

FCL-COD: Weakly Supervised Camouflaged Object Detection with Frequency-aware and Contrastive Learning

Jingchen Ni^{1*} Quan Zhang^{1*} Dan Jiang¹ Keyu Lv¹ Ke Zhang^{2†} Chun Yuan^{1†}

¹Tsinghua University ²Soochow University

njc24@mails.tsinghua.edu.cn, zhangqua22@tsinghua.org.cn,
kzhang19@suda.edu.cn, yuanc@sz.tsinghua.edu.cn

Abstract

Existing camouflage object detection (COD) methods typically rely on fully-supervised learning guided by mask annotations. However, obtaining mask annotations is time-consuming and labor-intensive. Compared to fully-supervised methods, existing weakly-supervised COD methods exhibit significantly poorer performance. Even for the Segment Anything Model (SAM), there are still challenges in handling weakly-supervised camouflage object detection (WSCOD), such as: a. non-camouflage target responses, b. local responses, c. extreme responses, and d. lack of refined boundary awareness, which leads to unsatisfactory results in camouflage scenes. To alleviate these issues, we propose a frequency-aware and contrastive learning-based WSCOD framework in this paper, named FCL-COD. To mitigate the problem of non-camouflaged object responses, we propose the Frequency-aware Low-rank Adaptation (FoRA) method, which incorporates frequency-aware camouflage scene knowledge into SAM. To overcome the challenges of local and extreme responses, we introduce a gradient-aware contrastive learning approach that effectively delineates precise foreground-background boundaries. Additionally, to address the lack of refined boundary perception, we present a multi-scale frequency-aware representation learning strategy that facilitates the modeling of more refined boundaries. We validate the effectiveness of our approach through extensive empirical experiments on three widely recognized COD benchmarks. The results confirm that our method surpasses both state-of-the-art weakly supervised and even fully supervised techniques.

1. Introduction

Camouflaged Object Detection (COD) aims to identify and segment objects concealed within their surrounding envi-

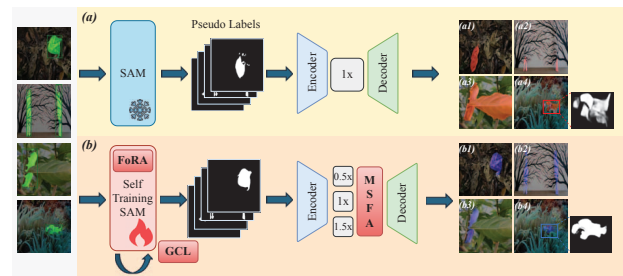


Figure 1. Comparison with the baseline, where (a) previous SAM-based method and (b) shows our proposed FCL-COD. By incorporating frequency awareness and contrastive learning, FCL-COD avoids the baseline’s drawbacks of a1) responses to non-camouflaged objects, a2) localized responses, a3) extreme responses, and a4) coarse boundaries.

ronment, attracting significant attention due to its potential applications in fields such as medical diagnosis[7], species conservation[19], and crop pest detection[28]. Unlike traditional object detection[8, 13], COD faces the challenge of high intrinsic similarity between the camouflaged object and its background, requiring the recognition of internal object information based on fine-grained details. Additionally, as a pixel-level classification task, COD demands precise boundary detection results.

In recent years, there has been an increasing amount of research on COD using data-driven deep learning techniques. Although fully supervised methods that rely on pixel-level annotations have made progress, they face inherent obstacles: manual pixel-level annotation of large-scale datasets is time-consuming and labor-intensive, and traditional methods treat each pixel in the target area equally, potentially failing to capture the essential structural features of the object[22]. To overcome these obstacles, sparse annotation methods have emerged to simplify dataset annotation and reduce overfitting. Exploring the use of sparse annotations as supervision in Weakly Supervised Camou-

*Equal contribution; † corresponding authors.

flaged Object Detection (WSCOD) has become a promising approach. Recent progress in weakly supervised, zero-shot, and multimodal localization under sparse supervision [42, 53, 54] further highlights the potential of learning reliable representations from limited annotations.

However, the limited annotation information available in WSCOD can severely hinder detection performance. While weakly supervised methods offer flexibility and reduce dataset annotation time and labor costs, they often fail to provide sufficient cues for accurate boundary inference. To address this issue, some studies [15] introduce consistency constraints, yet these losses are still computed on weakly represented image features and may therefore lead to imprecise object localization. Although boundary priors can partially improve contour quality, the cluttered backgrounds in camouflage scenes often introduce substantial non-target noise. As a result, existing WSCOD methods still suffer from non-camouflaged object responses, local responses, extreme responses, and insufficient boundary perception. Such challenges are also prevalent in broader weakly supervised and cross-modal settings [29, 44, 47, 56], motivating the design of more robust frequency-aware representations.

As a universal segmentation foundation model, SAM [20] has shown strong adaptability across downstream dense prediction tasks such as image restoration [52], image manipulation detection [55], and segmentation [41, 43]. Related advances in image manipulation detection and localization [5] further highlight the promise of foundation-model-based fine-grained visual understanding, making SAM a compelling backbone for weakly supervised scenarios. To address the issues present in previous Weakly Supervised Camouflaged Object Detection (WSCOD) methods, we propose a frequency-aware and contrastive learning-based weakly supervised camouflaged object detection framework, FCL-COD. To tackle the problem of non-camouflaged object responses, we introduce frequency-aware low-rank adaptation to inject camouflaged object scene knowledge into the pre-trained backbone model SAM. For the issues of local responses and extreme responses, we propose gradient-aware contrastive learning, which explores difficult background areas through a gradient-aware strategy and increases the representation distance between the foreground and background in the high-dimensional space. To address the lack of refined boundary perception, we propose a multi-scale frequency-aware attention module that combines multi-scale attention perception in both the frequency and spatial domains to uncover boundary-sensitive feature representations.

In summary, our contributions are as follows:

- We propose a frequency-aware and contrastive learning-based WSCOD method, which explores fine-grained object boundaries by mining high-dimensional frequency-domain differences, and uses contrastive learning to sep-

arate the object and background in the representation space.

- We introduce frequency-aware low-rank adaptation, which injects frequency-aware camouflaged object knowledge into SAM. Combined with gradient-aware contrastive learning, we mine easily confused background areas to push the object and background apart in the high-dimensional representation space.
- We propose a multi-scale frequency-aware attention module, which realizes boundary-sensitive representation learning through multi-scale interactions between the frequency and spatial domains.
- Extensive empirical experiments were conducted on four mainstream COD benchmarks. The results show that FCL-COD outperforms the state-of-the-art weakly supervised methods, and even fully supervised methods.

