Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2026.073119

ARTICLE

# Research on Camouflage Target Detection Method Based on Edge Guidance and Multi-Scale Feature Fusion

**Tianze Yu, Jianxun Zhang* and Hongji Chen**

Department of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

*Corresponding Author: Jianxun Zhang. Email: zjx@cqut.edu.cn

Received: 11 September 2025; Accepted: 02 December 2025; Published: 10 February 2026

**ABSTRACT:** Camouflaged Object Detection (COD) aims to identify objects that share highly similar patterns—such as texture, intensity, and color—with their surrounding environment. Due to their intrinsic resemblance to the background, camouflaged objects often exhibit vague boundaries and varying scales, making it challenging to accurately locate targets and delineate their indistinct edges. To address this, we propose a novel camouflaged object detection network called Edge-Guided and Multi-scale Fusion Network (EGMFNet), which leverages edge-guided multi-scale integration for enhanced performance. The model incorporates two innovative components: a Multi-scale Fusion Module (MSFM) and an Edge-Guided Attention Module (EGA). These designs exploit multi-scale features to uncover subtle cues between candidate objects and the background while emphasizing camouflaged object boundaries. Moreover, recognizing the rich contextual information in fused features, we introduce a Dual-Branch Global Context Module (DGCM) to refine features using extensive global context, thereby generating more informative representations. Experimental results on four benchmark datasets demonstrate that EGMFNet outperforms state-of-the-art methods across five evaluation metrics. Specifically, on COD10K, our EGMFNet-P improves $F_\beta$ by 4.8 points and reduces mean absolute error (MAE) by 0.006 compared with ZoomNeXt; on NC4K, it achieves a 3.6-point increase in $F_\beta$. On CAMO and CHAMELEON, it obtains 4.5-point increases in $F_\beta$, respectively. These consistent gains substantiate the superiority and robustness of EGMFNet.

**KEYWORDS:** Camouflaged object detection; multi-scale feature fusion; edge-guided; image segmentation

## 1 Introduction

To survive, wild animals have evolved sophisticated camouflage capabilities as a defense mechanism. Species such as chameleons, flounders, and cuttlefish often disguise themselves by altering their appearance, color, or patterns to blend seamlessly with their surroundings, thereby avoiding detection. In recent years, Camouflaged Object Detection (COD) has garnered increasing interest within the computer vision community, and its technical insights into low-contrast segmentation may benefit domains with analogous challenges, including search and rescue operations [1], medical image segmentation (e.g., polyp segmentation, lung infection segmentation, retinal image segmentation), agricultural management, and recreational art. Consequently, COD has become a prominent research focus in computer vision, given its significant applications and scientific value in medical and industrial fields. However, accurately and reliably detecting camouflage targets remains a challenging task for several reasons. In complex scenarios involving multiple objects or occlusions, the performance of existing methods degrades markedly. Compared with traditional salient object detection (SOD) and other segmentation tasks, camouflage object detection (COD) poses

greater difficulties due to three main factors: (i) the wide variation in the size of camouflaged objects, (ii) their indistinct and blurred boundaries, and (iii) the strong visual similarity between camouflaged objects and their surrounding environments. As illustrated in Fig. 1, current models often fail to identify all camouflaged objects correctly and exhibit poor robustness under occlusion. Experimental results further reveal that most COD models are highly sensitive to background interference and have difficulty capturing the subtle and discriminative semantic cues of camouflaged targets—particularly when their boundaries are missing or seamlessly blended with the environment. Under such circumstances, distinguishing camouflaged objects from cluttered backgrounds and uncertain low-confidence regions becomes exceedingly difficult. To address these challenges, this paper refines the COD problem into three core research questions: (i) how to alleviate the adverse effects of blurred boundaries; (ii) how to precisely locate camouflaged objects with inconspicuous appearances and diverse scales; and (iii) how to suppress dominant background interference to achieve more reliable detection of camouflaged objects. Despite recent progress, two key gaps remain underexplored: (i) contextual reasoning across scales that reconciles object–background ambiguity, and (ii) explicit boundary modeling that preserves fine contours under occlusions and scale variation. Our design directly targets these gaps by combining MSFM for selective multi-scale fusion, DGCM for complementary local–global context reasoning, and EGA for edge-guided refinement.

