on deep learning have been proposed. With a large amount of data available for training, these methods have the potential to automatically extract features and detect camouflaged objects, resulting in impressive performance. More recently, researchers have introduced various techniques for exploring useful information of input features from the spatial domain using boundary-guided [39, 57, 58], multi-scale strategy [6, 34, 38], uncertainty-aware [21, 47], distraction mining [31], and *etc.*

We have observed that the COD methods [7, 18, 21, 52, 58] mentioned above primarily focus on single spatial features. While these spatial features are advantageous for COD tasks, they are often susceptible to interference from complex backgrounds. This vulnerability arises from their reliance on pixel-level information, with a primary emphasis on the local intensity and spatial position of individual pixels. Furthermore, spatial features possess local properties, meaning that pixels within a feature may only exhibit certain correlations with surrounding pixels. That being said, relying solely on spatial features can make it challenging to distinguish subtle variations within concealed objects and backgrounds. Therefore, it is crucial to find ways to overcome the limitations of spatial features to achieve accurate COD results. Recently, frequency features generated through Fourier Transform have been shown to have global characteristics and have been proven to be beneficial for understanding image contents [3, 35, 42]. This can help break the bottleneck of spatial features.

Some recent COD methods [4, 13, 22, 27, 56] have begun to incorporate frequency clues in their approach. These methods can be divided into two categories based on their objectives. The first category (*e.g.*, FDNet [56] and EVP [27]) is to act directly on input images through different frequency transforms to extract frequency features, which are then combined with spatial features. However, camouflaged images often contain a lot of background noise, making the frequency features obtained from the image unreliable. When aggregated with spatial features, this may introduce some unnecessary background noises, resulting in under-segmented results (as depicted EVP [27] in Fig. 1). The second category [4, 13, 22] focuses on initial features from the encoder. For example, Cong *et al.* [4] designed a frequency-perception module to improve the detection of camouflaged objects by utilizing both high-frequency features and low-frequency features. And He *et al.* [13] proposed frequency attention modules that obtain important parts of corresponding features by considering both high-frequency and low-frequency components. Although these methods have shown promising results, they only focus on high-frequency and low-frequency features, overlooking some information that falls between these two frequencies. This can be seen in FPNet [4] and FEDER [13] in Fig. 1, where significant information within the frequency domain may be missed.

Based on the above discussion, we propose a novel method called Frequency-Spatial Entanglement Learning for accurate camouflaged object detection. Our method combines global frequency features and local spatial features to optimize the initial input features and enhance their discriminative ability. Specifically, we first establish a Frequency Self-attention to obtain discriminative global frequency features, which models the correlation between each frequency band and learns the dependency relationships between different frequencies in input bands. Moreover, we introduce entanglement learning between the frequency and spatial features in the Entanglement Transformer Block, allowing them to mutually learn and collaborate for optimization. Furthermore, we extend the applicability of global frequency features by utilizing the Joint Domain Perception Module and the Dual-domain Reverse Parser to optimize the input features and generate powerful representations that incorporate both frequency and spatial information. Extensive experiments on three widely-used benchmark datasets (*i.e.*, CAMO [20], COD10K [7], and NC4K [30]) demonstrate that FSEL consistently outperforms 21 state-of-the-art COD methods across different backbones.

The main contributions can be summarized as follows:

(1) We propose a Frequency-Spatial Entanglement Learning (FSEL) framework that utilizes both global frequency and local spatial features to enhance the detection of camouflaged objects.

(2) To improve the representation capability of frequency and spatial features, we have designed an Entanglement Transformer Block (ETB). This block allows for entanglement learning of frequency-spatial features, resulting in a more comprehensive understanding of the data.

(3) To reduce the sensitivity and locality limitation of spatial features, we have incorporated frequency domain transformations into both the Joint Domain Perception Module (JDPM) and the Dual-domain Reverse Parser (DRP).

2 Related work

Camouflaged Object Detection. Recently, with the public availability of datasets (*i.e.*, CAMO [20], COD10K [7], and NC4K [30]), deep learning-based COD methods have started to surface in large numbers, which can be broadly categorized into, including multi-scale strategies [34, 38, 59], edge-guidance [13, 39, 57, 58], uncertainty-aware [21, 47], multi-graph learning [52], iterative man-

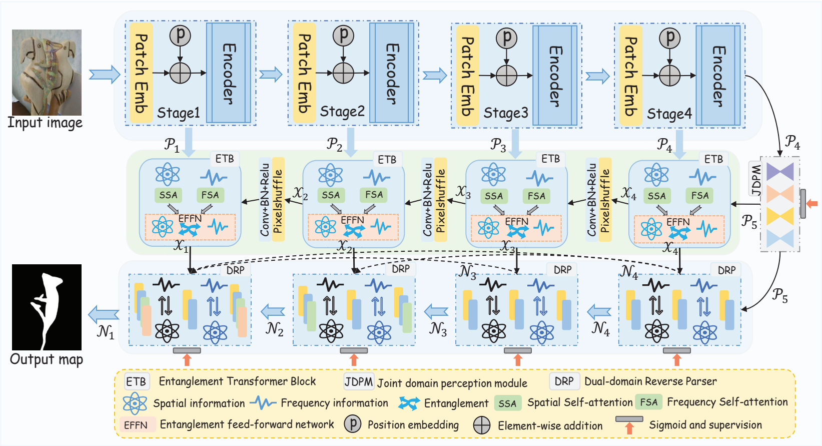

Fig. 2: Overview of the proposed FSEL model framework for camouflaged object detection. The proposed FSEL method generates predicted results through a Joint Domain Perception Module (JDPM), a series of stacked Entanglement Transformer Block (ETB), and a Dual-domain Reverse Parser (DRP).

ner [15, 18, 54], transformer [27, 37], and so on. Besides, other COD methods [4, 22, 27] considered frequency clues to help with reasoning camouflaged objects. Particularly, Zhong *et al.* [56] processed directly the camouflaged image through discrete cosine transform to obtain frequency information. He *et al.* [13] proposed frequency attention modules to filter out the noteworthy parts of corresponding features. After that, Cong *et al.* [4] designed a frequency-perception module by learning different frequency features to achieve coarse localization of camouflaged objects. However, these methods often focus on the high- and low-frequency information, ignoring the relationship between all bands in the frequency domain. Therefore, we conduct frequency analysis on different spectral features to achieve all frequency band interactions and importance allocation. Furthermore, we perform entanglement learning on global frequency and local spatial features, which is beneficial for obtaining powerful representations.

Vision Transformer. Transformer [40] utilized self-attention to model global semantic and long-range dependencies, which helps to understand the correlation between different regions in an image, and therefore it has been widely used in some computer vision tasks, including object detection [1, 45], image classification [43, 60], semantic segmentation [5, 17, 46], *etc.* For example, Yuan *et al.* [49] obtained long-range relationships from the sequence of image patches to perform the image classification. Next, Liu *et al.* [29] split input maps into non-overlapping local windows, and then transferred the information through shift operations between the windows to improve the efficiency of the model. In addition, other transformer models have been successful in computer visions, such as Restormer [51], CrossFormer [44], EfficientViT [28], MPFormer [53], and among others. Unlike these methods, which always model relationships based on the spatial domain, we transform spatial features into the frequency domain and combine them to perform dual-domain feature optimization.