2. Related Work

2.1. Camouflaged Object Detection

Traditional camouflaged object detection (COD) methods rely on handcrafted features such as color [18], texture [9], and other visual cues. However, their performance degrades in complex scenes where camouflaged objects closely resemble the background. With the rise of deep learning [32–36, 46] and the establishment of benchmark datasets [6], deep learning-based COD approaches have gained prominence. Some methods [6] draw inspiration from biological perception, such as SINet, which mimics predators’ two-stage hunting strategy. Others employ multi-task frameworks that jointly learn COD and edge detection, using graphical models to capture task dependencies [51]. To overcome the limitations of RGB-only input, recent studies [39, 58] integrate auxiliary cues like frequency-domain features to better distinguish subtle foreground–background differences. Despite the advances of fully supervised models, their dependence on dense pixel-level annotations often overlooks holistic target structures, motivating the development of weakly supervised COD methods that exploit sparse supervision for more efficient learning.

2.2. Contrastive Learning

Recently, contrastive learning has become a dominant self-supervised paradigm for learning discriminative representations from unlabeled data [1]. By pulling similar samples together and pushing dissimilar ones apart, it effectively enhances feature separability—an essential property for camouflaged object detection. Representative methods include SimCLR [3], which employs data augmentation and transformation prediction to learn invariant representations; MoCo [14], which introduces a momentum-updated encoder and dynamic queue to improve negative sampling; and SwAV [1], which integrates clustering with contrastive

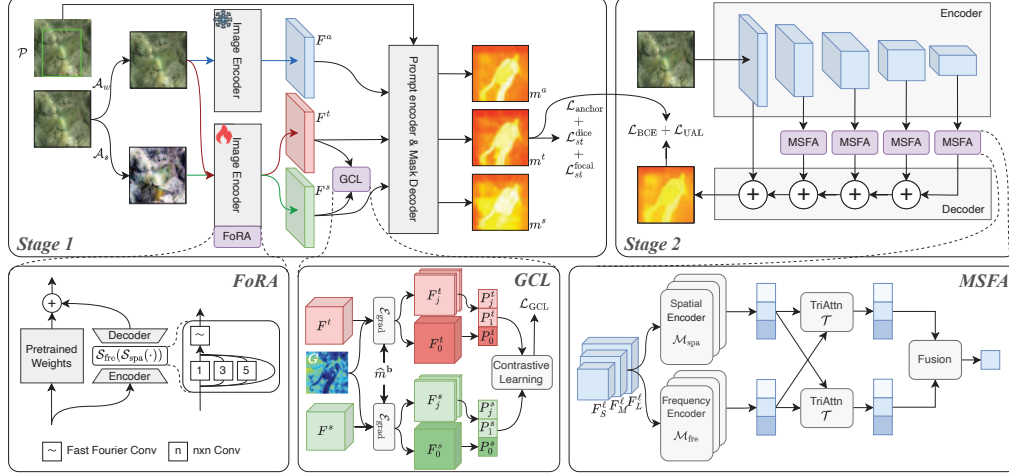


Figure 2. **FCL-COD Pipeline Overview.** Within a triadic teacher–student architecture, frequency-aware low-rank adaptation and gradient-aware contrastive learning jointly suppress the localized and extreme activations that plague earlier approaches. In the subsequent re-training phase, multi-scale frequency-aware attention is injected to excavate crisper, more delicate camouflaged boundaries.

objectives for fine-grained semantic learning. Moreover, InfoNCE [37] formalizes contrastive learning via mutual information maximization, reinforcing affinity preservation in unsupervised representation learning. These advances underscore contrastive learning’s strength in boosting generalization and feature discriminability. Nevertheless, its application to COD remains challenging due to the difficulty of defining positive/negative pairs and efficiently mining hard negatives—issues that merit further exploration.

3. Method

Our proposed **FCL-COD** employs a two-stage framework. In the first stage, SAM is adapted to produce high-quality pseudo-labels for camouflaged objects via *Triadic Teacher–Student Self-training* (§3.1), enhanced by *Frequency-aware Low-Rank Adaptation* (FoRA, §3.2) and *Gradient-aware Contrastive Learning* (GCL, §3.3). In the second stage, a lightweight encoder–decoder detector with *Multi-Scale Frequency-aware Attention* (MSFA, §3.4) is trained using these pseudo-labels. This design achieves an optimal trade-off between accuracy and efficiency: SAM offers robust priors for pseudo-labeling, while the MSFA-equipped lightweight detector enables real-time inference with boundary-sensitive features.

3.1. Triadic Teacher-Student Self-training

To obtain reliable pseudo-labels, three encoders are maintained: an anchor encoder $f^a(x; \Theta^a)$, a student encoder $f^s(x; \Theta^s)$, and a teacher encoder $f^t(x; \Theta^t)$, where $\Theta^s = \Theta^t$. As shown in Fig. 2, for each sample x_i , weak augmentation $\mathcal{A}_w$ is applied to anchor and teacher inputs, and

strong augmentation $\mathcal{A}_s$ to the student input. The encoders yield feature maps F^a, F^s , and $F^t \in \mathbb{R}^{D \times H \times W}$. Given a bounding box prompt $\mathcal{P}$, each branch produces N_p masks $y_j^a, y_j^s, y_j^t, j = 1, \dots, N_p$, which are normalized via sigmoid to $m_j \in [0, 1]^{H \times W}$ and binarized as $\hat{m}_j = \mathbb{1}(m_j > 0.5) \in \{0, 1\}^{H \times W}$.

The triadic self-training loss combines two terms. The *student–teacher loss* employs Focal Loss and Dice Loss to guide the student with the teacher’s pseudo-labels, where γ emphasizes hard pixels and ϵ prevents division by zero. The Focal Loss is formulated in Eq. (1).

$$\mathcal{L}_{st}^{\text{focal}} = -\frac{1}{HW} \sum_{j=1}^{N_p} \sum_{h,w} \mathbb{1}(\hat{m}_{jhw}^t = 1) (1 - m_{jhw}^s)^\gamma \log(m_{jhw}^s) + \mathbb{1}(\hat{m}_{jhw}^t = 0) (m_{jhw}^s)^\gamma \log(1 - m_{jhw}^s) \quad (1)$$

The Dice Loss is given by Eq. (2).

$$\mathcal{L}_{st}^{\text{dice}} = \sum_{j=1}^{N_p} 1 - \frac{2 \sum_{h,w} m_{jhw}^s \cdot \hat{m}_{jhw}^t + \epsilon}{\sum_{h,w} m_{jhw}^s + \sum_{h,w} \hat{m}_{jhw}^t + \epsilon} \quad (2)$$

