Frequency Director: Learnable Mixture of Frequency Experts for Unified Concealed Scene Segmentation

Guangqian Guo[1], Aixi Ren[1], Xuehui Yu[2], Pengxu Wei[3], Yong Guo[4], and Shan Gao[1⋆]

[1] Northwestern Polytechnical University, Xi'an, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Sun Yat-sen University, Guangzhou, China
[4] Max Planck Institute for Informatics, Saarbrücken, Germany
{guogq21, renaixi}@mail.nwpu.edu.cn, yuxuehui17@mails.ucas.ac.cn
weipx3@mail.sysu.edu.cn, guoyongcs@gmail.com, gaoshan@nwpu.edu.cn

Abstract. Concealed object segmentation aims to segment objects that blend into their surroundings, presenting significant challenges due to the high similarity between objects and backgrounds. Existing methods rely on task-specific designs and independent training, leading to limited generalization and structural redundancy. Although a unified parameter-shared model for diverse concealed scenarios is highly desirable and promising, it is hindered by two challenges: large representation gaps across scenarios and the intrinsic difficulty of concealed targets. From a frequency-domain perspective, we observe that different concealed scenarios exhibit more discriminative and interpretable spectral characteristics compared to those in the RGB domain. These findings motivate us to approach unified concealed scene segmentation from the perspective of frequency modulation. To this end, we propose ***Frequency Director (FreqDirect)***, a dynamic frequency adaptation framework that directs spectral representations across diverse concealed scenes. It features a *mixture of frequency experts* to perform adaptive routing over frequency components, capturing heterogeneous task-specific patterns, along with a *spatial commonality anchor* to complement the spatial and structure cues. These components collectively enable effective and efficient adaptation to diverse concealed segmentation tasks within a single model. Extensive experiments on eight concealed scenarios demonstrate that *FreqDirect* surpasses existing unified models and achieves performance comparable to, or better than, specialist models.

Keywords: Dynamic Frequency Experts · Unified Concealed Scene Segmentation · Frequency Learning

1 Introduction

Accurate visual perception [15, 18–21, 25, 46] under diverse and unpredictable real-world conditions remains a fundamental task in computer vision research.

⋆ Corresponding Author

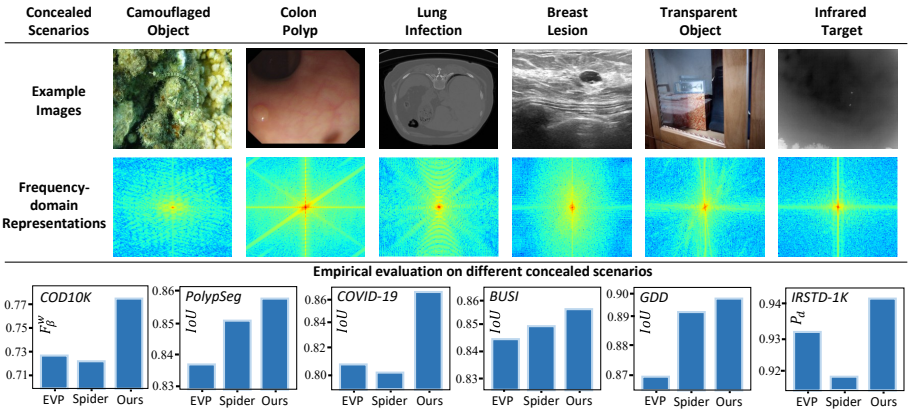

Fig. 1: Example images, frequency-domain representations, and cross-task performance across diverse concealed scenarios. **Top part**: Although these scenarios share the concealed attributes, they differ significantly in data modalities (*e.g.*, RGB, CT, ultrasound) and representational patterns. We observe that such discrepancies become more discriminative and interpretable in the frequency domain. **Bottom part**: Preliminary experiments on datasets [1, 11, 13, 14, 38, 55] corresponding to these scenarios show that previous unified models [32, 57] exhibit sub-optimal and unstable performance when jointly trained on tasks with these stark frequency-domain discrepancies.

This challenge becomes particularly acute when targets are deliberately or naturally hidden, motivating the study of concealed object segmentation (COS), which aims to segment objects that are visually blended with their surroundings [12]. COS is a broad concept that encompasses various sub-domains with concealed properties, including camouflaged object detection [11], polyp segmentation [2], lung infection detection [14], and transparent object segmentation [38], among others. It is particularly challenging due to the intrinsic similarity between the object and its background. Existing methods typically adopt task-specific architectures [14, 38, 44, 55] and train models separately for each task [32, 33], thereby overlooking the commonality of shared knowledge across different concealed scenarios. This approach limits generalization and leads to structural redundancy and inefficient data utilization. To overcome these limitations, developing a single, parameter-shared model capable of generalizing across these diverse scenarios is a highly desirable goal [57]. However, this direction faces two core challenges: ***i) Significant differences in data representation.*** While certain common knowledge may be shared across these concealed scenarios, different concealed scenarios exhibit distinct pattern structures and modality variations (*e.g.*, RGB and CT). That causes large distribution gaps and makes it difficult for a single model to learn a unified representation across them; ***ii) Intrinsic concealment properties increase learning difficulty,*** as they differ fundamentally from general segmentation tasks, making feature discrimination and localization considerably harder.

To address this challenge, the previous generalist model, Spider [57] attempts to unify multiple context-dependent concept segmentation tasks through generic semantic prompts. However, its in-context prompts make it less capable of capturing subtle, low-contrast features, leading to sub-optimal performance (bottom of Fig. 1) on concealed scenarios. These limitations motivate us to seek a more fundamental perspective. To this end, we *shift our perspective to the semantic-decoupled frequency domain* and perform a preliminary analysis (top of Fig. 1) on six representative COS datasets. Note that, unlike the RGB domain with complex semantic dependencies, the frequency domain, governed by physical properties, offers a more direct and interpretable representation [8, 30, 58]. The frequency analysis reveals significant spectral discrepancies across scenarios. Such pronounced frequency variations inspire us to ***reconsider the unification problem from a frequency perspective.***

Building upon the above observations, we argue that a more effective solution for unified COS should enable the model to *modulate its frequency representations* to accommodate the feature divergence across diverse COS tasks. Motivated by this insight, we propose ***FreqDirect***, a lightweight and prompt-free framework that enhances feature diversity. The core component of *FreqDirect* is the *Mixture of Frequency Experts* (MoFE), a multi-expert structure [42, 60] that operates in the frequency-domain to modulate latent representations, enabling adaptive specialization across diverse scenarios. Specifically, each expert in MoFE learns heterogeneous frequency representations, which are then adaptively integrated to model task-specific concealed patterns. However, frequency transformations are inherently global, which can weaken local structural consistency. Since COS relies heavily on structural topology, purely frequency adaptation lacks spatial constraints. Thus, to complement MoFE, a simple Spatial Commonality Anchor (SCA) is introduced to provide spatial regularization. These designs balance task-aware frequency adaptation with task-invariant structural preservation, enhancing cross-domain generalization.