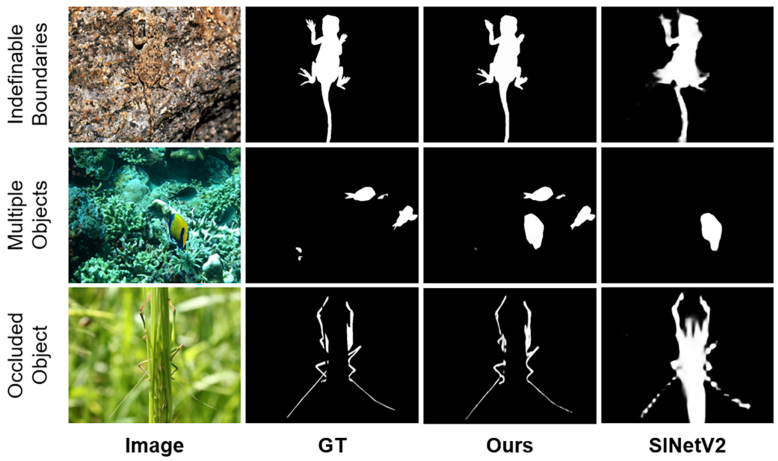


**Figure 1:** The three challenging camouflage scenarios are listed from top to bottom, which include undefined boundaries, multiple objects, and occluded objects. Our model outperforms SINet [2] in these challenging situations

To address the challenge of accurately detecting camouflaged objects, we propose a novel network, EGMFNet, which integrates edge guidance and multi-scale feature fusion. The framework comprises two key modules: (i) the Multi-scale Fusion Module (MSFM), which selects and integrates multi-scale features, and (ii) the Edge-Guided Attention (EGA) module, which leverages edge information of camouflaged objects to guide the network in enhancing boundary segmentation. To achieve precise target localization, we employ scale-space theory [3–5] to simulate zoom-in and zoom-out strategies. Specifically, for camouflaged objects, our model first extracts discriminative features across different "scaled" resolutions using a tripartite architecture. These multi-scale features are subsequently refined and fused using the MSFM. Within this module, we introduce Multi-scale Grouped Dilated Convolution (MSGDC) within each scale branch to further facilitate local multi-scale feature extraction. Furthermore, we employ a Hypergraph Convolution (HGC) Block to model the high-order semantic relationships among pixels by constructing a hypergraph. This approach enables the effective aggregation of discrete features belonging to the same object from across a global context, thereby forming a more holistic feature representation. Consequently, the model

is empowered to capture the critical cues of camouflaged objects with greater efficiency and accuracy. Next, we employ the Dual-Branch Global Context Module (DGCM) to extract rich contextual information from the fused features. The DGCM transforms input features into multi-scale representations through two parallel branches, where attention coefficients are calculated using the Multi-Scale Contextual Attention (MSCA) component. These attention coefficients guide the integration of features, resulting in refined feature representations. The integrated features are then passed through the decoding stage, where we incorporate the Edge-Guided Attention (EGA) module. This module operates on boundary information, decoding features, and multi-level predictions simultaneously. By emphasizing complete edge details and consistent boundary representation across scales, this combined approach allows the model to highlight the intricate boundaries of camouflaged objects effectively. This strategic integration of modules enables the model to excel at capturing both fine-grained semantic details and complete boundary information, ensuring robust performance across varying scales and challenging backgrounds.

Our contributions can be summarized as follows:

1. To extract discriminative feature representations of camouflaged objects, we designed **the Multi-Scale Fusion Module (MSFM)**. MSFM leverages Multi-scale Grouped Dilated Convolution (MSGDC) and Hypergraph Convolution (HGC) to extract, aggregate, and enhance features across different scales. This comprehensive approach enables the acquisition of features possessing fine-grained semantic representations, which effectively boosts the accuracy of Camouflaged Object Detection (COD), while simultaneously reducing both model parameters and computational complexity.
2. We designed the **Hypergraph Convolution Block (HGCBlock)**, a module that models high-order correlations to infer latent dependencies from the global context. This process serves a dual purpose: it enhances the feature consistency among different regions within a camouflaged object to form a more holistic representation, and it improves the discriminability between the object and background features.
3. We proposed an edge-guided approach, namely the **Edge-Guided Attention (EGA) module**, which utilizes edge information to guide network learning and improve segmentation performance for camouflaged objects with weak boundaries.
4. Our model outperforms 30 recent state-of-the-art methods across three benchmark COD datasets, demonstrating its effectiveness and robustness.