Second, to mitigate error accumulation caused by imperfect teacher pseudo-labels, we introduce an *anchor loss* for regularization. A frozen anchor network preserves the original SAM knowledge and constrains both the student and teacher, preventing excessive deviation from the pre-trained model. The anchor loss is formulated in Eq. (3).

$$\mathcal{L}_{\text{anchor}} = \lambda_{\text{stu}}^{\text{dice}} \mathcal{L}_{\text{stu}}^{\text{dice}}(m^s, \hat{m}^a) + \lambda_{\text{tea}}^{\text{dice}} \mathcal{L}_{\text{tea}}^{\text{dice}}(m^t, \hat{m}^a) \quad (3)$$

3.2. Frequency-aware Low-Rank Adaptation

Low-Rank Adaptation (LoRA) constrains parameter updates to a low-dimensional subspace via an encoder-decoder parameterization. While freezing the pre-trained parameters, it injects a pair of lightweight, trainable rank-decomposition matrices into each Transformer layer. Given a linear projection weight $W_0 \in \mathbb{R}^{b \times a}$, LoRA introduces two trainable matrices $W_e \in \mathbb{R}^{r \times a}$ and $W_d \in \mathbb{R}^{b \times r}$, where $r \ll \min(a, b)$. The forward propagation is given by Eq. (4).

$$h = W_0 x + W_d W_e x \quad (4)$$

The low-rank branch $W_d W_e x$ serves as a residual update to the frozen projection $W_0 x$.

However, in camouflaged-object scenarios, where boundaries are ambiguous and transparent targets often overlap with cluttered backgrounds, conventional low-rank adaptation struggles to inject camouflage-specific priors into SAM, leading to weak discrimination and spurious activations in non-camouflaged regions. To address this, we propose Frequency-aware Low-Rank Adaptation (FoRA), which extends LoRA by inserting a two-stage transformation between the encoder and decoder to enrich the encoded features in both spatial and frequency domains.

Formally, the FoRA-enhanced forward propagation is defined in Eq. (5).

$$h = W_0 x + W_d \mathcal{S}_{\text{fre}}(\mathcal{S}_{\text{spa}}(W_e x)) \quad (5)$$

The spatial operator precedes the frequency operator within the low-rank pathway. Here, $\mathcal{S}_{\text{spa}}(\cdot)$ denotes the spatial enhancement stage and $\mathcal{S}_{\text{fre}}(\cdot)$ denotes the frequency modulation stage. The spatial enhancement stage captures multi-scale contextual dependencies via convolutions with different receptive fields. Given the encoded feature $F_e = W_e x$, we define the spatial enhancement in Eq. (6), which aggregates multi-scale responses and adds a residual connection.

$$\mathcal{S}_{\text{spa}}(F_e) = \frac{\Phi_{1 \times 1}(F_e) + \Phi_{3 \times 3}(F_e) + \Phi_{5 \times 5}(F_e)}{3} + F_e \quad (6)$$

Subsequently, the frequency modulation stage applies the Fourier transform, performs convolution in the frequency domain, and reconstructs the representation via the inverse Fourier transform. The transformation is given in Eq. (7).

$$\mathcal{S}_{\text{fre}}(F) = \text{IFFT}(\Phi_{3 \times 3}(\text{FFT}(F))) \quad (7)$$

By explicitly modeling spatial and frequency cues in a cascaded manner as shown in Eq. (5), FoRA augments the base projection with spatial-frequency enriched low-rank updates and more effectively injects camouflage-specific priors while preserving the generalization ability of the foundation model.

3.3. Gradient-aware Contrastive Learning

Camouflaged scenes often contain ambiguous boundaries between foreground and background, which can lead to partial detections and a high rate of false positives. While the losses in the triadic teacher-student framework operate at the output level, they are insufficient to enforce separability in the feature space. To further enhance the distinction between foreground and background representations, we introduce a Gradient-aware Contrastive Learning (GCL) objective.

During contrastive learning, the quality of sampled features plays a crucial role. To emphasize ambiguous background regions that are easily confused with the foreground, we derive a gradient activation map G^t via Grad-CAM from the teacher feature map F^t , which provides more stable and reliable guidance than the student. The gradient activation map is computed as shown in Eq. (8), and we construct a gradient-weighted background mask according to Eq. (9):

$$G^t = \Phi_{\text{GC}}(F^t), \quad G^t \in [0, 1]^{H \times W} \quad (8)$$

$$\tilde{m}_0 = \hat{m}_0 \odot G^t \quad (9)$$

where $\odot$ denotes element-wise multiplication, $\hat{m}_0$ is the binary background mask, and $\tilde{m}_0$ is its gradient-weighted counterpart.

We then compute branch-specific instance prototypes via masked average pooling. For each branch $\mathbf{b} \in \{s, t\}$, the background prototype and the j -th foreground prototype are defined as. Here, the superscripts s and t indicate the student and teacher branches, respectively; the corresponding feature maps are F^s and F^t , and the instance prototypes are I_j^s and I_j^t .

$$I_0^{\mathbf{b}} = \frac{\sum_{h,w} F_{hw}^{\mathbf{b}} \cdot \tilde{m}_{0hw}}{\sum_{h,w} \tilde{m}_{0hw}}, \quad \mathbf{b} \in \{s, t\} \quad (10)$$

$$I_j^{\mathbf{b}} = \frac{\sum_{h,w} F_{hw}^{\mathbf{b}} \cdot \hat{m}_{jhw}}{\sum_{h,w} \hat{m}_{jhw}}, \quad \mathbf{b} \in \{s, t\}, j = 1, \dots, N \quad (11)$$

Here, N denotes the number of foreground instances in the current image (variable across samples), and index 0 refers to the background prototype.

Here, $F^{\mathbf{b}}$ denotes the L2-normalized feature map, i.e., $F_{hw}^{\mathbf{b}} = F_{hw}^{\mathbf{b}} / \|F_{hw}^{\mathbf{b}}\|_2$. In Eq. (12), positive pairs are constructed between I_j^s and I_j^t , while negatives include other foreground instances as well as the gradient-aware background prototype I_0^t . Weighting the background with $\tilde{m}_0$ (Eqs. (9)–(10)) encourages the model to focus on Grad-CAM-highlighted background regions that are most prone to confusion with the foreground. The contrastive objective is formulated as:

$$\mathcal{L}_{\text{GCL}} = -\log \frac{\sum_{j=1}^N \exp(I_j^s \cdot I_j^t / \tau)}{\sum_{j=1}^N \sum_{k=0, k \neq j}^N \exp(I_j^s \cdot I_k^t / \tau)} \quad (12)$$

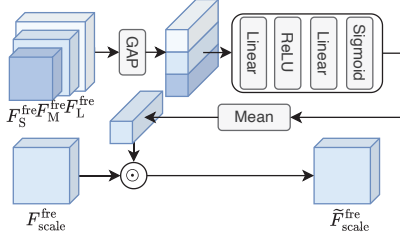


Figure 3. Detailed structure of Tri-Channel Attention mechanism in MSFA

where τ is the temperature hyperparameter.