Our method can be seamlessly embedded into intermediate layers of various encoders to regulate feature representations and is fully compatible with standard segmentation frameworks. During training, the pre-trained encoder remains frozen, and only a small number of parameters (approximately 5M) are fine-tuned for efficient adaptation. We validate the effectiveness and generalization ability of our approach through extensive experiments on eight diverse concealed datasets. Overall, our contributions are summarized as follows:

- We propose *FreqDirect*, a new perspective from the frequency domain for unifying cross-domain COS tasks. By fine-tuning only a small number of parameters, it forms a single, parameter-shared model capable of generalizing across diverse concealed segmentatio scenarios.
- We design the Mixture of Frequency Experts method, which dynamically modulates heterogeneous frequency components to enhance the representation diversity of concealed patterns in the frequency domain.
- Without any task-specific designs, our method achieves superior performance over existing unified methods on eight concealed datasets and exhibits strong

generalization to unseen scenarios. Extensive qualitative and quantitative results validate its effectiveness, universality, and adaptability, offering a new robust solution for diverse concealed segmentation.

2 Related Work

2.1 Concealed Object Segmentation

Accurate perception and understanding of concealed scenarios [1,2,11,17,38,55] remain a challenging issue. Some works solve this problem by designing task-specific structures. For example, for the camouflaged object detection [3,4,10,24, 40], SINet [11] designs a bio-inspired network to gradually search and locate camouflaged objects. For polyp segmentation [13, 43, 49, 51], PraNet [13] integrates the reverse attention module to accentuate the boundaries between polyps and their surroundings. However, these methods rely on meticulous model designs, which limit their generalization ability and lead to inefficient data utilization. These expert approaches are computationally expensive and scale poorly, highlighting the need for a single, parameter-shared model that can generalize across diverse concealed tasks.

2.2 Unified Segmentation Models

Recent research has shown a clear shift toward developing general-purpose models. It aims to design a generalist segmentation model capable of adapting to different segmentation tasks [32, 33, 36, 48, 54, 57]. One relevant line of work is EVP [32,33], which injects high-frequency components as a static visual prompt to adapt to different tasks. FOCUS [54] further enhances representation learning by introducing edge priors and utilizing large foundation models [39, 41]. However, these methods adopt static adaptation strategies, which are insufficient for diverse concealed tasks. In contrast, Spider [57] unifies a broad range of context-dependent concept segmentation tasks with multiple semantic concept prompts. While effective for semantic unification, its generic prompt mechanism is not designed to capture or adapt to these fundamental frequency-domain discrepancies explicitly, resulting in sub-optimal performance. In contrast to previous methods, *FreqDirect* is designed to perform frequency modulation within a single, parameter-shared model, enabling to effectively adapt to diverse scenarios without additional in-context prompts.

2.3 Frequency Learning

Compressed representations in the frequency domain contain rich patterns for different vision tasks [8, 16, 22, 28, 58, 59]. Compared to the spatial domain, the features of concealed targets and backgrounds exhibit greater discriminability in the frequency domain. Therefore, in the field of concealed object segmentation, a line of studies [8, 23, 30, 58] have explored frequency features to enhance

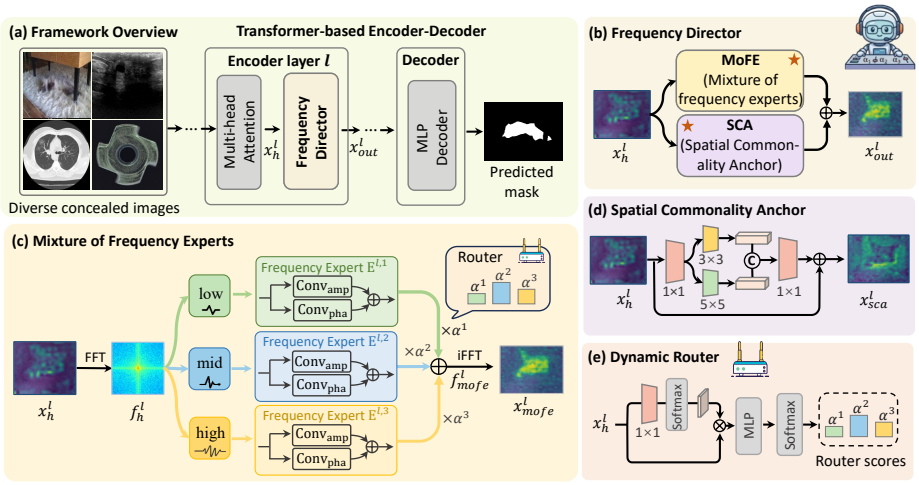

Fig. 2: Our framework (a) retains the generic encoder-decoder structure with the proposed *FreqDirect* (b) embedded into transformer layers, enabling the model to dynamically learn the peculiarities and commonalities of different concealed scenes. Specifically, Mixture of Frequency Experts (c) with Dynamic Router (e) and Spatial Commonality Anchor (d) are introduced for task-specific frequency adaptation and task-invariant spatial regularization, respectively.

model representation. [58] adopts the offline discrete cosine transform to extract frequency features, and then fuses the features from the RGB domain and the frequency domain. [30] aggregates multiscale features from a frequency perspective and enhances the features of the learned important frequency components.

Beyond concealed scenes, recent dense-prediction methods improve the quality of frequency features through local or layer-wise operations. FreqFusion [5] employs adaptive low-pass and high-pass filters to suppress intra-class inconsistency and sharpen blurred boundaries during feature fusion, while FADC [7] adaptively adjusts dilation rates according to local frequency content. To preserve high-frequency details against aliasing during downsampling, SFM [6] modulates high-frequency components to a lower band before downsampling and demodulates them afterwards. Inspired by these studies, we further explore frequency-domain priors under a unified concealed scene segmentation paradigm, demonstrating that frequency modulation in the latent representation can effectively bridge cross-scenario discrepancies.

3 Frequency Director

In the following, we introduce the proposed ***FreqDirect*** for diverse concealed scene segmentation tasks. The overall framework is illustrated in Fig. 2. Our method acts as a *frequency director to modulate the latent representations in the frequency domain* to enable the model to learn the peculiarities of diverse concealed scenes, as overviewed in Sec. 3.1. To dynamically capture the distinct

characteristics of different concealed scenarios from a frequency-domain perspective, we introduce the core design of the Mixture of Frequency Experts (MoFE) for *task-specific frequency adaptation* (Sec. 3.2). To complement the spatial and structure cues, we further insert a Spatial Commonality Anchor (SCA) for *task-invariant spatial regularization* (Sec. 3.3). Finally, the overall training method is outlined in Sec. 3.4.