## 2 Related Work

### *2.1 Camouflaged Object Detection*

Identifying and segmenting objects that seamlessly blend into their surroundings defines the core challenge of Camouflaged Object Detection (COD). Historically, early attempts tackled this problem through hand-engineered features, quantifying regional similarities between foreground and background patterns. Such conventional strategies, however, demonstrated limited efficacy—only thriving under simplistic or homogeneous backdrop conditions. In complex scenes where foreground-background distinction becomes subtle, these models suffer precipitous performance degradation, underscoring the inherent difficulty of reliable camouflage discovery.

The advent of deep learning has catalyzed a paradigm shift in COD research. SINet [2] exemplifies this transition with a dual-stage architecture: a localization submodule coarsely pinpoints candidate regions, which are subsequently refined by a discriminative submodule for precise segmentation. MirrorNet [6] diverges from this pure data-driven approach, drawing inspiration from human perceptual behavior—specifically, viewpoint alternation—to incorporate a flipped image stream that enhances region saliency. In

parallel, C2Net [7] advances feature integration through attention-mediated cross-layer fusion, dynamically aggregating contextual cues across hierarchical representations to bolster detection fidelity. Addressing scale variability, HGINet [8] deploys a hierarchical graph interaction transformer coupled with dynamic token clustering, enabling adaptive capture of multi-scale object characteristics. Complementarily, AGLNet [9] implements adaptive guidance learning that modulates feature extraction conditioned on object appearance and contextual semantics, thereby improving adaptability to diverse camouflage strategies. Beyond holistic object localization, boundary ambiguity remains a critical bottleneck. BGNet [10] confronts this by embedding edge priors directly into segmentation features through a specialized edge-guided module. BSA-Net [11] pursues analogous objectives via adaptive spatial normalization, achieving more selective background suppression. Similarly, EAMNet [12] formulates edge detection and object segmentation as a mutual refinement process within an edge-aware mirror architecture, fostering cross-task synergies. More recently, WSSCOD [13] circumvented annotation scarcity by harnessing noisy pseudo-labels under weak supervision, while CamoTeacher [14] introduced dual-rotation consistency for semi-supervised learning with limited labeled data.

Collectively, these innovations underscore a trend toward hybrid architectures—unifying multi-scale reasoning, edge-informed refinement, and adaptive supervision—to progressively disentangle camouflaged objects from challenging backgrounds.

### *2.2 Scale Space Integration*

Scale-space theory provides a foundational framework for multi-scale image analysis, aiming to improve the understanding of image structures and extract richer visual information by examining features across different scales. It offers a robust and theoretically grounded means of handling natural variations in scale. The central idea is to progressively blur or downscale an image to generate multi-scale representations, allowing the capture of structures and details at varying levels of granularity. This concept has been extensively applied in computer vision, forming the basis of techniques such as image pyramids and feature pyramids.

Because features at different scales exhibit diverse structural and semantic characteristics, they contribute differently to image representation. However, conventional inverted pyramid architectures often lead to the loss of fine-grained texture and appearance details, which adversely affects dense prediction tasks that rely on the integrity of emphasized regions and boundaries. To mitigate this limitation, recent CNN-based approaches in COD and SOD have investigated cross-layer feature fusion strategies to enhance feature representation, thereby enabling more accurate localization and segmentation of target objects and improving dense prediction performance.

Nevertheless, in COD tasks, existing methods still struggle with the inherent structural ambiguity of camouflaged data, making it difficult for any single scale to capture camouflaged objects comprehensively. To overcome this challenge, we introduce a multi-scale strategy inspired by magnification and reduction mechanisms, designed to model the distinct relationships between objects and backgrounds across scales. This facilitates a more complete perception of camouflaged targets within complex scenes. Furthermore, we explore fine-grained interactions among features in the scale-space across channels to strengthen the model's perceptual capability and overall performance.