Finally, the overall training objective integrates the gradient-aware contrastive loss with the teacher–student self-training and anchor losses, as defined in Eq. (13).

$$\mathcal{L} = \mathcal{L}_{st}^{dice} + \lambda_1 \mathcal{L}_{anchor} + \lambda_2 \mathcal{L}_{GCL} + \lambda_3 \mathcal{L}_{st}^{focal} \quad (13)$$

By incorporating the gradient-aware contrastive term, the model achieves better separation between foreground and background features in the embedding space, leading to more robust discrimination of camouflaged targets.

3.4. Multi-Scale Frequency-aware Attention

The high-quality pseudo-labels generated in the first stage enable efficient training of a lightweight detector. However, camouflaged objects exhibit ambiguous boundaries that require fine-grained boundary-sensitive features. To address this, we propose MSFA as the second-stage module inserted between the encoder and decoder. For each encoder layer ℓ , we extract multi-scale features at three different resolutions: small-scale F_S^ℓ , medium-scale F_M^ℓ , and large-scale F_L^ℓ , which capture fine details, intermediate structures, and global context, respectively. These three-scale features are processed by MSFA before decoding.

MSFA adopts a dual-branch design: the spatial branch $\mathcal{M}_{spa}$ enhances local context via stacked 3×3 convolutions, while the frequency branch $\mathcal{M}_{fre}$ models spectral cues in the Fourier domain, as defined in Eq. (14).

$$\begin{aligned} \mathcal{M}_{spa}(F) &= \Phi_{3 \times 3}(\Phi_{3 \times 3}(F)) \\ \mathcal{M}_{fre}(F) &= \text{IFFT}(\Phi_{1 \times 1}([\Re(\text{FFT}(F)), \Im(\text{FFT}(F))])) \end{aligned} \quad (14)$$

Here, $\Phi_{k \times k}$ denotes a $k \times k$ convolution, and $\Re(\cdot)$, $\Im(\cdot)$ denote the real and imaginary parts of the FFT output, respectively.

To enable cross-domain interaction, we employ a Tri-Channel Attention mechanism, denoted as $\mathcal{T}$, to gate each branch with multi-scale context from the other domain. For each scale, we first apply the two branches to obtain $F_{scale}^{spa} = \mathcal{M}_{spa}(F_{scale}^\ell)$ and $F_{scale}^{fre} = \mathcal{M}_{fre}(F_{scale}^\ell)$. The gating operator

$\mathcal{T}$ is defined in Eq. (15).

$$\mathcal{T}_i(x | \{y_1, y_2, y_3\}) = x \odot \frac{1}{3} \sum_{j=1}^3 \sigma(W_{i,j}^{(2)} \cdot \text{ReLU}(W_{i,j}^{(1)} \cdot \text{GAP}(y_j))) \quad (15)$$

where $\text{GAP}(\cdot)$ denotes global average pooling, $W_{i,j}^{(1)}$ and $W_{i,j}^{(2)}$ are scale-specific learnable weight matrices for channel reduction and expansion, σ is the sigmoid activation, and the summation averages attention weights from three scales. Applying Eq. (15), we obtain the gated features as shown in Eq. (16).

$$\begin{aligned} \tilde{F}_{scale}^{fre} &= \mathcal{T}_{scale}(F_{scale}^{fre} | \{F_S^{spa}, F_M^{spa}, F_L^{spa}\}) \\ \tilde{F}_{scale}^{spa} &= \mathcal{T}_{scale}(F_{scale}^{spa} | \{F_S^{fre}, F_M^{fre}, F_L^{fre}\}) \end{aligned} \quad (16)$$

Here, the first line gates frequency features with spatial context, while the second line gates spatial features with frequency context, for each scale $\in \{S, M, L\}$. Finally, the gated features from both branches are aggregated and fused as shown in Eq. (17).

$$F_{MSFA}^\ell = \Phi_{1 \times 1}^{spa}([\tilde{F}_S^{spa}, \tilde{F}_M^{spa}, \tilde{F}_L^{spa}]) + \Phi_{1 \times 1}^{fre}([\tilde{F}_S^{fre}, \tilde{F}_M^{fre}, \tilde{F}_L^{fre}]) \quad (17)$$

where $[\cdot]$ denotes channel-wise concatenation, and $\Phi_{1 \times 1}^{spa}$ and $\Phi_{1 \times 1}^{fre}$ are branch-specific projection convolutions. This design enables the model to capture boundary-sensitive representations through multi-scale spatial–frequency interaction.

Training Objective. In the second stage, we train the MSFA-enhanced detector using the pseudo-labels generated from the first stage. Given the predicted probability map p and the pseudo-label mask $\hat{m}$, the training objective combines binary cross-entropy loss and uncertainty-aware loss, defined in Eq. (18).

$$\begin{aligned} \mathcal{L}_{stage2} &= \mathcal{L}_{BCE}(p, \hat{m}) + \alpha(t) \cdot \mathcal{L}_{UAL}(p) \\ \mathcal{L}_{UAL}(p) &= (1 - |2p - 1|^2) \end{aligned} \quad (18)$$

where $\mathcal{L}_{BCE}$ is the binary cross-entropy loss, the uncertainty-aware loss $\mathcal{L}_{UAL}$ penalizes uncertain predictions, and $\alpha(t) = \cos(\pi t/2)$ is a cosine-annealed coefficient with $t \in [0, 1]$ denoting the training progress. This design encourages the model to produce confident predictions while leveraging the high-quality pseudo-labels from the first stage.

4. Experiment

4.1. Experimental Setup

Dataset: We conducted experiments on four widely used COD benchmarks: CAMO [21], CHAMELEON [38], COD10K [6], and NC4K [26]. CAMO includes 1,250 camouflaged and 1,250 non-camouflaged images, while CHAMELEON provides 76 finely annotated samples.

Table 1. Comprehensive evaluation of FCL-COD against competing methods across four benchmarks. **Sup.**: F = fully supervised, S = scribble, B = bounding box, P = point, - = zero-shot/no task-specific training.