3.1 Framework Overview

As shown in Fig. 2, the core component in *FreqDirect* is a dynamic MoFE $\Phi^l_{\text{mofe}}(\cdot)$ (where l indicates the encoder layer index), which is responsible for learning task-specific peculiarities. It is formulated as a multi-expert structure, consisting of multiple frequency-domain experts that model representations in different frequency bands, along with a dynamic routing mechanism to adaptively aggregate expert outputs. To capture task-invariant spatial details, a SCA $\Phi^l_{\text{sca}}(\cdot)$ is introduced for spatial regularization. The task-specific and shared representations are then integrated to form the output features. In detail, we first map latent features of the l-th layer $x^l \in \mathbb{R}^{C\times H\times W}$ to a lower-dimensional space ($x^l_{\text{h}} \in \mathbb{R}^{\frac{C}{\gamma}\times H\times W}$) for parameter efficiency, where γ is the scale factor to control the dimension. Then, x^l_{h} is sent to the parallel frequency and spatial paths.

FreqDirect can be embedded into intermediate layers of a segmentation encoder, enabling dynamic feature adaptation during joint training across multiple domains. Our overall framework follows a standard encoder–decoder architecture. In the following, we detail each component of the network.

3.2 Task-Specific Frequency Adaptation

To capture the peculiarities of different scenarios in the frequency domain, we formalize MoFE (as shown in Fig. 2 (c)) as a dynamic multi-expert structure that enhances feature encoding diversity by introducing multiple frequency experts. This design is inspired by a key observation, *i.e.*, different concealed scenarios exhibit explicit and distinctive spectral characteristics (Fig. 1). Based on this insight, we transform features into the frequency domain and construct a dynamic network to adaptively enhance task-specific frequency representations. Each expert specializes in learning representations for particular frequency components (*e.g.*, high-frequency or low-frequency) and then their outputs are adaptively assembled to adapt to different scenarios.

Frequency Expert Learning. Specifically, given the feature of the l-th layer x^l_h, we use the Fast Fourier Transform (FFT) to transfer it from the spatial domain to the frequency domain: $f^l = \mathcal{F}(x^l_h)$, where $f^l \in \mathbb{C}^{\frac{C}{\gamma}\times H\times W}$ denotes the complex-valued frequency representation and $\mathcal{F}$ denotes the FFT operation. We partition the frequency spectrum into N disjoint bands using a set of predefined radial thresholds $\{\rho_0, \rho_1, ..., \rho_N\}$, with $\rho_0 = 0$ and ρ_N set to the maximum radial distance. Based on these radial thresholds, N binary masks $\{\mathrm{M}^i, i \in [1, N]\}$ are

defined. The mask for the i-th frequency band is formulated as:

$$\mathrm{M}^i(u,v) = \begin{cases} 1 & , \text{ if } \rho_{i-1} < r(u,v) \leq \rho_i \\ 0 & , \text{ otherwise} \end{cases} \tag{1}$$

The radial distance $r(u,v)$ is computed as:

$$r(u,v) = \max(|u - \frac{H}{2}|, |v - \frac{W}{2}|), \tag{2}$$

where H and W denote the height and width of the frequency map, respectively. Each mask corresponds to a specific frequency band (*e.g.*, low, middle, or high frequency regions). The i-th frequency subspace is obtained by:

$$f^{l,i} = \mathrm{M}^i \odot f^l, \tag{3}$$

where $\odot$ denotes element-wise multiplication. This operation decomposes f^l into a set of band-wise frequency features.

We further define a group of frequency-domain experts $\{\mathrm{E}^{l,i}(\cdot), i \in [1,N]\}$, each responsible for modeling a specific frequency band. Rather than directly operating on complex-valued features, each expert performs separate affine transformations on the amplitude and phase components. Let $f^{l,i}_{\mathrm{amp}}$ and $f^{l,i}_{\mathrm{pha}}$ denote the amplitude and phase maps, respectively. The i-th expert is formulated as:

$$\mathrm{E}^{l,i}(f^{l,i}) = \mathrm{Complex}(\mathrm{Conv}_{\mathrm{amp}}(f^{l,i}_{\mathrm{amp}}), \mathrm{Conv}_{\mathrm{pha}}(f^{l,i}_{\mathrm{pha}})), \tag{4}$$

where $\mathrm{Conv}_{\mathrm{amp}}(\cdot)$ and $\mathrm{Conv}_{\mathrm{pha}}(\cdot)$ denote learnable transformations applied to amplitude and phase, respectively, and $\mathrm{Complex}(\cdot,\cdot)$ reconstructs the complex representation. Through this process, these experts form a set of *frequency representation bases*. Discriminative features for different scenarios are then obtained by adaptively assembling these representative frequency bases.

Router Score Computing. As shown in Fig. 2 (e), we introduce a router network $\mathcal{R}^l(\cdot)$ to adaptively control the contribution of each expert by computing router score $\alpha^l \in \mathbb{R}^N$. Given the input feature of x^l_{h}, we first obtain the spatial attention weight $w^l \in \mathbb{R}^{HW\times 1}$ by employing 1×1 convolotion which compresses the channel dimension of the input to 1, followed by the softmax on the spatial dimension, and then reshaping it to 2D vector:

$$w^l = \mathrm{Reshape}(\mathrm{Softmax}(\mathrm{Conv}(x^l_{\mathrm{h}}))). \tag{5}$$

We then apply w^l on the x^l_{h} to perform spatially weighted summation to obtain a channel vector that will go through a MLP and a softmax function to generate the router score:

$$\alpha^l = \mathcal{R}^l(x^l_{\mathrm{h}}) = \mathrm{Softmax}(\mathrm{MLP}(w^l \times \mathrm{Reshape}(x^l_{\mathrm{h}}))), \tag{6}$$

where we reshape x^l_{h} to 2D for dimension alignment.

Dynamically Assembling. Finally, the frequency representations produced by different experts are adaptively combined according to the router scores:

$$f^l_{\text{mofe}} = \sum_{i=1}^{N} (\alpha^{l,i} * E^{l,i}(f^{l,i})). \tag{7}$$

The aggregated frequency feature is then transformed back to the spatial domain via the inverse Fast Fourier Transform (iFFT, $\mathcal{F}^{-1}(\cdot)$), as: $x^l_{\text{mofe}} = \mathcal{F}^{-1}(f^l_{\text{mofe}})$.

3.3 Task-Invariant Spatial Regularization

While MoFE is designed to capture task-specific peculiarities, such dynamic routing inherently increases representation variance across domains. More importantly, frequency-domain operations, *i.e.*, the Fourier transform, convert local spatial details into global coefficients, thereby weakening explicit spatial constraints. Pure frequency adaptation lacks an explicit mechanism to preserve these spatial properties, especially for subtle boundaries and small concealed targets, which is very detrimental to the perception of concealed targets that require structural information. To address this issue, we introduce the Spatial Commonality Anchor (SCA), $\Phi^l_{\text{sca}}(\cdot)$, as illustrated in Fig. 2 (d). Unlike task-specific modulation, SCA acts as a spatial regularization that explicitly preserves structural priors. By aggregating features from multiple receptive fields, SCA extracts task-shared spatial cues, such as boundary continuity and shape consistency, into a unified representation. Consequently, the proposed dual-path design balances frequency adaptability with spatial stability, leading to improved cross-domain generalization. Empirical validation (Tab. 3) and feature visualization (Fig. 5) demonstrates that this branch effectively captures spatial cuesand yields substantial performance improvements.