Frequency Learning. The frequency domain is very important for signal analysis, and recently it has been gradually applied in computer vision tasks. Particularly, Qin *et al.* [35] assumed channel attention as a compression problem and introduced frequency transformation in the channel attention. Yun *et al.* [50] handled the balancing problem of different frequency components of visual features. Wang *et al.* [42] proposed a frequency shortcut perspective in image classification. In addition, some frequency domain-based methods [3,4,13,23,27,41] have achieved great performance. In this paper, we extend global frequency features to different applications, involving multi-receptive fields perception, transformer, and reverse attention.

3 Method

3.1 Framework Architecture

Camouflaged objects exhibit a high level of visual similarity to their backgrounds, achieved through adaptive changes in color, texture, and shape. This creates challenges in distinguishing between object and background pixels in the spatial domain. Additionally, the locality of features in the spatial domain is limited in understanding camouflaged objects. To address this issue, we have implemented several strategies: 1) We have expanded beyond the spatial domain and utilized Fourier transformation to map features to the frequency domain, allowing for a more global perspective; 2) We have analyzed the relationships between all frequency bands to combine global frequency features with local spatial features; 3) We have extended frequency features to multiple components to fully utilize the global understanding of the object.

The complete architecture of our FSEL model is shown in Fig. 2. Given an input image $I_c \in \mathbb{R}^{H\times W\times 3}$, we first use the basic encoder (*i.e.*, PVTv2 [43]/ ResNet [14]/ Res2Net [12]) to extract initial feature $\mathcal{P}{=}\{\mathcal{P}_i\}_{i=1}^4$ with the resolution of $\frac{W}{2^{i+1}}\times\frac{H}{2^{i+1}}$. Then the JDPM (Sec. 3.2) captures a higher-level semantic feature $\mathcal{P}_5$ to guide location by integrating multi-receptive field information from frequency-spatial domains. After that, the ETB (Sec. 3.3) models the cross-long-range relationships and performs entanglement learning on the frequency and spatial domains from initial features to generate discriminative feature $\mathcal{X}{=}\{\mathcal{X}_i\}_{i=1}^4$. To ensure the quality of predicted map $\mathcal{N}{=}\{\mathcal{N}_i\}_{i=1}^4$, we design a DRP (Sec. 3.4) to aggregate feature flows through auxiliary optimization in the frequency and spatial domains.

3.2 Joint Domain Perception Module

Multi-scale information is beneficial for contextual understanding in different regions. We observe that these methods [2,26,48] often generate multi-scale features through different convolutions with multiple receptive fields in the spatial domain. However, the receptive field of convolution operations in the spatial

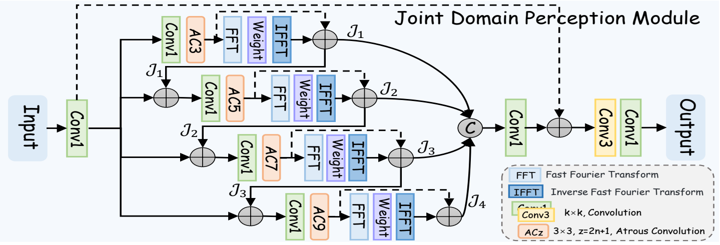

Fig. 3: Details of the joint domain perception module.

domain is limited, and in the process of data processing, tiny fluctuations may be overlooked, resulting in sub-optimized outcomes.

Therefore, we propose a Joint Domain Perception Module (JDPM) that reconstructs multi-receptive field information by introducing frequency transformation in multi-scale features. As depicted in Fig. 3, our JDRM uses the hierarchical structure to extract frequency-spatial information of different receptive fields. Technically, we use feature $\mathcal{P}_4$ as input and first reduce its channel numbers using 1×1 convolution (C_1), *i.e.*, $\mathcal{P}_4^{128} = C_1\mathcal{P}_4$. Then, we construct a set of 3×3 atrous convolutions ($\mathcal{AC}_z$) with filling rate z to capture local multi-scale spatial feature $\{\mathcal{J}_n^s\}_{n=1}^4$, *i.e.*, $\mathcal{J}_n^s = C_1\mathcal{AC}_z(\mathcal{P}_4^{128} + \mathcal{J}_{n-1})$, where $z = 2n+1$ and $n-1 \geq 1$. Next, we transform local spatial features into the frequency domain using the Fast Fourier Transform ($fft(\cdot)$) and perform redundancy filtering. We then use the Inverse Fast Fourier Transform ($ifft(\cdot)$) and the modulus of complex features to obtain global frequency features $\left\{\mathcal{J}_n^f\right\}_{n=1}^4$, which is defined as:

$$\begin{aligned}
\mathcal{J}_n^f &= \varPhi\left\|ifft(\sigma(fft(\mathcal{J}_n^s)) * fft(\mathcal{J}_n^s))\right\|, n = 1, 2, 3, 4, \\
fft(u, v) &= \sum_{x=0}^{W-1}\sum_{y=0}^{H-1}\mathcal{J}_n^s(x, y)e^{-2\pi i(\frac{ux}{W}+\frac{vy}{H})}, \\
ifft(x, y) &= \frac{1}{WH}\sum_{u=0}^{W-1}\sum_{v=0}^{H-1}fft(u, v)e^{2\pi i(\frac{ux}{W}+\frac{vy}{H})},
\end{aligned} \tag{1}$$

where $\varPhi\left\|\cdot\right\|$ and "$*$" denote the modulus operation and the element-wise multiplication. $\sigma(\cdot)$ presents a set of weight coefficients, which sequentially contains a convolution, a batch normalization, a ReLU, a convolution, and a sigmoid function. (u, v) and (x, y) denote frequency domain coordinates and spatial domain coordinates, i represents the imaginary part. After that, we aggregate global frequency features with local spatial features to generate intermediate multi-scale features $\{\mathcal{J}_n\}_{n=1}^4$, that is, $\mathcal{J}_n = J_n^s + \mathcal{J}_n^f, n = 1, 2, 3, 4$.