Methods	Sup.	CAMO				COD10K				NC4K				CHAMELEON			
		MAE↓	S_m ↑	E_m ↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	F_β^w ↑
UGTR[48]	F	0.086	0.784	0.822	0.684	0.036	0.817	0.852	0.666	0.052	0.839	0.874	0.747	0.030	0.891	0.955	0.833
ZoomNet[30]		0.066	0.820	0.892	0.752	0.029	0.838	0.911	0.729	0.043	0.853	0.896	0.784	0.023	0.902	0.958	0.845
SAM-Adapter[4]		0.070	0.847	0.873	0.765	0.025	0.883	0.918	0.801	-	-	-	-	0.033	0.896	0.919	0.824
CamoFormer-R[49]		0.067	0.817	0.885	0.752	0.029	0.838	0.930	0.724	0.042	0.855	0.914	0.788	0.025	0.898	0.956	0.847
FEDER[11]		0.071	0.802	0.873	0.738	0.032	0.822	0.905	0.716	0.044	0.847	0.915	0.789	0.030	0.887	0.954	0.835
CamoFormer-P[49]		0.046	0.872	0.938	0.831	0.023	0.869	0.939	0.786	0.030	0.892	0.946	0.847	0.022	0.910	0.966	0.865
MSCAF-Net[25]		0.046	0.873	0.937	0.828	0.024	0.865	0.936	0.775	0.032	0.887	0.942	0.838	0.022	0.912	0.970	0.865
HitNet[16]		0.055	0.849	0.910	0.809	0.023	0.871	0.938	0.806	0.037	0.875	0.929	0.834	0.019	0.921	0.972	0.897
FSPNet[17]		0.050	0.856	0.928	0.799	0.026	0.851	0.930	0.735	0.035	0.878	0.937	0.816	0.023	0.908	0.965	0.851
SARNet[45]		0.046	0.874	0.935	0.844	0.021	0.885	0.947	0.820	0.032	0.889	0.940	0.851	0.017	0.933	0.978	0.909
SAM[20] (SAM-H)	-	0.132	0.684	0.687	0.606	0.050	0.783	0.798	0.701	0.078	0.767	0.776	0.696	0.081	0.727	0.734	0.639
SCSOD[50]	S	0.102	0.713	0.795	0.618	0.055	0.710	0.805	0.546	-	-	-	-	0.053	0.792	0.881	0.714
CRNet[15]		0.092	0.735	0.815	0.641	0.049	0.733	0.832	0.576	0.063	0.775	0.855	0.688	0.046	0.818	0.897	0.791
SAM-S[20] (SAM-H)		0.105	0.731	0.774	-	0.046	0.772	0.828	-	0.071	0.763	0.832	-	0.076	0.650	0.820	0.729
WS-SAM[12] (SAM-H)		0.092	0.759	0.818	-	0.038	0.803	0.878	-	0.052	0.829	0.886	-	0.046	0.824	0.897	0.777
SAM-P[20] (SAM-H)	P	0.123	0.677	0.693	-	0.069	0.765	0.796	-	0.082	0.776	0.786	-	0.101	0.697	0.745	0.696
WS-SAM[12] (SAM-H)		0.102	0.718	0.757	-	0.039	0.790	0.856	-	0.057	0.813	0.859	-	0.056	0.805	0.868	0.767
SAM-COD[2](SAM-H)	B	0.062	0.837	0.901	0.786	0.028	0.842	0.914	0.745	0.037	0.867	0.923	0.813	-	-	-	-
FCL-COD(SAM-B)		0.060	0.841	0.899	0.795	0.027	0.859	0.924	0.774	0.041	0.867	0.919	0.817	0.039	0.866	0.928	0.799
FCL-COD(SAM-L)		0.054	0.856	0.910	0.818	0.022	0.881	0.938	0.812	0.034	0.886	0.930	0.847	0.026	0.901	0.954	0.856
FCL-COD(SAM-H)		0.050	0.862	0.915	0.824	0.022	0.878	0.934	0.808	0.033	0.885	0.928	0.846	0.038	0.882	0.932	0.842

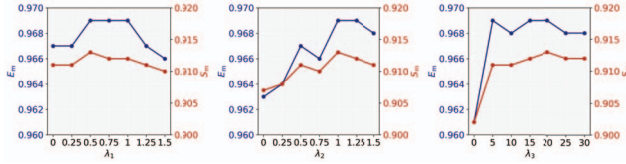


Figure 4. Hyperparameter analysis of loss-function weights.

COD10K is a large-scale dataset with 5,066 camouflaged, 3,000 background, and 1,934 non-camouflaged images, offering high diversity. NC4K contains 4,121 images collected from multiple online sources. Following [27], we use only camouflaged images. Specifically, 3,040 images from COD10K and 1,000 from CAMO form the training set, while the remaining camouflaged images from all four datasets serve as the test set.

Implementation Details: Our method is implemented in PyTorch on two NVIDIA H20 GPUs. Stage 1 adapts SAM via FoRA for pseudo-label generation; Stage 2 uses PVT-B4 as the lightweight encoder-decoder backbone. Training uses SGD with momentum 0.9, weight decay 5×10^{-4} , and cosine annealing from 1×10^{-3} over 60 epochs with batch size 8. Bounding-box prompts are derived from ground-truth mask bounding boxes; no pixel-level annotations are used. At inference, Stage 2 operates without prompt input.

4.2. Comparison with State-of-the-art Methods

Quantitative Comparison: As demonstrated in Tab. 1, our method delivers substantial improvements over existing approaches. When compared to the state-of-the-art weakly-

supervised COD method, SAM-COD, our method consistently outperforms across all evaluation metrics. Specifically, on the CAMO dataset, our approach results in a reduction of the MAE by 0.012, alongside improvements in S_m , E_m , and F_β^w by 0.025, 0.014, and 0.038, respectively. These performance gains are not only observed in CAMO but also extend consistently across the remaining three datasets. Furthermore, when contrasted with fully-supervised methods such as ZoomNet and CamoFormer, our method achieves noticeable performance advancements across all four datasets.

Qualitative Comparison: As illustrated in Fig. 5, the predicted maps produced by our method demonstrate clearer, more coherent object regions with more defined contours. These results surpass the performance of the state-of-the-art weakly-supervised COD method, SAM-COD, and the fully-supervised COD method, ZoomNet. Our approach effectively addresses limitations observed in prior methods, including the presence of non-camouflaged target responses, extreme or partial responses, and rough boundary delineations.