3.4 Training Method and Details

To demonstrate the general applicability of *FreqDirect*, we integrate it into several transformer-based backbones, including the MiT series (MiT-B4 and MiT-B5) [50], the Swin series (Swin-B and Swin-L) [35], and PVTv2 [47]. We adopt a simple and efficient MLP decoder [50] in our model, without any task-specific modifications. During training, the pre-trained image encoder is frozen, and only the proposed modules and the decoder are fine-tuned. Our model maintains 100% parameter sharing across tasks for feature extraction and masks prediction. Our method requires only training once, where a minimal set of parameters (5% of the base model) is tuned to adapt to diverse concealed scenarios. Joint training is conducted on the unified concealed object segmentation dataset, ensuring balanced learning across different scenarios by sampling data evenly from all the scenarios within each batch. Binary Cross-Entropy (BCE) and IoU losses are employed as segmentation loss functions, as:

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{BCE}}(m_{\text{p}}, m_{\text{g}}) + \mathcal{L}_{\text{IoU}}(m_{\text{p}}, m_{\text{g}}), \tag{8}$$

where m_{p} and m_{g} indicate predicted and GT masks.

Table 1: Quantitative comparison among our method, other unified models, and task-specific models on six concealed datasets. Our method outperforms other unified models and achieves superior or comparable performance to specialized models. We make FOCUS Gray because it uses a large foundation model. The words in **boldface** and underline indicate the best and second-best results, respectively.

Method	COD10K		PolypSeg		COVID-19		BUSI		GDD		IRSTD-1K	
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑	M ↓	$F_a \downarrow$	$P_d \uparrow$
Specialized Models:												
HitNet [26]	0.8470	0.7900	-	-	-	-	-	-	-	-	-	-
FocusDiff [56]	0.8630	0.7850	-	-	-	-	-	-	-	-	-	-
FSEL [44]	0.8730	0.8000	-	-	-	-	-	-	-	-	-	-
ASPS [29]	-	-	0.7994	0.8784	-	-	-	-	-	-	-	-
RollingUNet [34]	-	-	0.7833	0.6920	0.7636	0.6275	0.7959	0.7384	-	-	-	-
CMUNet [45]	-	-	-	-	-	-	0.8302	0.5452	-	-	-	-
GDNet [38]	-	-	-	-	-	-	-	-	0.8760	0.0630	-	-
GhostingNet [52]	-	-	-	-	-	-	-	-	0.8930	0.0540	-	-
ISNet [55]	-	-	-	-	-	-	-	-	-	-	27.92	92.59
MSHNet [31]	-	-	-	-	-	-	-	-	-	-	15.03	93.88
Unified Models (One model for multiple tasks):												
SegGPT (zero-shot) [48]	0.6677	0.4539	0.6304	0.4906	0.5930	0.2724	0.5266	0.1594	0.3517	0.4118	240.9	18.05
EVPv1 (MiT-B4) [32]	0.8384	0.7381	0.8392	0.7833	0.8190	0.6391	0.8501	0.8190	0.8706	0.0687	17.82	91.93
EVPv2 (MiT-B4) [33]	0.8375	0.7373	0.8415	0.7856	0.8358	0.6547	0.8519	0.8218	0.8717	0.0673	11.76	91.72
Spider (Swin-B) [57]	0.8464	0.7225	0.8303	0.7840	0.8068	0.6836	0.8499	0.8132	0.8924	0.0556	27.18	84.10
FOCUS (DINOv2-L) [54]	0.8515	0.7719	0.8479	0.8078	0.6986	0.5305	0.8253	0.7817	0.9013	0.0494	80.09	47.66
FreqDirect (PVTv2)	0.8534	0.7732	0.8542	0.8082	0.8551	0.6755	0.8456	0.8127	0.8896	0.0574	22.05	93.14
FreqDirect (MiT-B4)	0.8542	0.7667	0.8488	0.7980	0.8632	0.6837	0.8512	0.8168	0.8938	0.0545	**7.07**	91.17
FreqDirect (MiT-B5)	0.8574	0.7717	0.8491	0.7984	0.8608	0.6864	**0.8567**	**0.8269**	0.8863	0.0590	20.55	93.34
FreqDirect (Swin-B)	0.8505	0.7609	0.8519	0.7989	0.8638	**0.6922**	0.8520	0.8251	0.8951	0.0553	20.66	91.81
FreqDirect (Swin-L)	**0.8811**	**0.8166**	**0.8624**	**0.8202**	**0.8646**	0.6910	0.8526	0.8226	**0.9051**	**0.0485**	17.68	**94.38**

4 Experiment

4.1 Experimental Setup

Dataset. To enable a single model to learn robust characteristics from diverse concealed scenarios, we construct a unified dataset for joint training and evaluation by merging several widely used datasets in the community. For training, we combine all training samples from six concealed scenarios into a single dataset for joint learning, including COD10K [11], PolypSeg [13], COVID-19 [14], BUSI [1], GDD [38], IRSTD-1K [55]. For evaluation, performance is assessed separately on each scenario. Additionally, we select two unseen scenarios: Industrial Defect [12] and Camouflaged Plant [53], for zero-shot evaluation to test the model's generalization ability to unseen concealed objects. More details about the datasets and task are shown in the *supplementary material*.

Evaluation metrics. We use the same evaluation metrics as the expert models for each task. Specifically, the camouflaged object segmentation task is evaluated using the S-measure S_m [9] and weighted F-measure F_β^w [37]. The colon polyp, lung infection, and breast lesion image segmentation tasks are evaluated using mean Intersection over Union (mIoU) and Dice Coefficient (Dice). GDD is evaluated using mIoU and Mean Squared Error (MSE). The infrared target detection task is evaluated using F_α and P_d.

Implement details. Our model is trained using the AdamW [27] optimizer. During training, data from the six scenarios are mixed within each iteration,

Table 2: Zero-shot evaluation on unseen scenarios. Our method consistently outperforms other unified models.