Finally, we concatenate all multi-scale features and introduce residual connections to generate a coarse feature map $\mathcal{P}_5$ with 1-channel through a 3 × 3 and a 1 × 1 convolutions, which can be expressed as:

$$\mathcal{P}_5 = C_3C_1(C_1Cat(\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3, \mathcal{J}_4) + \mathcal{P}_4^{128}), \tag{2}$$

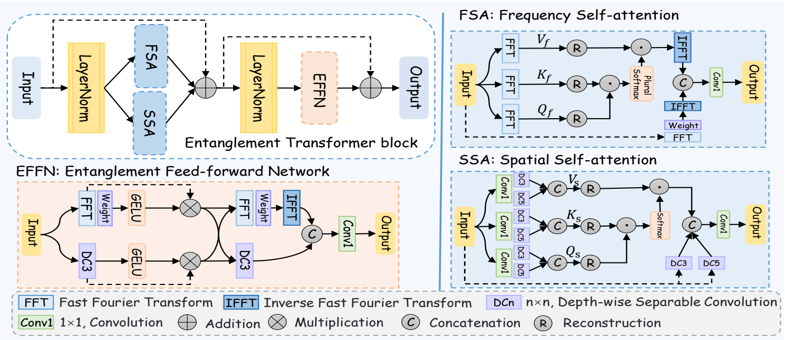

Fig. 4: Details of the entanglement transformer block.

where C_k presents $k \times k$ convolution. $Cat(\cdot,\cdot,\cdot,\cdot)$ and "+" denote concatenation and element-wise addition.

3.3 Entanglement Transformer Block

Unlike previous methods [4,16,29], which only model long-range dependencies based on local features in the spatial domain, our ETB incorporates different relationships from the frequency and spatial domains. In addition, we propose entanglement learning for different domain features in the ETB, allowing for the integration of information such as color, texture, edge, spectral, amplitude, and energy. This approach is beneficial for learning discriminative representations by considering various types of information. As depicted in Fig. 4, our ETB consists of three key components: frequency self-attention (FSA), spatial self-attention (SSA), and entanglement feed-forward network (EFFN).

Frequency self-attention. To better analyze frequency signals, we introduce a self-attention structure, which allows the model to obtain the relationships and interactions among different frequency bands, learn importance weights, and perform adaptive fusion. Technically, the feature $\mathcal{P}_{\psi}^{128}$ with 128-channel is taken as an input, which comes from the initial features at the same level and the optimized high-level features (as depicted in Fig. 2), and a layer normalization to generate feature $\hat{\mathcal{P}}(\hat{\mathcal{P}}{=}\mathcal{LN}(\mathcal{P}_{\psi}^{128}))$. After that, we use a Fast Fourier Transform to obtain *query* $Q_f{=}fft^Q(\hat{\mathcal{P}})$, *key* $K_f{=}fft^K(\hat{\mathcal{P}})$, *value* $V_f{=}fft^V(\hat{\mathcal{P}})$ in the frequency domain, where $fft^{(\cdot)}(\cdot)$ denotes the Fast Fourier Transform, and then we reshape *query* ($\widetilde{Q}_f{\in}\mathbb{R}^{C\times HW}$) and *key* ($\widetilde{K}_f{\in}\mathbb{R}^{HW\times C}$) projections that their dot-product generates transpose-attention map ($\mathit{\Lambda}_f$, $\mathit{\Lambda}_f{=}\widetilde{Q}_f \odot \widetilde{K}_f$), where $\odot$ presents the matrix multiplication. Different from these models [15,16] that directly utilize the Softmax function to activate the attention map with the real type, the frequency attention map ($\mathit{\Lambda}_f$) is a complex type that cannot be activated directly. Therefore, we extract the real part ($\mathit{\Lambda}_f^{re}$, $\mathit{\Lambda}_f^{re}{=}(\mathit{\Lambda}_f + conj(\mathit{\Lambda}_f))/2$) and imaginary part ($\mathit{\Lambda}_f^{im}$, $\mathit{\Lambda}_f^{im}{=}(\mathit{\Lambda}_f - conj(\mathit{\Lambda}_f))/2i$), where $conj(\cdot)$ denotes conjugate complex number, and then activate and merge the real and imaginary parts to obtain the activated attention maps ($a\mathit{\Lambda}_f$), as follows:

$$a\mathit{\Lambda}_f = \mathit{\Theta}(Sof(\mathit{\Lambda}_f^{re}), Sof(\mathit{\Lambda}_f^{im})), \tag{3}$$

where $\mathit{\Theta}(\cdot,\cdot)$ presents a combination function that combines the imaginary and real parts into a complex number. $Sof(\cdot)$ denotes a Softmax function. Subsequently, we use the attention map $a\mathit{\Lambda}_f$ to optimize the weights on the frequency feature V_f and then use the Inverse Fast Fourier Transform ($ifft(\cdot)$) to convert it to an original domain and employ the modulus operation to obtain the frequency attention feature. In addition, we introduce a frequency residual connection to increase frequency information ($\hat{\mathcal{P}}_f^r$, $\hat{\mathcal{P}}_f^r$=$\mathit{\Phi}||ifft(\sigma(fft(\hat{\mathcal{P}})) * fft(\hat{\mathcal{P}}))||$), and finally fuse features to produce the frequency feature $\mathcal{X}_f^1$, which is formulated as:

$$\mathcal{X}_f^1 = C_1 Cat(\mathit{\Phi}||ifft(a\mathit{\Lambda}_f \odot \widetilde{V}_f)||, \hat{\mathcal{P}}_f^r), \tag{4}$$

where $Cat(\cdot,\cdot)$ and $\odot$ are concatenation and matrix multiplication. $\mathit{\Phi}\,\|\cdot\|$ presents the modulus operation. $\widetilde{V}_f$ is the reshaped V_f.

Spatial self-attention. Considering the unfixed size of camouflaged objects, we embed abundant contextual information into spatial self-attention. As shown in the bottom right of Fig. 4, similar to the FSA, we take the feature $\hat{\mathcal{P}}$ as the input and encode the position information using a 1×1 convolution (C_1), and then we obtain the *query* Q_s, *key* K_s, and *value* V_s required by the self-attention by utilizing two depth-wise separable convolution with 3×3 ($\mathcal{DC}_3$) and 5×5 ($\mathcal{DC}_5$). After that, we generate the attention map ($a\mathit{\Lambda}_s$, $a\mathit{\Lambda}_s$= $Sof(\widetilde{Q}_s \odot \widetilde{K}_s)$) through the reconstructed $\widetilde{Q}_s$ and $\widetilde{K}_s$ and activate it using Softmax function. Subsequently, the activated attention map $a\mathit{\Lambda}_s$ is used to correct the weights of V_s. Besides, to increase the spatial local information ($\hat{\mathcal{P}}_s^r$, $\hat{\mathcal{P}}_s^r$=$Cat(\mathcal{DC}_3\hat{\mathcal{P}}, \mathcal{DC}_5\hat{\mathcal{P}})$), we perform a residual connection to generate spatial feature $\mathcal{X}_s^r$, as shown in:

$$\mathcal{X}_s^r = C_1 Cat(a\mathit{\Lambda}_s \odot \widetilde{V}_s, \hat{\mathcal{P}}_s^r), \tag{5}$$

where C_1, $Cat(\cdot,\cdot)$, and $\odot$ are the same as in Eq. (4). $\widetilde{V}_s$ is the reshaped V_s.