4.3. Ablation Study

Ablation experiment of components. We conducted a progressive ablation study to evaluate each component’s contribution, as shown in Tab. 2. The pseudo-label quality on **COD-train** steadily improves with the integration of **FoRA** and **GCL**, boosting the E_m from 0.959 to 0.969. This enhancement directly benefits the final lightweight model, where the full system with **MSFA** achieves the best results

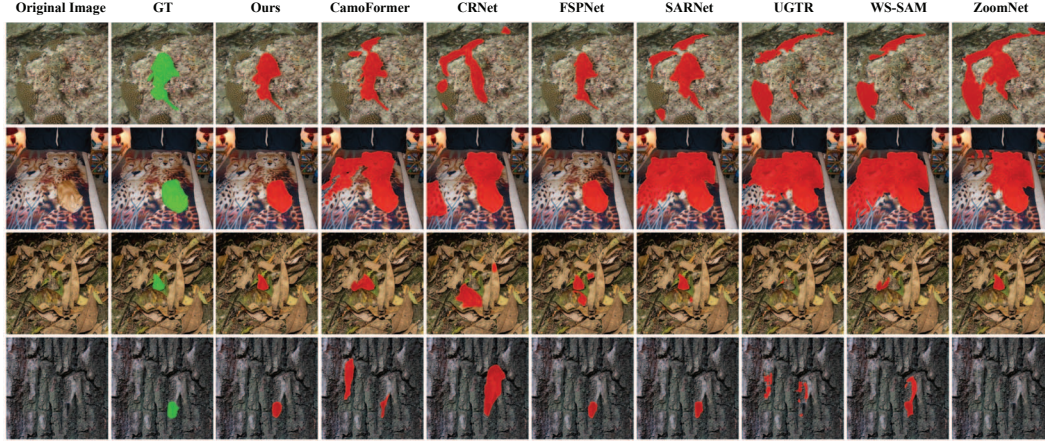


Figure 5. Qualitative Comparison between FCL-COD and Competing Methods

Table 2. Ablation study of Individual Components

FoRA	GCL	MSFA	COD-Train						CHAMELEON						COD10K					
			MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑	MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑	MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑
×	×	×	0.017	0.900	0.959	0.865	0.799	0.874	0.041	0.864	0.927	0.801	0.745	0.826	0.025	0.856	0.919	0.766	0.702	0.794
✓	×	×	0.015	0.907	0.963	0.878	0.812	0.884	0.041	0.868	0.928	0.809	0.753	0.833	0.024	0.860	0.923	0.775	0.709	0.800
✓	✓	×	0.013	0.913	0.969	0.887	0.820	0.892	0.029	0.888	0.947	0.839	0.778	0.857	0.024	0.863	0.926	0.782	0.716	0.807
✓	✓	✓	-	-	-	-	-	-	0.026	0.901	0.954	0.856	0.801	0.873	0.022	0.881	0.938	0.812	0.750	0.836

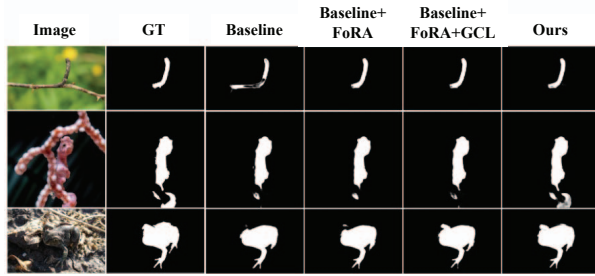


Figure 6. Qualitative ablation study of each component.

(E_m =0.938 on **COD10K**, 0.954 on **CHAMELEON**), validating the synergy of all modules.

Hyperparameter Analysis. We analyze the loss-weight parameters in Fig. 4. Optimal values are λ_1 =0.50, λ_2 =1.00, and λ_3 =20.

Ablation Study of FoRA. As shown in Tab. 3, the baseline attains an E_m of 0.965. Incorporating either the frequency modulation stage S_{fre} or the spatial enhancement stage S_{spa} improves performance to 0.967 and 0.966, respectively. The complete FoRA, combining both stages, achieves the highest E_m of 0.969, validating the complementary contributions of spatial and frequency cues to fine-grained segmentation.

Ablation Study of GCL. To assess Gradient-aware Contrastive Learning (GCL), we conduct a controlled ablation in Tab. 3. The baseline (**noCL**) yields an E_m of 0.963, while standard **CL** increases it to 0.968. Our **GCL** lever-

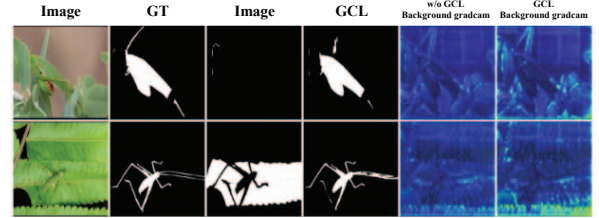


Figure 7. Qualitative analysis of GCL.

ages gradient cues to highlight hard samples, further improving performance to 0.969.

Ablation Study of MSFA. Table 4 shows that the **Base** model reaches E_m scores of 0.926 (COD10K) and 0.944 (CHAMELEON). Introducing the single-scale Tri-Channel Attention $\mathcal{T}$ provides moderate gains, while adding the spatial branch $\mathcal{M}_{spa}$ and frequency branch $\mathcal{M}_{fre}$ yields further improvements. The full multi-scale design delivers the largest boost, with MSFA (**Ours**) achieving 0.938 and 0.954, confirming that multi-scale fusion drives the majority of the performance gains, complemented by dual-branch attention.

Qualitative Analysis for Ablation Study. Figure 6 illustrates the progressive gains from each component. The **Baseline** produces coarse and incomplete structures. Incorporating **FoRA** improves core-region localization, yielding more complete predictions. Adding **GCL** further enhances object-background separation and contour refine-

Table 3. Ablation study on FoRA and GCL

FoRA		COD-Train					
S_{spa}	S_{fre}	MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑
×	×	0.014	0.91	0.965	0.881	0.814	0.885
×	✓	0.014	0.91	0.967	0.883	0.816	0.888
✓	×	0.014	0.91	0.966	0.882	0.817	0.888
✓	✓	0.013	0.913	0.969	0.887	0.82	0.892
GCL		COD-Train					
CL	Grad-aware	MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑
×	×	0.015	0.907	0.963	0.878	0.812	0.884
✓	×	0.014	0.912	0.968	0.883	0.817	0.89
✓	✓	0.013	0.913	0.969	0.887	0.82	0.892