Dataset	Metric	SegGPT	EVPv1	EVPv2	Spider	FOCUS	FreqDirect (Ours)				
		zero-shot	MiT-B4	MiT-B4	Swin-B	DINOv2-L	PVTv2	MiT-B4	MiT-B5	Swin-B	Swin-L
CDS2K	$S_\alpha \uparrow$	0.4887	0.5536	0.5428	0.5178	0.5672	<u>0.5571</u>	0.5567	0.5575	**0.5766**	0.5607
	$F_\beta^w \uparrow$	0.0155	0.2390	0.2124	0.2315	**0.2793**	0.2488	<u>0.2627</u>	0.2757	0.2641	0.2365
PlantCAMO	$S_\alpha \uparrow$	0.5710	0.6801	0.6841	0.6710	0.6734	0.6745	**0.6919**	0.6750	0.6832	<u>0.6881</u>
	$F_\beta^w \uparrow$	0.3143	0.5435	0.5148	0.5208	0.5231	0.5221	**0.5463**	0.5212	0.5237	<u>0.5445</u>

with three samples per task, resulting in a total batch size of 18. The initial learning rate is set to 4×10^{-4} with a cosine decay schedule. Training is conducted for 20 epochs and takes approximately 8 hours on a single RTX 4090 GPU. For frequency implementation, we apply a 2D FFT over the spatial dimensions (H, W) in a channel-wise manner, using orthonormal normalization. FFT are performed at intermediate encoder stages with reduced spatial resolution and channel dimension, which keeps the computational overhead modest. The frequency spectrum is partitioned based on radial distance after FFT shift.

4.2 Performance Evaluation

In Tab. 1, we evaluate the performance of the proposed method on six scenarios. We compare our model with leading unified models, including SegGPT [48], EVPv1/v2 [32,33], Spider [57], and FOCUS [54]. Results for other unified baselines are re-implemented by us using their official code under our joint training setting for a fair comparison. Our method achieves consistently superior performance across all tasks. Although FOCUS [54] adopts large pre-trained foundation models (DINOv2-L+CLIP [39, 41]), our results remain competitive. the performance gains are most pronounced on tasks with distinctive frequency characteristics that differ from standard RGB images. For example, on the COVID-19 (CT) dataset, our method achieves about 6 and 4 points of improvement over Spider and EVP with the same backbone, respectively. This observation suggests that generic or static adaptation strategies struggle to handle the severe modality and frequency shifts present in CT images. Remarkably, our unified model, with fully shared parameters and trained only once, achieves performance comparable to or even surpassing these expert models within their respective domains.

Additionally, in Tab. 2, we evaluate the zero-shot segmentation performance on two unseen datasets: CDS2K [12] and PlantCAMO [53]. Our method demonstrates stronger generalization than other competitors. Fig. 3 showcases the qualitative comparisons of our method on eight concealed scenarios, showcasing its adaptability in handling diverse environments.

4.3 Ablation Study

In this part, we conduct extensive experiments in Tab. 3 to evaluate the contribution of each component in our framework and to determine the optimal

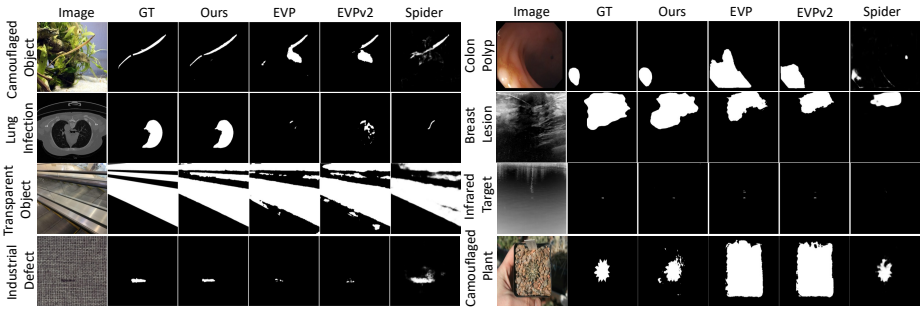

Fig. 3: Qualitative comparisons across diverse concealed scenarios demonstrate that our approach produces more accurate segmentation.

hyper-parameter settings. Unless otherwise specified, all ablation experiments are conducted using the MiT-B4 [50] backbone.

Effect of each component. We first conduct a progressive ablation study in the first part of Tab. 3. Two additional baselines, *i.e.*, *Spatial MoE* and *Static Freq Adapter* are introduced to disentangle the contributions of dynamic adaptation and frequency-domain modeling. ***i)*** We begin with the simplest baseline, which embeds a standard adapter into the encoder and fine-tunes both the adapter and the decoder. This approach provides insufficient adaptation, resulting in the lowest performance. ***ii)*** To verify the role of frequency information, we design a *Static Frequency Adapter (SFA)* that injects pre-computed frequency components into the features, similar to EVP [32]. This yields a moderate improvement, confirming that frequency-domain representations are indeed helpful for modeling concealed patterns. However, its static design cannot adapt to the distinct frequency distributions across scenarios. ***iii)*** To evaluate the effect of dynamic routing, we construct a *Spatial MoE* baseline by replacing the frequency experts in our MoFE with three 3×3 convolutional experts operating in the spatial domain while retaining the same routing module. This dynamic structure improves performance, indicating that dynamically combining expert outputs enhances representational flexibility. But the results also underperform. We attribute this to the fact that all experts operate on the entire feature space, ignoring frequency-domain priors across tasks and leading to redundant learning among experts. ***iv)*** Next, we introduce proposed MoFE, which performs dynamic routing directly in the frequency domain. This design surpasses both static and spatial dynamic baselines with a comparable parameter budget. The result confirms that our performance gain stems not merely from the MoE structure, but from adaptive frequency selection that effectively aligns diverse concealed scenes. ***v)*** When further combined with the simple SCA module to complement spatial priors, the complete FreqDirect framework achieves the best performance, benefiting from complementary task-shared and task-specific feature learning.

Evaluation of the dynamic router. We further evaluate the role of dynamic routing. As a control experiment, we replace the adaptive router with a fixed routing strategy, where the routing weights are uniformly averaged across ex-

Table 3: Ablation experiments to evaluate the effect of various components of the proposed method. The performance is measured across six datasets. The gray background indicates our default setting. The words in **boldface** indicate the best results.