Entanglement feed-forward network. Frequency and spatial features usually contain different information. The frequency domain focuses on the global energy distribution and variation of signals, while spatial information acts on local pixel-level details and spatial structures, all of which are crucial for comprehending camouflaged objects. In our EFFN, these features are considered as two kinds of states that can perform entanglement learning to obtain more robust and powerful representations during the entanglement process.

Specifically, we first entangle the global frequency feature $\mathcal{X}_f^1$ and the local spatial feature $\mathcal{X}_s^1$ to adapt them to each other, followed by the residual connection to acquire the comprehensive feature $\mathcal{X}_c^1$, that is, $\mathcal{X}_c^1$=$\mathcal{X}_s^1$+$\mathcal{X}_f^1$+$\mathcal{P}_\psi^{128}$, which performs the layer normalization to improve the stability, and then the normalized feature $\hat{\mathcal{X}}_c^1$ ($\hat{\mathcal{X}}_c^1$=$\mathcal{LN}(\mathcal{X}_c^1)$) is subjected to non-linearity entanglement learning in the EFFN. Technically, the EFFN consists of two phases, the first

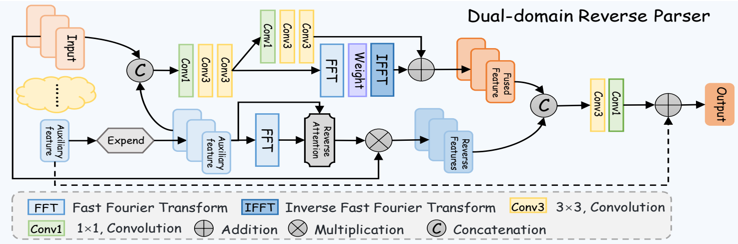

Fig. 5: Details of the dual-domain reverse parser.

stage projects the feature $\hat{\mathcal{X}}_c^1$ to the frequency and spatial domains, and utilizes the GELU function for nonlinear activation and a gate mechanism to obtain global frequency feature $\hat{\mathcal{X}}_f^2$ and local spatial feature $\hat{\mathcal{X}}_s^2$, which can be written as follows:

$$\begin{aligned}
&\hat{\mathcal{X}}_f^2 = GE(\Phi||\sigma(fft(\hat{\mathcal{X}}_c^1)) * fft(\hat{\mathcal{X}}_c^1)||) * \Phi||\sigma(fft(\hat{\mathcal{X}}_c^1)) * fft(\hat{\mathcal{X}}_c^1)||,\\
&\hat{\mathcal{X}}_s^2 = GE(\mathcal{DC}_3\hat{\mathcal{X}}_c^1) * \mathcal{DC}_3\hat{\mathcal{X}}_c^1,
\end{aligned} \tag{6}$$

where $GE(\cdot)$ denotes the GELU function. Subsequently, in the second stage, the frequency and spatial features from the first stage are again entangled by interacting with each other by transferring information from different domains, and the entangled frequency-spatial features are optimized independently. They are then aggregated and reduced channels to generate comprehensive feature $\hat{\mathcal{X}}_c^3$, which can be formulated as follows:

$$\begin{aligned}
&\hat{\mathcal{X}}_c^3 = C_1Cat(\hat{\mathcal{X}}_f^3, \hat{\mathcal{X}}_s^3) + \mathcal{X}_c^1,\\
&\hat{\mathcal{X}}_f^3 = \Phi||ifft(\sigma(fft(Cat(\hat{\mathcal{X}}_f^2, \hat{\mathcal{X}}_s^2))) * fft(Cat(\hat{\mathcal{X}}_f^2, \hat{\mathcal{X}}_s^2)))||,\\
&\hat{\mathcal{X}}_s^3 = \mathcal{DC}_3Cat(\hat{\mathcal{X}}_f^2, \hat{\mathcal{X}}_s^2),
\end{aligned} \tag{7}$$

where C_1, $Cat(\cdot,\cdot)$ and $\Phi||\cdot||$ are the same as in Eq. (4). $fft(\cdot)$ and $ifft(\cdot)$ present the Fast Fourier Transform and the Inverse Fast Fourier Transform. $\mathcal{DC}_3$ denotes the depth-wise separable convolution with 3×3 kernel. Finally, we introduce residual connections to obtain the final feature $\mathcal{X}$ with 128-channel in the ETB, *i.e.*, $\mathcal{X}$=$C_1Cat(\hat{\mathcal{X}}_c^3, \mathcal{P}_\psi^{128}) + \mathcal{P}_\psi^{128}$. Through multiple aggregation interactions, global frequency and local spatial features interact and depend on each other, leading to the entanglement of features from different states, forming rich and comprehensive representations.

3.4 Dual-domain Reverse Parser

Different from these methods [6, 16, 58] that integrate multi-level features based on the spatial domain, we propose the dual-domain reverse parser (DRP), which optimizes and aggregates diverse information from multi-level feature $\mathcal{X}$ in both frequency and spatial domains. As depicted in Fig. 2, we first take the feature $\mathcal{X}_4$ from the ETB as the optimization objective and use the higher-level semantic feature $\mathcal{P}_5$ as the auxiliary objective.

The DRP consists of two branches (as shown in Fig. 5), in the first branch, we first expand the channel of auxiliary feature $\mathcal{P}_5$ to match the dimension of the optimization objective and aggregate these feature to obtain feature $\mathcal{I}_4$, *i.e.*, $\mathcal{I}_4$=$\mathcal{C}on(Cat(Ex(\mathcal{P}_5), \mathcal{X}_4))$, where $Ex(\cdot)$ denotes to expand the channel to 128, $\mathcal{C}on(\cdot)$ presents a 1×1 convolution and two 3×3 convolutions. And then feature $\mathcal{I}_4$ is separated into the spatial and frequency domains. We perform the Fast Fourier Transform $(fft(\cdot))$ and Inverse Fast Fourier Transform $(ifft(\cdot))$ in the frequency domain and adopt a series of convolution operations $(\mathcal{C}on(\cdot))$ to optimize features in the spatial domain. Subsequently, they are aggregated to obtain the fused feature $\mathcal{N}_4^1$, that is,

$$\mathcal{N}_4^1 = \Phi||ifft(\sigma(fft(\mathcal{I}_4)) * fft(\mathcal{I}_4))|| + \mathcal{C}on(\mathcal{I}_4), \tag{8}$$

where $\Phi \, \|\cdot\|$ denotes the modulus operation. "+" presents element-wise addition. In the second branch, we produce the hybrid reverse attention map ($\mathcal{A}_r$, $\mathcal{A}_r$=(1− $Sig(\mathcal{P}_5)) + (1 - Sig(\Phi||fft(\mathcal{P}_5)||)))$) using the auxiliary feature, where $Sig(\cdot)$ denotes the Sigmoid function. Unlike other methods [6,13], our the reverse attention map $(\mathcal{A}_r)$ contains abundant the frequency-spatial information to efficiently obtain the reverse feature $\mathcal{N}_4^2$, *i.e.*, $\mathcal{N}_4^2$=$\mathcal{A}_r * \mathcal{X}_4$. Next, we integrate the features $\mathcal{N}_4^1$ and $\mathcal{N}_4^2$ to generate final feature $\mathcal{N}_4$, that is, $\mathcal{N}_4$=$C_3Cat(\mathcal{N}_4^1,\mathcal{N}_4^2)$+$\mathcal{P}_5$. Subsequently, $\mathcal{N}_{i+1}$ will continue to optimize features $\mathcal{X}_i (i = 1, 2, 3)$ as an auxiliary objective in the proposed DRP. Note that there must be at least one auxiliary feature used for optimizing feature $\mathcal{X}_i$ to generate feature $\mathcal{N}_i$, and auxiliary features are input in the dense connection manner.