Table 4. Ablation Study on MSFA

MSFA			COD10K						CHAMELEON					
$\mathcal{T}$	$\mathcal{M}_{spa}$	$\mathcal{M}_{fre}$	MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑	MAE↓	S_m ↑	E_m ↑	F_{β}^w ↑	mIOU↑	mF1↑
×	×	×	0.024	0.863	0.926	0.782	0.716	0.807	0.03	0.886	0.944	0.835	0.773	0.85
✓	×	×	0.024	0.865	0.927	0.785	0.718	0.809	0.03	0.887	0.945	0.836	0.774	0.852
✓	✓	×	0.022	0.88	0.936	0.811	0.749	0.834	0.027	0.898	0.947	0.852	0.796	0.869
✓	×	✓	0.022	0.877	0.934	0.807	0.743	0.829	0.027	0.895	0.95	0.85	0.792	0.865
✓	✓	✓	0.022	0.881	0.938	0.812	0.75	0.836	0.026	0.901	0.954	0.856	0.801	0.873

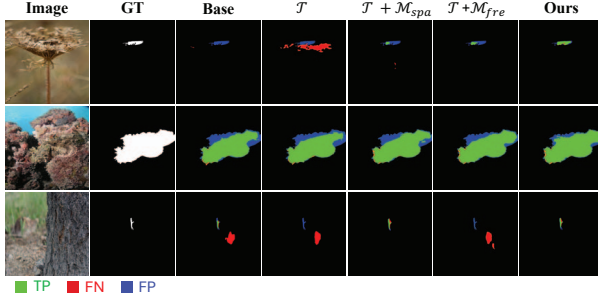


Figure 8. Qualitative analysis of MSFA.

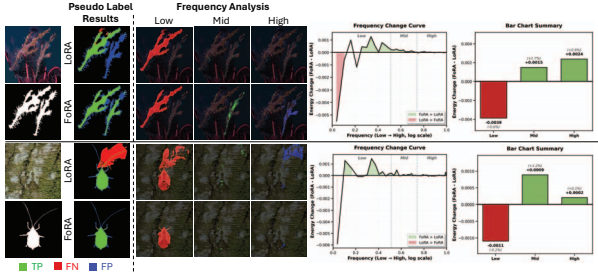


Figure 9. Qualitative analysis of frequency awareness achieved by FoRA.

ment. With MSFA integrated, the **Full Model** achieves accurate segmentation with well-defined boundaries through multi-scale feature fusion. These results align with quantitative trends, validating each module’s contribution.

Effectiveness of the GCL module with Grad-CAM.

Fig. 7 demonstrates the impact of GCL. Grad-CAM results show that **GCL** effectively guides attention to background regions that resemble the foreground—critical hard negatives for contrastive learning—whereas **noGCL** suffers from dispersed focus. This targeted hard-sample awareness leads to more accurate and robust segmentation.

Qualitative Analysis of MSFA. As shown in Fig. 8, the Base model suffers from localized responses (row 1 & 3) and coarse boundaries. Progressively integrating spatial and frequency branches reduces false positives, while the full MSFA achieves complete object coverage with refined boundaries through multi-scale spatial–frequency fusion.

Qualitative Analysis of FoRA. Fig. 9 demonstrates that

Table 5. Extended Experiments of FCL-COD on Weakly-Supervised Salient Object Detection

	Label	ECSSD		DUT-O		HKU-IS		DUTS-TE	
		MAE↓	S_m ↑	MAE↓	S_m ↑	MAE↓	S_m ↑	MAE↓	S_m ↑
AFNet[24]	F	0.042	0.913	0.057	0.826	0.036	0.905	0.046	0.867
BASNet[31]	F	0.037	0.916	0.057	0.836	0.032	0.909	0.048	0.866
GateNet[57]	F	0.040	0.920	0.055	0.838	0.033	0.915	0.040	0.885
ICON-R[59]	F	0.032	0.928	0.057	0.845	0.029	0.920	0.037	0.890
MENet[40]	F	0.031	0.928	0.045	0.850	0.023	0.927	0.028	0.905
VST-S++ [23]	F	0.027	0.939	0.050	0.859	0.025	0.932	0.029	0.909
SCSOD[50]	S	0.049	0.881	0.060	0.811	0.038	0.882	0.049	0.853
PSOD[10]	P	0.036	0.913	0.064	0.824	0.033	0.901	0.045	0.853
SAM-COD[2]	B	0.031	0.929	0.051	0.844	0.023	0.952	0.033	0.899
FCL-COD	B	0.023	0.945	0.043	0.872	0.020	0.942	0.030	0.907

vanilla LoRA produces extreme responses (excessive false negatives in rows 1-3) and non-camouflaged responses. In contrast, FoRA substantially mitigates these issues by suppressing low-frequency texture interference while preserving discriminative mid–high frequency details, as evidenced by the feature frequency amplitude spectra, which show notable energy redistribution toward mid-to-high frequency components.

Generalization to salient object detection. Tab. 5 further shows that our framework generalizes well to salient object detection (SOD). Enhanced frequency perception and contrastive learning strengthen SAM’s adaptability to complex scenes, highlighting the versatility of the proposed weakly supervised paradigm.

5. Conclusion

We propose **FCL-COD**, a frequency-aware and contrastive learning-based framework that addresses key limitations in weakly supervised camouflaged object detection: non-camouflaged responses, localized and extreme responses, and coarse boundaries. Through frequency-aware low-rank adaptation (FoRA), gradient-aware contrastive learning (GCL), and multi-scale frequency-aware attention (MSFA), FCL-COD effectively adapts SAM to camouflage scenarios with sparse annotations. Extensive experiments demonstrate that our method surpasses existing weakly supervised approaches and rivals fully supervised methods, highlighting the potential of frequency-domain modeling and contrastive learning for challenging perception tasks.

Acknowledgement

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008).