Method	Learnable Params	COD10K		PolypSeg		COVID-19		BUSI		GDD		IRSTD-1K	
		$S_\alpha \uparrow$	$F_\beta^w \uparrow$	IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑	M ↓	$F_\alpha \downarrow$	$P_d \uparrow$
Effect of Each Component:													
Simple Adapter	3.97 M	0.8172	0.7120	0.8132	0.7422	0.8014	0.6143	0.8235	0.7845	0.8215	0.0952	31.45	87.98
Static Freq. Adap.	4.21 M	0.8312	0.7253	0.8220	0.7601	0.8141	0.6276	0.8370	0.8005	0.8427	0.0797	16.47	89.36
Sptial MoE	5.08 M	0.8245	0.7054	0.8117	0.7386	0.8230	0.6239	0.8403	0.8047	0.8356	0.0898	18.71	89.20
MoFE	5.10 M	0.8457	0.7501	0.8366	0.7816	0.8516	0.6721	0.8425	0.8145	0.8605	0.0735	13.62	90.24
MoFE & SCA	5.26 M	**0.8542**	**0.7667**	**0.8488**	**0.7980**	**0.8632**	**0.6837**	**0.8512**	**0.8168**	**0.8938**	**0.0545**	**7.07**	**91.17**
Evaluation of the Dynamic Router:													
Fixed-router	5.26 M	0.8344	0.7273	0.8338	0.7523	0.8165	0.6343	0.8386	0.8021	0.8296	0.0777	17.45	89.35
Dynamic router	5.26 M	**0.8542**	**0.7667**	**0.8488**	**0.7980**	**0.8632**	**0.6837**	**0.8512**	**0.8168**	**0.8938**	**0.0545**	**7.07**	**91.17**
Evaluation of the Number of Experts:													
2	5.24 M	0.8419	0.7466	**0.8533**	**0.8064**	0.8492	0.6747	0.8493	0.8169	0.8834	0.0606	9.11	90.85
3	5.26 M	**0.8542**	**0.7667**	0.8488	0.7980	**0.8632**	**0.6837**	0.8512	0.8168	**0.8938**	**0.0545**	**7.07**	91.17
4	5.27 M	0.8525	0.7619	0.8509	0.8034	0.8496	0.6820	**0.8543**	**0.8255**	0.8843	0.0603	11.10	**91.37**
Evaluation of Insertion Position:													
Even-num Layers	4.18 M	0.8532	0.7647	0.8499	**0.8025**	0.8569	0.6810	0.8487	0.8126	0.8857	0.0591	18.80	90.83
Odd-num Layers	4.08 M	0.8519	0.7621	0.8477	0.7990	0.8548	0.6748	**0.8525**	**0.8199**	0.8858	0.0586	11.67	**92.25**
Stage 4	3.47 M	0.8109	0.6928	0.7716	0.6681	0.8323	0.6438	0.8190	0.7669	0.8083	0.1051	13.45	90.02
Stage 4, 3	5.15 M	0.8481	0.7530	**0.8506**	0.8018	0.8390	0.6657	0.8490	0.8143	0.8794	0.0626	9.91	91.81
Stage 4, 3, 2	5.24 M	0.8519	0.7615	0.8459	0.7959	0.8562	0.6742	0.8507	0.8167	0.8826	0.0603	9.71	91.12
Stage 4, 3, 2, 1	5.26 M	**0.8542**	**0.7667**	0.8488	0.7980	**0.8632**	**0.6837**	0.8512	0.8168	**0.8938**	**0.0545**	**7.07**	91.17

perts. In this setting, the router cannot adaptively adjust frequency components based on the input. The results show that fixed routing leads to clear performance degradation, while dynamic routing significantly improves performance. This observation confirms that the gains stem from input-aware frequency modulation, rather than merely increasing model capacity through additional experts.

Evaluation of the number of experts. We conduct an ablation study on the number of frequency experts in the MoFE module. Each expert processes an evenly partitioned subset of the frequency components. While increasing the number of experts enhances the diversity of feature representations, it also introduces greater training complexity. As shown in the third part of Tab. 3, the configuration with three experts achieves the best trade-off between performance and computational cost. This setup effectively captures diverse input patterns without incurring excessive model complexity.

Evaluation of insertion position. We explore different insertion positions for the proposed *FreqDirect*, as summarized in the fourth part of Tab. 3. Instead of placing it in every transformer layer, we first examine inserting it into only the even- or odd-numbered layers. The result shows that inserting into even-numbered layers achieves comparable performance to all-layer insertion while reducing parameters from 5.26M to 4.18M. We further analyze the impact of incorporating *FreqDirect* at different stages. Progressive inclusion from stage 4 to stage 1 shows steady performance improvements. Notably, inserting into stages 4, 3 already achieves a strong performance with only 5.15M parameters.

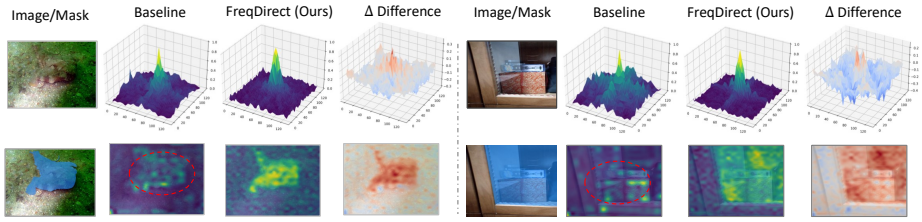

Fig. 4: Comparison of frequency spectra (top) and spatial response maps (bottom) between the baseline and our proposed *FreqDirect*. Our method exhibits more selective spectral modeling and structurally coherent spatial responses, enhancing object-relevant activations and suppressing background noise.

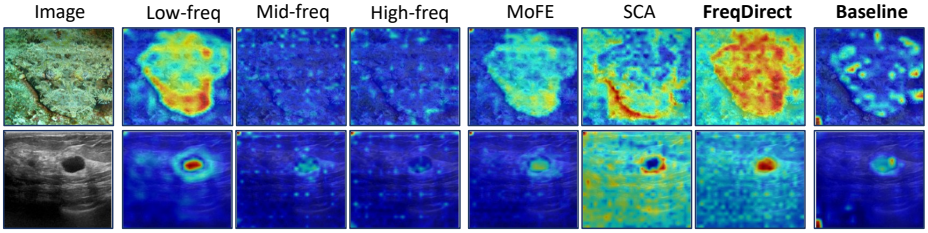

Fig. 5: Visualization of feature activations from different components of FreqDirect. Frequency experts (low, mid, and high) capture complementary spectral patterns, SCA emphasizes structural boundaries, and their integration leads to more discriminative and coherent object representations.

4.4 Model Analysis and Discussion

Model analysis. The effectiveness of our approach stems from operating in the frequency domain, a semantically decoupled space governed by physical properties rather than high-level semantic correlations. Compared with the spatial domain, where object/background semantics are entangled and difficult to separate, the frequency domain provides more direct, interpretable, and discriminative cues for concealed scenes. By transforming features into the spectral space, the model only needs to optimize over frequency attributes, substantially simplifying the learning process and making scenario-specific differences more accessible. Fig. 4 visualizes the frequency spectra and corresponding spatial activation maps of the baseline and *FreqDirect*, along with their differences. Compared with the baseline, *FreqDirect* produces more concentrated and structurally coherent spatial feature responses while suppressing spurious activations in irrelevant regions. The frequency difference maps reveal *enhanced target-related components* and *attenuated peripheral high-frequency noise*, indicating more *selective spectral modeling*. By jointly amplifying discriminative object signals and filtering task-irrelevant interference, *FreqDirect* yields feature representations with improved structural consistency and higher signal-to-noise characteristics.