3.5 Loss function

In the proposed FSEL method, we supervise multi-level feature $\mathcal{N}$ to produce an accurately predicted map. Specifically, we adopt the weighted binary cross-entropy (BCE) and the weighted intersection over union (IoU) [36] as the overall loss function to optimize the model based on ground truth (G). The loss function can be defined as:

$$\mathcal{L}_{all} = \sum_{i=1}^{5} \frac{1}{2^{i-1}} (\mathcal{L}_{bce}^{w}(\mathcal{N}_i, G) + \mathcal{L}_{iou}^{w}(\mathcal{N}_i, G)), \tag{9}$$

where $\mathcal{L}_{bce}^{w}$ and $\mathcal{L}_{iou}^{w}$ denote the weighted BCE and IoU functions. $\mathcal{N}_5$ is the feature $\mathcal{P}_5$ from the JDPM.

4 Experiment

4.1 Experimental Setups

Datasets. We evaluate our FSEL model on three benchmark datasets: CAMO [20], COD10K [7], and NC4K [30]. CAMO [20] is an early dataset that contains

Table 1: Quantitative results on three COD datasets. The best result is shown in **blod**. “Ours-R50”, “Ours-R2N”, and “Ours-Pvt” present ResNet50 [14]/Res2Net [12]/PVTv2 [43] as backbone. “Ours-R50†”denotes using the same input strategy as ZoomNet [34].

Method	Year, Pub.	CAMO (250 images)						COD10K (2026 images)						NC4K(4121 images)					
		$\mathcal{M}\downarrow$	$F^m_\varphi\uparrow$	$F^a_\varphi\uparrow$	$F^w_\varphi\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$	$F^m_\varphi\uparrow$	$F^a_\varphi\uparrow$	$F^w_\varphi\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$	$F^m_\varphi\uparrow$	$F^a_\varphi\uparrow$	$F^w_\varphi\uparrow$	$S_m\uparrow$	$E_m\uparrow$
SINet [7]	2020, CVPR	0.100	0.762	0.709	0.606	0.751	0.835	0.051	0.708	0.593	0.551	0.770	0.797	0.058	0.805	0.768	0.723	0.807	0.883
UGTR [47]	2021, ICCV	0.086	0.800	0.748	0.684	0.784	0.858	0.036	0.772	0.671	0.666	0.815	0.850	0.052	0.833	0.778	0.747	0.839	0.888
JSOCOD [21]	2021, CVPR	0.073	0.812	0.779	0.728	0.800	0.872	0.035	0.762	0.705	0.684	0.807	0.882	0.047	0.838	0.803	0.771	0.841	0.906
MGL-S [52]	2021, CVPR	0.089	0.791	0.733	0.664	0.772	0.850	0.037	0.765	0.667	0.655	0.808	0.851	0.055	0.826	0.771	0.731	0.828	0.885
LSR [30]	2021, CVPR	0.080	0.791	0.756	0.696	0.787	0.859	0.037	0.756	0.699	0.673	0.802	0.883	0.048	0.836	0.802	0.766	0.839	0.904
PFNet [31]	2021, CVPR	0.085	0.795	0.751	0.695	0.782	0.855	0.040	0.748	0.676	0.660	0.798	0.868	0.053	0.821	0.779	0.745	0.828	0.894
SegMaR$_1$ [18]	2022, CVPR	0.072	0.821	0.772	0.728	0.808	0.870	0.035	0.765	0.699	0.682	0.811	0.881	—	—	—	—	—	—
PreyNet [54]	2022, MM	0.077	0.803	0.764	0.708	0.789	0.856	0.034	0.775	0.731	0.697	0.810	0.894	—	—	—	—	—	—
FEDER [13]	2023, CVPR	0.071	0.824	0.786	0.738	0.802	0.877	0.032	0.788	0.740	0.716	0.820	**0.901**	0.044	0.852	**0.822**	0.789	0.846	0.913
Ours-R50	—	**0.067**	**0.833**	**0.799**	**0.758**	**0.821**	**0.893**	**0.031**	**0.802**	**0.743**	**0.728**	**0.830**	0.898	**0.042**	**0.855**	0.818	**0.792**	**0.854**	**0.914**
ZoomNet [34]	2022, CVPR	**0.066**	0.832	0.792	0.752	0.820	0.883	**0.029**	0.810	0.741	0.729	0.835	0.893	0.043	0.851	0.815	0.784	0.852	0.907
Ours-R50†	-	0.068	**0.837**	**0.799**	**0.765**	**0.826**	**0.890**	**0.029**	**0.814**	**0.754**	**0.743**	**0.839**	**0.903**	**0.040**	**0.863**	**0.828**	**0.802**	**0.861**	**0.917**
C2FNet [38]	2021, IJCAI	0.080	0.803	0.764	0.719	0.796	0.865	0.036	0.764	0.703	0.686	0.811	0.886	0.049	0.832	0.788	0.762	0.838	0.901
FAPNet [57]	2022, TIP	0.076	0.823	0.776	0.734	0.815	0.877	0.036	0.781	0.707	0.694	0.820	0.875	0.047	0.846	0.804	0.775	0.850	0.903
SINet$_{v2}$ [6]	2022, TPAMI	0.071	0.820	0.779	0.743	0.820	0.884	0.037	0.770	0.682	0.680	0.813	0.864	0.048	0.842	0.792	0.770	0.847	0.901
BSANet [58]	2022, AAAI	0.079	0.804	0.768	0.717	0.794	0.866	0.034	0.776	0.724	0.699	0.815	0.894	0.048	0.839	0.805	0.771	0.841	0.906
BGNet [39]	2022, IJCAI	0.073	0.825	0.786	0.749	0.811	0.878	0.033	0.795	**0.739**	0.722	0.828	**0.902**	0.044	0.851	0.813	0.788	0.850	0.911
Ours-R2N	—	**0.065**	**0.844**	**0.803**	**0.771**	**0.831**	**0.895**	**0.030**	**0.803**	**0.739**	**0.731**	**0.837**	0.899	**0.042**	**0.858**	**0.821**	**0.798**	**0.860**	**0.915**
VST [25]	2021, ICCV	0.081	0.812	0.753	0.713	0.808	0.853	0.037	0.779	0.721	0.698	0.817	0.882	0.048	0.840	0.801	0.768	0.844	0.899
EVP [27]	2023, CVPR	0.067	0.836	0.800	0.762	0.831	0.896	0.032	0.802	0.708	0.726	0.835	0.877	—	—	—	—	—	—
FPNet [4]	2023, MM	0.056	0.863	0.838	0.802	0.851	0.912	0.029	0.817	0.765	0.755	0.847	0.909	—	—	—	—	—	—
HiNet [15]	2023, AAAI	0.055	0.857	0.833	0.809	0.849	0.910	0.023	0.850	**0.818**	**0.806**	0.868	**0.936**	0.037	0.879	0.854	0.834	0.874	0.928
FSPNet [16]	2023, CVPR	0.050	0.869	0.829	0.799	0.855	0.919	0.026	0.816	0.736	0.735	0.847	0.900	0.035	0.878	0.826	0.816	0.878	0.923
SAM [19]	2023, ICCV	—	—	—	—	—	—	0.050	0.844	0.758	0.701	0.778	0.800	0.078	0.852	0.754	0.696	0.765	0.778
Ours-Pvt	—	**0.040**	**0.891**	**0.864**	**0.851**	**0.885**	**0.942**	**0.021**	**0.853**	0.796	0.800	**0.873**	0.928	**0.030**	**0.895**	**0.864**	**0.853**	**0.892**	**0.941**