References

- [1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2
- [2] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod+: Sam-guided unified framework for weakly-supervised camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6, 8
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2
- [4] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?—sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023. 6
- [5] Yirui Chen, Xudong Huang, Quan Zhang, Wei Li, Mingjian Zhu, Qiangyu Yan, Simiao Li, Hanting Chen, Hailin Hu, Jie Yang, et al. Gim: A million-scale benchmark for generative image manipulation detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2311–2319, 2025. 2
- [6] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 2, 5
- [7] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE transactions on medical imaging*, 39(8):2626–2637, 2020. 1
- [8] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2344–2366, 2022. 1
- [9] Galun, Sharon, Basri, and Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *Proceedings Ninth IEEE international conference on computer vision*, pages 716–723. IEEE, 2003. 2
- [10] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervision. In *Proceedings of the AAAI conference on artificial intelligence*, pages 670–678, 2022. 8
- [11] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22046–22055, 2023. 6
- [12] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36, 2024. 6
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2
- [15] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 781–789, 2023. 2, 6
- [16] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 881–889, 2023. 6
- [17] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5557–5566, 2023. 6
- [18] Iván Huerta, Daniel Rowe, Mikhail Mozerov, and Jordi González. Improving background subtraction based on a causality of colour-motion segmentation problems. In *Iberian conference on pattern recognition and image analysis*, pages 475–482. Springer, 2007. 2
- [19] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. 1
- [20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 6
- [21] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabran network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019. 5
- [22] Liu Liu, Rujing Wang, Chengjun Xie, Po Yang, Fangyuan Wang, Sud Sudirman, and Wancai Liu. Pestnet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification. *Ieee Access*, 7:45301–45312, 2019. 1
- [23] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE*

- Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7300–7316, 2024. 8
- [24] Rui Liu, Li Mi, and Zhenzhong Chen. Afnet: Adaptive fusion network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7871–7886, 2020. 8
- [25] Yu Liu, Haihang Li, Juan Cheng, and Xun Chen. Mscaf-net: A general framework for camouflaged object detection via learning multi-scale context-aware features. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4934–4947, 2023. 6
- [26] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11601, 2021. 5
- [27] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021. 6
- [28] Melia G Nafus, Jennifer M Germano, Jeanette A Perry, Brian D Todd, Allyson Walsh, and Ronald R Swaisgood. Hiding in plain sight: a study on camouflage and habitat selection in a slow-moving desert herbivore. *Behavioral Ecology*, 26(5):1389–1394, 2015. 1
- [29] Jingchen Ni, Keyu Lyu, Yu Guo, and Chun Yuan. Semantic alignment and hard sample retraining for visible-infrared person re-identification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025. 2
- [30] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2160–2170, 2022. 6
- [31] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489, 2019. 8
- [32] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. In *Proc. VLDB Endow.*, pages 2363–2377, 2024. 2
- [33] Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, Jilin Hu, Christian S. Jensen, and Bin Yang. TAB: Unified benchmarking of time series anomaly detection methods. In *Proc. VLDB Endow.*, pages 2775–2789, 2025.
- [34] Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. Dbloss: Decomposition-based loss function for time series forecasting. In *NeurIPS*, 2025.
- [35] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pages 1185–1196, 2025.
- [36] Xiangfei Qiu, Yuhao Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. Dag: A dual correlation network for time series forecasting with exogenous variables. *arXiv preprint arXiv:2509.14933*, 2025. 2
- [37] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143*, 2024. 3
- [38] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 5
- [39] Qingwei Wang, Jinyu Yang, Xiaosheng Yu, Fangyi Wang, Peng Chen, and Feng Zheng. Depth-aided camouflaged object detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3297–3306, 2023. 2
- [40] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10031–10040, 2023. 8
- [41] Yuji Wang, Ruojun Zhao, Shicai Wei, Jingchen Ni, Meng Wu, Yang Luo, and Chunbo Luo. Convolution meets transformer: Efficient hybrid transformer for semantic segmentation with very high resolution imagery. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 9688–9691. IEEE, 2024. 2
- [42] Yuji Wang, Jingchen Ni, Yong Liu, Chun Yuan, and Yansong Tang. Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8159–8168, 2025. 2
- [43] Yuji Wang, Haoran Xu, Yong Liu, Jiaze Li, and Yansong Tang. Sam2-love: Segment anything model 2 in language-aided audio-visual scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28932–28941, 2025. 2
- [44] Rui Xia, Dan Jiang, Quan Zhang, Ke Zhang, and Chun Yuan. Clip-ae: Clip-assisted cross-view audio-visual enhancement for unsupervised temporal action localization. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 2014–2018, 2025. 2
- [45] Haozhe Xing, Shuyong Gao, Yan Wang, Xujun Wei, Hao Tang, and Wenqiang Zhang. Go closer to see better: Camouflaged object detection via object area amplification and figure-ground conversion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5444–5457, 2023. 6
- [46] Shannan Yan, Jingchen Ni, Leqi Zheng, Jiajun Zhang, Peixi Wu, Dacheng Yin, Jing Lyu, Chun Yuan, and Fengyun Rao. Adamem: Adaptive user-centric memory for long-horizon dialogue agents. *arXiv preprint arXiv:2603.16496*, 2026. 2
- [47] Shannan Yan, Leqi Zheng, Keyu Lv, Jingchen Ni, Hongyang Wei, Jiajun Zhang, Guangting Wang, Jing Lyu, Chun Yuan, and Fengyun Rao. Learning cross-view object correspon-

- dence via cycle-consistent mask prediction. *arXiv preprint arXiv:2602.18996*, 2026. [2](#)
- [48] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4146–4155, 2021. [6](#)
- [49] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camo-former: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [6](#)
- [50] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3234–3242, 2021. [6](#), [8](#)
- [51] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12997–13007, 2021. [2](#)
- [52] Quan Zhang, Xiaoyu Liu, Wei Li, Hanting Chen, Junchao Liu, Jie Hu, Zhiwei Xiong, Chun Yuan, and Yunhe Wang. Distilling semantic priors from sam to efficient image restoration models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25409–25419, 2024. [2](#)
- [53] Quan Zhang, Jinwei Fang, Yuxin Qi, Mingyang Wan, Guojun Ma, Ke Zhang, and Chun Yuan. Eav-mamba: Efficient audio-visual representation learning for weakly-supervised temporal action localization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. [2](#)
- [54] Quan Zhang, Jinwei Fang, Rui Yuan, Xi Tang, Yuxin Qi, Ke Zhang, and Chun Yuan. Weakly supervised temporal action localization via dual-prior collaborative learning guided by multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24139–24148, 2025. [2](#)
- [55] Quan Zhang, Yuxin Qi, Xi Tang, Jinwei Fang, Xi Lin, Ke Zhang, and Chun Yuan. IMDPrompter: Adapting SAM to image manipulation detection by cross-view automated prompt learning. In *The Thirteenth International Conference on Learning Representations*, 2025. [2](#)
- [56] Quan Zhang, Yuxin Qi, Xi Tang, Rui Yuan, Xi Lin, Ke Zhang, and Chun Yuan. Rethinking pseudo-label guided learning for weakly supervised temporal action localization from the perspective of noise correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10085–10093, 2025. [2](#)
- [57] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *International Journal of Computer Vision*, 132(10):4157–4234, 2024. [8](#)
- [58] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4504–4513, 2022. [2](#)
- [59] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2022. [8](#)