Analysis of feature visualization. Fig. 5 illustrates feature activations from different components of *FreqDirect*, including low-, mid-, and high-frequency ex-

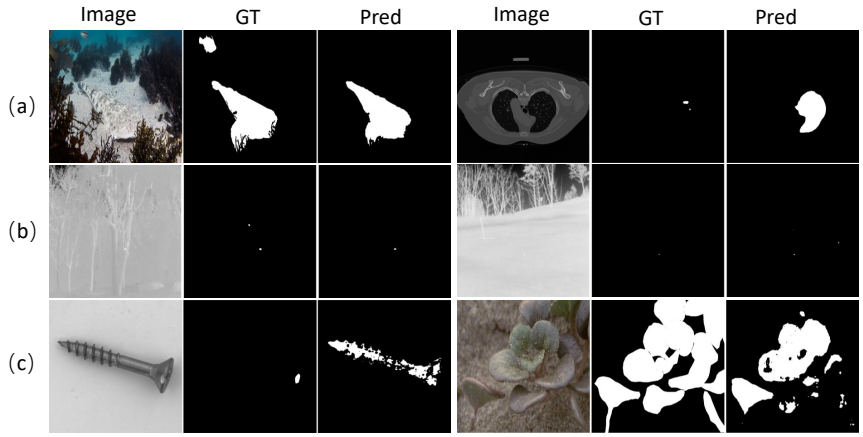

Fig. 6: Failure cases of *FreqDirect* across multiple concealed scenarios.

Table 4: Comparison of parameters, efficiency, and model complexity with other unified models.

Method	Params	MACs	Latency	Avg. IoU
EVP (MiT-B4) [32]	64.5M	20.4G	0.055	.842
EVPv2 (MiT-B4) [33]	64.5M	20.4G	0.060	.844
Spider (Swin-B) [57]	167.0M	90.7G	0.097	.845
FOCUS (DINOv2) [54]	500.7M	975.7G	0.080	.818
FreqDirect (MiT-B4)	64.6M	20.4G	0.076	.864
FreqDirect (Swin-B)	92.5M	62.9G	0.079	**.866**

perts, the aggregated MoFE output, the SCA branch, and the final prediction, with the baseline shown for comparison. Distinct frequency experts exhibit heterogeneous response patterns, capturing complementary spectral characteristics of concealed targets. The SCA branch highlights clear structural cues, particularly boundary continuity, indicating effective spatial regularization. After integration, *FreqDirect* produces more concentrated and coherent object-aware activations than the baseline, demonstrating that each component contributes complementary information and jointly enhances representation discriminability.

Comparison on parameters, model complexity, and efficiency. As shown in Tab. 4, our proposed method consistently outperforms prior approaches in terms of average IoU, while maintaining comparable computational cost. Notably, *FreqDirect* achieves the best IoU of 0.864, demonstrating both high segmentation accuracy and strong parameter efficiency. This highlights the effectiveness of our adaptive frequency learning in balancing performance and model complexity.

Comparison of joint training and separate training. This study compares the performance of joint training (JT) and separate training (ST) strategies across four datasets. Specifically, CDS2K and PlantCAMO are used for zero-shot

Table 5: Comparative analysis of joint-training (JT) and separate training (ST). CDS2K and PlantCAMO are used for zero-shot test. Results highlight the superior generalization capability of JT, demonstrating consistent performance improvements compared to ST across diverse scenarios.

	PolypSeg		COVID-19		CDS2K		PlantCAMO	
	IoU ↑	Dice ↑	IoU ↑	Dice ↑	S_α ↑	F_β^w ↑	S_α ↑	F_β^w ↑
ST	0.8471	0.7962	-	-	0.4982	0.0936	0.5783	0.3771
	-	-	0.8412	0.6525	0.4919	0.0635	0.4198	0.0885
JT	**0.8513**	**0.8014**	**0.8502**	**0.6782**	**0.5551**	**0.2540**	**0.6453**	**0.4521**

evaluation. We train each model separately on each task with the same number of iterations and architectures as done in joint training. As shown in Tab. 5, the jointly trained models consistently outperform the separately trained ones on all the datasets, especially in zero-shot testing.

Analysis of failure cases. Despite its strong performance, *FreqDirect* still exhibits limitations in certain challenging cases. As shown in Fig. 6, the model struggles to detect objects with highly subtle boundaries, such as camouflaged animals or blurred lesions, where foreground and background share nearly indistinguishable frequency cues. In infrared small-target scenarios, extremely weak and sparsely activated signals are easily suppressed by thermal noise, leading to missed detections. These cases highlight the inherent difficulty of modeling ultra-weak or unstable frequency patterns and point to future directions such as finer-grained frequency decomposition and improved spatial–frequency coupling.

5 Conclusion

In this work, we presented *FreqDirect*, a frequency-aware framework for unified concealed scene segmentation. Guided by the frequency-domain analysis, we seek to solve the cross-scenario segmentation from the frequency perspective. Our method dynamically modulates latent features through a Mixture of Frequency Experts and complements this with lightweight spatial commonality anchor to capture spatial details. It introduces minimal additional parameters while significantly improving adaptability across scenarios. Extensive experiments on eight datasets demonstrate its superior performance over unified and specialist models alike, along with strong generalization to unseen scenarios. Our findings highlight the value of frequency-domain modeling for multi-scenario learning and establish *FreqDirect* as a promising direction for advancing concealed scene understanding.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No.62372382, 62376292, and Guangdong Provincial General Fund No. 2024A1515010208.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Biswas, R.: Polyp-sam++: Can a text guided sam perform better for polyp segmentation? arXiv preprint arXiv:2308.06623 (2023)
3. Chen, H., Shao, D., Guo, G., Gao, S.: Just a hint: Point-supervised camouflaged object detection. In: European Conference on Computer Vision. pp. 332–348. Springer (2024)
4. Chen, H., Wei, P., Guo, G., Gao, S.: Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In: European Conference on Computer Vision. pp. 315–331. Springer (2024)
5. Chen, L., Fu, Y., Gu, L., Yan, C., Harada, T., Huang, G.: Frequency-aware feature fusion for dense image prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2024)
6. Chen, L., Fu, Y., Gu, L., Zheng, D., Dai, J.: Spatial frequency modulation for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2025)
7. Chen, L., Gu, L., Fu, Y.: Frequency-adaptive dilated convolution for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
8. Cong, R., Sun, M., Zhang, S., Zhou, X., Zhang, W., Zhao, Y.: Frequency perception network for camouflaged object detection. In: Proceedings of the 31st ACM international conference on multimedia. pp. 1179–1189 (2023)
9. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
10. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6024–6042 (2021)
11. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2777–2787 (2020)
12. Fan, D.P., Ji, G.P., Xu, P., Cheng, M.M., Sakaridis, C., Van Gool, L.: Advances in deep concealed scene understanding. Visual Intelligence **1**(1), 16 (2023)
13. 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: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
14. 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. IEEE transactions on medical imaging **39**(8), 2626–2637 (2020)
15. Gao, S., Guo, G., Huang, H., Chen, C.P.: Go deep or broad? exploit hybrid network architecture for weakly supervised object classification and localization. IEEE Transactions on Neural Networks and Learning Systems **36**(3), 3976–3989 (2023)
16. Gueguen, L., Sergeev, A., Kadlec, B., Liu, R., Yosinski, J.: Faster neural networks straight from jpeg. Advances in Neural Information Processing Systems **31** (2018)
17. Guo, G., Chen, P., Guo, Y., Chen, H., Zhang, B., Gao, S.: Boosting segment anything model to generalize visually non-salient scenarios. IEEE Transactions on Image Processing **35**, 1143–1157 (2026)