Table 2: Efficiency analysis of our FSEL and multiple COD methods.

	SINet [7]	PFNet [31]	MGL [52]	UGTR [47]	JSOCOD [21]	C2FNet [38]	ZoomNet [34]	SegMaR [18]	FSPNet [16]	HitNet [15]	FEDER [13]	Ours-R50	Ours-R2N	Ours-Pvt
Parameters (M)	48.95	46.50	63.60	48.87	217.98	28.41	32.38	55.62	273.79	**25.73**	37.37	29.15	29.31	67.13
FLOPs (G)	38.75	53.22	553.94	1000.01	112.34	26.17	203.50	33.65	283.31	56.55	**23.98**	35.64	37.07	54.73

1,250 camouflaged images with 1,000 training images and 250 testing images. COD10K [7] is a currently large dataset of camouflaged objects, consisting of 3,040 training images and 2,026 testing images. NC4K [30] is the largest COD dataset for testing, containing 4,121 images of camouflaged objects. We use 4,040 images from CAMO [20] and COD10K [7] as training samples to train the FSEL.

Implementation details. The proposed FSEL model is implemented in the PyTorch framework on four NVIDIA GTX 4090 GPUs with 24GB. We utilize the pre-trained PVTv2 [43]/ResNet50 [14]/Res2Net [12] as the encoder to extract initial features. Following [6, 13], we also employ data augmentation techniques such as random flipping and random clipping to enhance training data. We use the Adam optimizer with an initial learning rate of 1e-4 and decay the rates by 10 every 60 epochs. All input images are resized to 416×416, and the batch size is set to 40 for 180 epochs of training progressing.

Evaluation metrics. We use six well-known evaluation metrics, including Mean Absolute Error ($\mathcal{M}$), Maximum F-measure (F^m_φ), Average F-measure (F^a_φ), Weighted F-measure (F^w_φ), S-measure (S_m), and E-measure (E_m).

4.2 Comparisons with the SOTAs

We conduct a comparison of our FSEL with twenty-one COD methods, including SINet [7], C2FNet [38], UGTR [47], JSOCOD [21], MGL-S [52], LSR [30], PFNet [31], VST [25], FAPNet [57], SINet$_{v2}$ [6], BSANet [58], SegMaR [18],

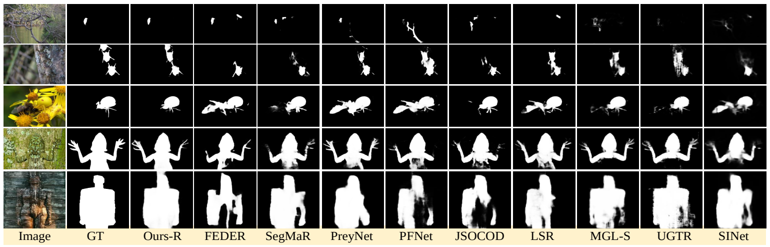

Fig. 6: Qualitative comparisons of the proposed FSEL and nine COD methods.

ZoomNet [34], BGNet [39], PreyNet [54], FEDER [13], EVP [27], FPNet [4], HitNet [15], FSPNet [16], and SAM [19]. Note that the predicted maps from all methods are provided by the authors or obtained from open-source codes.

Quantitative Evaluation. Table 1 summarizes the quantitative result of our FSEL and other 21 SOTA models. From Table 1, we can observe that the FSEL model achieves excellent performance across different backbone networks. Particularly, compared to the recently proposed FEDER [13] method, our FSEL with ResNet50 [14] backbone overall surpasses 5.97%, 3.23%, and 4.76% on three public datasets under the $\mathcal{M}$ metric. Besides, in the Res2Net [12] backbone, our FSEL achieves average performance gains of 9.23%, 2.30%, 2.16%, 2.94%, 1.34%, and 1.24% over the second-best method in terms of six public evaluation metrics on CAMO [20] dataset. Moreover, compared to the frequency-based FPNet [4] and EVP [27] methods, FSEL method with PVTv2 [43] backbone achieves average performance gains of 38.10%, 4.41%, 4.05%, 5.96%, 3.07%, and 2.09% over FPNet [4] and 52.38%, 6.36%, 12.43%, 10.19%, 4.55%, and 5.82% over EVP [27] in terms of $\mathcal{M}$, F_{φ}^{m}, F_{φ}^{a}, F_{φ}^{w}, S_m, and E_m on the COD10K [7] dataset. Furthermore, FSEL achieves excellent performance when adopting the same input strategy with ZoomNet [34]. The superiority in performance benefits from the joint optimization of the ETB, JDPM, and DRP for input features in the frequency and spatial domains. In addition, we provide the parameters and FLOPs in Table 2. It can be seen that the proposed FSEL method parameters and FLOPs are at a medium to high level, however, our performance far exceeds that of methods with similar parameters and FLOPs.

Qualitative Evalation. Fig. 6 gives the visual comparisons between our FSEL and several COD method in different scenarios. As depicted in Fig. 6, the proposed FSEL method exhibits accurate and complete segmentation for camouflaged objects with different sizes compared to current COD methods (*i.e.*, HitNet [15], FSPNet [16], and FPNet [4]). These visual results demonstrate the superiority of the FSEL method for detecting camouflaged objects through the frequency-spatial domain optimization strategy.

4.3 Ablation Study

Effectiveness of proposed each component. We provide the quantitative results of different components in the proposed FSEL model, shown in Table 3.

Table 3: Ablation analysis of our FSEL structure.