18. Guo, G., Chen, P., Yu, X., Han, Z., Ye, Q., Gao, S.: Save the tiny, save the all: Hierarchical activation network for tiny object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(1), 221–234 (2024)
19. Guo, G., Guo, Y., Yu, X., Li, W., Wang, Y., Gao, S.: Segment any-quality images with generative latent space enhancement. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 2366–2376 (June 2025)
20. Guo, G., Ren, A., Guo, Y., Yu, X., Tian, J., Li, W., Wang, C., Wang, Y., Gao, S.: Towards any-quality image segmentation via generative and adaptive latent space enhancement. arXiv preprint arXiv:2601.02018 (2026)
21. Guo, G., Shao, D., Zhu, C., Meng, S., Wang, X., Gao, S.: P2p: Transforming from point supervision to explicit visual prompt for object detection and segmentation. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (2024)
22. Guo, H., Dai, T., Bai, Y., Chen, B., Ren, X., Zhu, Z., Xia, S.T.: Parameter efficient adaptation for image restoration with heterogeneous mixture-of-experts. Advances in Neural Information Processing Systems **37**, 13522–13547 (2025)
23. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 22046–22055 (2023)
24. He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y., Kong, L., Fan, D.P., Li, K., Farsiu, S.: Run: Reversible unfolding network for concealed object segmentation. In: Forty-second International Conference on Machine Learning (2025)
25. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
26. 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: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Li, C., Guo, C.L., Liang, Z., Zhou, S., Feng, R., Loy, C.C., et al.: Embedding fourier for ultra-high-definition low-light image enhancement. In: The Eleventh International Conference on Learning Representations (2023)
29. Li, H., Zhang, D., Yao, J., Han, L., Li, Z., Han, J.: Asps: Augmented segment anything model for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 118–128. Springer (2024)
30. Lin, J., Tan, X., Xu, K., Ma, L., Lau, R.W.: Frequency-aware camouflaged object detection. ACM Transactions on Multimedia Computing, Communications and Applications **19**(2), 1–16 (2023)
31. Liu, Q., Liu, R., Zheng, B., Wang, H., Fu, Y.: Infrared small target detection with scale and location sensitivity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17490–17499 (2024)
32. Liu, W., Shen, X., Pun, C.M., Cun, X.: Explicit visual prompting for low-level structure segmentations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19434–19445 (2023)
33. Liu, W., Shen, X., Pun, C.M., Cun, X.: Explicit visual prompting for universal foreground segmentations. arXiv preprint arXiv:2305.18476 (2023)
34. Liu, Y., Zhu, H., Liu, M., Yu, H., Chen, Z., Gao, J.: Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence (2024)

35. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
36. Luo, Z., Liu, N., Zhao, W., Yang, X., Zhang, D., Fan, D.P., Khan, F., Han, J.: Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17169–17180 (2024)
37. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 248–255 (2014)
38. Mei, H., Yang, X., Wang, Y., Liu, Y., He, S., Zhang, Q., Wei, X., Lau, R.W.: Don't hit me! glass detection in real-world scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3687–3696 (2020)
39. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (2024)
40. 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: Proceedings of the IEEE Conference on computer vision and pattern recognition. pp. 2160–2170 (2022)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
42. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
43. Shao, H., Zhang, Y., Hou, Q.: Polyper: Boundary sensitive polyp segmentation. In: Proceedings of the AAAI conference on artificial intelligence. pp. 4731–4739 (2024)
44. Sun, Y., Xu, C., Yang, J., Xuan, H., Luo, L.: Frequency-spatial entanglement learning for camouflaged object detection. In: European Conference on Computer Vision. pp. 343–360. Springer (2024)
45. Tang, F., Wang, L., Ning, C., Xian, M., Ding, J.: Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI). pp. 1–5. IEEE (2023)
46. Wang, C., Guo, G., Liu, C., Shao, D., Gao, S.: Effective rotate: Learning rotation-robust prototype for aerial object detection. IEEE transactions on geoscience and remote sensing **62**, 1–14 (2024)
47. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
48. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1130–1140 (2023)
49. Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z.: Weakpolyp: You only look bounding box for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 757–766. Springer (2023)

50. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
51. Xu, Z., Tang, F., Chen, Z., Zhou, Z., Wu, W., Yang, Y., Liang, Y., Jiang, J., Cai, X., Su, J.: Polyp-mamba: Polyp segmentation with visual mamba. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 510–521. Springer (2024)
52. Yan, T., Gao, J., Xu, K., Zhu, X., Huang, H., Li, H., Wah, B., Lau, R.W.: Ghostingnet: A novel approach for glass surface detection with ghosting cues. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
53. Yang, J., Wang, Q., Zheng, F., Chen, P., Leonardis, A., Fan, D.P.: Plantcamo: Plant camouflage detection. arXiv preprint arXiv:2410.17598 (2024)
54. You, Z., Kong, L., Meng, L., Wu, Z.: Focus: Towards universal foreground segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
55. Zhang, M., Zhang, R., Yang, Y., Bai, H., Zhang, J., Guo, J.: Isnet: Shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 877–886 (2022)
56. Zhao, J., Li, X., Yang, F., Zhai, Q., Luo, A., Jiao, Z., Cheng, H.: Focusdiffuser: Perceiving local disparities for camouflaged object detection. In: European Conference on Computer Vision. pp. 181–198. Springer (2024)
57. Zhao, X., Pang, Y., Ji, W., Sheng, B., Zuo, J., Zhang, L., Lu, H.: Spider: a unified framework for context-dependent concept segmentation. In: Proceedings of the 41st International Conference on Machine Learning. pp. 60906–60926 (2024)
58. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4504–4513 (2022)
59. Zhou, M., Huang, J., Guo, C.L., Li, C.: Fourmer: An efficient global modeling paradigm for image restoration. In: International conference on machine learning. pp. 42589–42601. PMLR (2023)
60. Zhu, J., Zhu, X., Wang, W., Wang, X., Li, H., Wang, X., Dai, J.: Uni-perceiver-moe: Learning sparse generalist models with conditional moes. Advances in Neural Information Processing Systems **35**, 2664–2678 (2022)