No.	Structure Setting				CAMO(250 images)						COD10K(2026 images)					
	Baseline	ETB	DRP	JDPM	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$F_\varphi^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$F_\varphi^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$
(a)	✓				0.093	0.784	0.712	0.663	0.767	0.847	0.046	0.749	0.617	0.610	0.778	0.818
(b)	✓	✓			0.076	0.811	0.763	0.723	0.801	0.873	0.034	0.789	0.712	0.702	0.821	0.886
(c)	✓		✓		0.074	0.820	0.778	0.735	0.810	0.876	0.034	0.794	0.722	0.713	0.826	0.887
(d)	✓			✓	0.081	0.797	0.743	0.697	0.787	0.863	0.039	0.769	0.676	0.668	0.804	0.863
(e)	✓	✓	✓		0.074	0.823	0.782	0.731	0.807	0.875	0.032	0.798	0.723	0.714	0.828	0.888
(f)	✓	✓		✓	0.071	0.807	0.767	0.728	0.802	0.878	0.033	0.781	0.719	0.702	0.815	0.895
(g)	✓		✓	✓	0.071	0.830	0.789	0.742	0.810	0.879	**0.031**	0.796	0.730	0.718	0.827	0.893
(h)	✓	✓	✓	✓	**0.067**	**0.833**	**0.799**	**0.758**	**0.821**	**0.893**	**0.031**	**0.802**	**0.743**	**0.728**	**0.830**	**0.898**

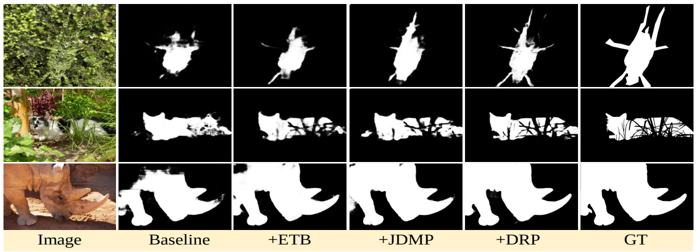

Fig. 7: Visual results of the effectiveness of our modules.

Specifically, we first adopt "ResNet50 [14] - FPN [24]" as "Baseline" (Tab. 3(a)) to detect camouflaged objects. And then we independently validate the effectiveness of "ETB" (Tab. 3(b)), "DRP" (Tab. 3(c)) and "JDPM" (Tab. 3(d)), and it can be seen that the performance of the predicted map increases significantly when the proposed component is embedded in the "Baseline" (Tab. 3(a)). Additionally, we validate the compatibility among all modules. From Tab. 3(e), Tab. 3(f), and Tab. 3(g), it can be observed that the three components are compatible with each other. Subsequently, all components are integrated, and the performance of the model is improved once again, as shown in Tab. 3(h). Additionally, in Fig. 7, we show the visual results obtained by progressively adding the proposed components (*i.e.*, ETB, JDPM, and DRP), generating that the predicted map gradually approaches the ground truth (GT). The above results demonstrate the effectiveness of our proposed modules in detecting camouflaged objects.

Effectiveness of frequency-spatial information within the ETB. Do we really need frequency information? To answer this question, we perform a series of experiments in the internal part of the ETB. Specifically, the ETB is first divided into two parts, with "ETB-S" (Tab. 4(a)) containing only spatial information, and "ETB-F" (Tab. 4(b)) presenting that it includes only frequency information. Based on Table 4, the performance of the separate frequency and spatial domains exhibits certain differences compared to the complete ETB (Tab. 4(g)). Besides, we investigate the entanglement learning of frequency-spatial information in the proposed ETB. In Tab. 4(c)-(f), it can be seen that the frequency and spatial features interact fusion to achieve entanglement between two states, enhancing the model's reasoning ability of camouflaged objects.

Table 4: Ablation analysis within the ETB structure.

No.	ETB				CAMO(250 images)						COD10K(2026 images)					
	FSA	SFA	FFFN	SFFA	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$F_\varphi^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$F_\varphi^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$
(a)		✓		✓	0.079	0.798	0.752	0.706	**0.802**	0.864	0.037	0.770	0.697	0.679	0.816	0.874
(b)	✓		✓		**0.075**	0.810	**0.770**	**0.725**	0.798	**0.880**	**0.034**	0.778	**0.713**	0.695	0.812	**0.890**
(c)	✓		✓	✓	**0.075**	0.807	0.759	0.720	0.795	0.876	0.035	0.778	0.692	0.687	0.811	0.873
(d)		✓	✓	✓	0.078	**0.811**	0.764	0.723	0.800	0.875	**0.034**	0.777	0.703	0.692	0.814	0.882
(e)	✓	✓	✓		0.076	0.808	0.764	0.712	0.795	0.872	**0.034**	0.782	0.702	0.692	0.819	0.878
(f)	✓	✓		✓	0.077	0.801	0.757	0.712	0.794	0.873	0.036	0.778	0.697	0.686	0.814	0.875
(g)	✓	✓	✓	✓	0.076	**0.811**	0.763	0.723	0.801	0.873	**0.034**	**0.789**	0.712	**0.702**	**0.821**	0.886

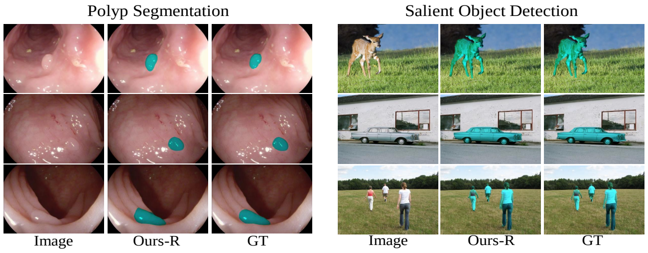

Fig. 8: Visual results of the expanded application.

4.4 Expanded application

To demonstrate the generalization ability of our FSEL model, we extend the FSEL model to salient object detection and polyp segmentation tasks. As shown in Fig. 8, the proposed FSEL method achieves highly accurate segmentation for both salient objects and polyps, benefiting from the complementary utilization of frequency domain and spatial information. More details and data are presented in the **supplementary materials**.

5 Conclusion

In this paper, we introduce a new approach for detecting camouflaged objects called Frequency-Spatial Entanglement Learning (FSEL). The key to FSEL is to extract important information from both the frequency and spatial domains. To achieve this, we have developed a Joint Domain Perception Module that combines multi-scale information from frequency-spatial features to accurately localize regions. Additionally, we have created an Entanglement Transformer Block that can be easily integrated into existing methods to improve their performance by modeling long-range dependencies in the hybrid domain. Furthermore, we have designed a Dual-Domain Reverse Parser that interacts with diverse information in multi-layer features to achieve more precise segmentation. Our extensive comparison experiments demonstrate that FSEL outperforms 21 state-of-the-art COD methods on three popular benchmark datasets.

Acknowledgments. This work was supported in part by the National Science Fund of China (No. 62276135, 62361166670, 62372238, and 62302006).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI **40**(4), 834–848 (2017)
3. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS. vol. 33, pp. 4479–4488 (2020)
4. Cong, R., Sun, M., Zhang, S., Zhou, X., Zhang, W., Zhao, Y.: Frequency perception network for camouflaged object detection. In: ACM MM (2023)
5. Ding, J., Xue, N., Xia, G.S., Schiele, B., Dai, D.: Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In: CVPR. pp. 15413–15423 (2023)
6. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. TPAMI **44**(10), 6024–6042 (2022)
7. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR. pp. 2777–2787 (2020)
8. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263–273 (2020)
9. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. TMI **39**(8), 2626–2637 (2020)
10. Fang, F., Li, L., Gu, Y., Zhu, H., Lim, J.H.: A novel hybrid approach for crack detection. PR **107**, 107474 (2020)
11. Galun, Sharon, Basri, Brandt: Texture segmentation by multiscale aggregation of filter responses and shape elements. In: ICCV. pp. 716–723 (2003)
12. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. TPAMI **43**(2), 652–662 (2019)
13. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: CVPR. pp. 22046–22055 (2023)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
15. Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., Shao, L.: High-resolution iterative feedback network for camouflaged object detection. In: AAAI. vol. 37, pp. 881–889 (2023)
16. Huang, Z., Dai, H., Xiang, T.Z., Wang, S., Chen, H.X., Qin, J., Xiong, H.: Feature shrinkage pyramid for camouflaged object detection with transformers. In: CVPR. pp. 5557–5566 (2023)
17. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: CVPR. pp. 2989–2998 (2023)
18. Jia, Q., Yao, S., Liu, Y., Fan, X., Liu, R., Luo, Z.: Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In: CVPR. pp. 4713–4722 (2022)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: ICCV. pp. 4015–4026 (2023)

20. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Computer Vision and Image Understanding **184**, 45–56 (2019)
21. Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., Dai, Y.: Uncertainty-aware joint salient object and camouflaged object detection. In: CVPR. pp. 10071–10081 (2021)
22. Lin, J., Tan, X., Xu, K., Ma, L., Lau, R.W.: Frequency-aware camouflaged object detection. ACM TMCCA **19**(2), 1–16 (2023)
23. Lin, S., Zhang, Z., Huang, Z., Lu, Y., Lan, C., Chu, P., You, Q., Wang, J., Liu, Z., Parulkar, A., et al.: Deep frequency filtering for domain generalization. In: CVPR. pp. 11797–11807 (2023)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: ICCV. pp. 2117–2125 (2017)
25. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: ICCV. pp. 4722–4732 (2021)
26. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: ECCV. pp. 385–400 (2018)
27. Liu, W., Shen, X., Pun, C.M., Cun, X.: Explicit visual prompting for low-level structure segmentations. In: CVPR. pp. 19434–19445 (2023)
28. Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y.: Efficientvit: Memory efficient vision transformer with cascaded group attention. In: CVPR. pp. 14420–14430 (2023)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
30. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR. pp. 11591–11601 (2021)
31. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: CVPR. pp. 8772–8781 (2021)
32. Mondal, A., Ghosh, S., Ghosh, A.: Partially camouflaged object tracking using modified probabilistic neural network and fuzzy energy based active contour. IJCV **122**, 116–148 (2017)
33. Nafus, M.G., Germano, J.M., Perry, J.A., Todd, B.D., Walsh, A., Swaisgood, R.R.: Hiding in plain sight: a study on camouflage and habitat selection in a slow-moving desert herbivore. BE **26**(5), 1389–1394 (2015)
34. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: CVPR. pp. 2160–2170 (2022)
35. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: ICCV. pp. 783–792 (2021)
36. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation pp. 234–244 (2016)
37. Song, Z., Kang, X., Wei, X., Liu, H., Dian, R., Li, S.: Fsnet: Focus scanning network for camouflaged object detection. TIP (2023)
38. Sun, Y., Chen, G., Zhou, T., Zhang, Y., Liu, N.: Context-aware cross-level fusion network for camouflaged object detection. In: IJCAI. pp. 1025–1031 (2021)
39. Sun, Y., Wang, S., Chen, C., Xiang, T.Z.: Boundary-guided camouflaged object detection. In: IJCAI (2022)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
41. Wang, C., Jiang, J., Zhong, Z., Liu, X.: Spatial-frequency mutual learning for face super-resolution. In: CVPR. pp. 22356–22366 (2023)

42. Wang, S., Veldhuis, R., Brune, C., Strisciuglio, N.: What do neural networks learn in image classification? a frequency shortcut perspective. In: ICCV. pp. 1433–1442 (2023)
43. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. CVM **8**(3), 415–424 (2022)
44. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: CVPR. pp. 22–31 (2021)
45. Wu, Y., Liang, L., Zhao, Y., Zhang, K.: Object-aware calibrated depth-guided transformer for rgb-d co-salient object detection. In: ICME. pp. 1121–1126 (2023)
46. Wu, Y., Song, H., Liu, B., Zhang, K., Liu, D.: Co-salient object detection with uncertainty-aware group exchange-masking. In: CVPR. pp. 19639–19648 (2023)
47. Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., Fan, D.P.: Uncertainty-guided transformer reasoning for camouflaged object detection. In: ICCV. pp. 4146–4155 (2021)
48. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: CVPR. pp. 3684–3692 (2018)
49. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: ICCV. pp. 558–567 (2021)
50. Yun, G., Yoo, J., Kim, K., Lee, J., Kim, D.H.: Spanet: Frequency-balancing token mixer using spectral pooling aggregation modulation. In: ICCV. pp. 6113–6124 (2023)
51. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. pp. 5728–5739 (2022)
52. Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., Fan, D.P.: Mutual graph learning for camouflaged object detection. In: CVPR. pp. 12997–13007 (2021)
53. Zhang, H., Li, F., Xu, H., Huang, S., Liu, S., Ni, L.M., Zhang, L.: Mp-former: Mask-piloted transformer for image segmentation. In: CVPR. pp. 18074–18083 (2023)
54. Zhang, M., Xu, S., Piao, Y., Shi, D., Lin, S., Lu, H.: Preynet: Preying on camouflaged objects. In: ACM MM. pp. 5323–5332 (2022)
55. Zhang, X., Zhu, C., Wang, S., Liu, Y., Ye, M.: A bayesian approach to camouflaged moving object detection. TCSVT **27**(9), 2001–2013 (2016)
56. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: CVPR. pp. 4504–4513 (2022)
57. Zhou, T., Zhou, Y., Gong, C., Yang, J., Zhang, Y.: Feature aggregation and propagation network for camouflaged object detection. TIP **31**, 7036–7047 (2022)
58. Zhu, H., Li, P., Xie, H., Yan, X., Liang, D., Chen, D., Wei, M., Qin, J.: I can find you! boundary-guided separated attention network for camouflaged object detection. In: AAAI. vol. 36, pp. 3608–3616 (2022)
59. Zhu, J., Zhang, X., Zhang, S., Liu, J.: Inferring camouflaged objects by texture-aware interactive guidance network. In: AAAI. vol. 35, pp. 3599–3607 (2021)
60. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W.: Biformer: Vision transformer with bi-level routing attention. In: CVPR. pp. 10323–10333 (2023)