### *2.3 Edge-Aware*

How to leverage additional edge information to mitigate the impact of sparse edge pixels has long been a challenge in the design of camouflaged object segmentation (COD) networks. In recent years, some salient object detection (SOD) methods have used edge cues to aid saliency inference. Luo et al. [15] proposed a U-shaped model with an IoU-based edge loss, directly optimizing the edges of the salient object map. Feng et al. [16] introduced a boundary-aware loss between predicted and annotated saliency maps. Xu et al. [17] utilized an edge-aware module to capture edge priors, guiding the network in detecting salient objects. Zhao et al. [18] proposed a network leveraging the complementarity between salient edge and salient object information. Zhang et al. [19] designed a dual-stream model that utilizes a saliency branch and an edge branch to detect the interiors and boundaries of salient objects, respectively. Feature enhancement is achieved by embedding edge information into the saliency branch as spatial weights. Luo et al. [20] proposed a semantic-edge interaction network that provides edge-enhanced semantics and semantic-enhanced edges for SOD through interaction. Bi et al. [21] proposed the STEG-Net method, which leverages extracted edge information to simultaneously guide the extraction of spatial and temporal features. This approach enables the mutual reinforcement of edge information and salient features during the feature extraction process. Xu et al. [22] utilized synthetic images and labels to learn precise boundaries of salient objects without any additional auxiliary data. They achieved saliency detection through a global integral branch and a boundary-aware branch, significantly improving the performance of weakly supervised methods and narrowing the gap with fully supervised methods. It is worth noting that although some salient object detection methods have explored cross-guidance between edge detection and salient object segmentation, their core design philosophy focuses on detecting salient objects that strongly contrast with their surrounding environment. However, due to the fundamental differences between "camouflage" and "salience," these methods are difficult to directly apply to the field of Camouflaged Object Detection (COD). Unlike the aforementioned approaches, we propose a novel scheme to achieve boundary guidance. The most distinct feature and advantage of our method lie in the introduction of multi-level boundary guidance modules during the decoder stage, We no longer treat edges as an independent supervisory branch. Instead, we propose a cross-layer, prediction-driven Edge-Guided Aggregation (EGA) mechanism. In this design, high-level predictions are reused to generate reverse attention and Laplacian boundary cues, which are then employed to gate and refine each decoding stage. This approach differs fundamentally from methods that inject edge information only once or use it merely as an auxiliary loss decoupled from the decoding process. enabling more promising results. Moreover, we demonstrate that boundary guidance significantly enhances the performance of COD.

## 3 Method

### *3.1 Overall Architecture*

Unlike existing camouflage object detection frameworks that rely solely on either multi-scale feature fusion or edge refinement branches, our EGMFNet decouples local-scale feature extraction from high-order semantic reasoning. This separation allows the network to jointly enhance boundary precision and interior integrity, providing a more balanced representation under heavy camouflage. The overall architecture of the proposed method is illustrated in Fig. 2. We simulate a scaling strategy to summarize input images at varying scales, capturing differentiated information across these scales. Specifically, given an input image, a pyramid of images is generated, consisting of the primary scale (input scale) and two auxiliary scales, which simulate zoom-in and zoom-out operations. Features at various scales are first extracted via a shared feature encoder. Subsequently, within a scale-merging sub-network, we employ a Multi-scale Fusion Module (MSFM) that integrates Multi-scale Grouped Dilated Convolution (MSGDC) with our adaptive

hypergraph computation paradigm. This module aligns and effectively aggregates scale-specific information by consolidating auxiliary scales into a primary feature stream. Such a process enhances the model's capability to extract critical information and semantic cues, thereby improving the discriminability between object and background features and boosting the detection performance for small camouflaged objects. The fused features are then passed to the Dual-Branch Global Context Module (DGCM) to extract global and local contextual information. During the decoding phase, Rich Granularity Perception Units (RGPUs) are used to progressively integrate multi-layer features in a top-down manner. To fully utilize edge information across multiple scales, we design the Edge-Guided Attention Module (EGA) module, which interacts with the decoder using reverse attention and boundary attention mechanisms to refine and enhance boundary information. This ensures that boundary details are preserved and amplified throughout the network structure. Finally, to address the uncertainties inherent in the data that affect prediction accuracy, we use an Uncertainty-Aware Loss (UAL) to complement the Binary Cross-Entropy (BCE) loss. This enables the model to effectively distinguish uncertain regions and produce precise and reliable predictions.

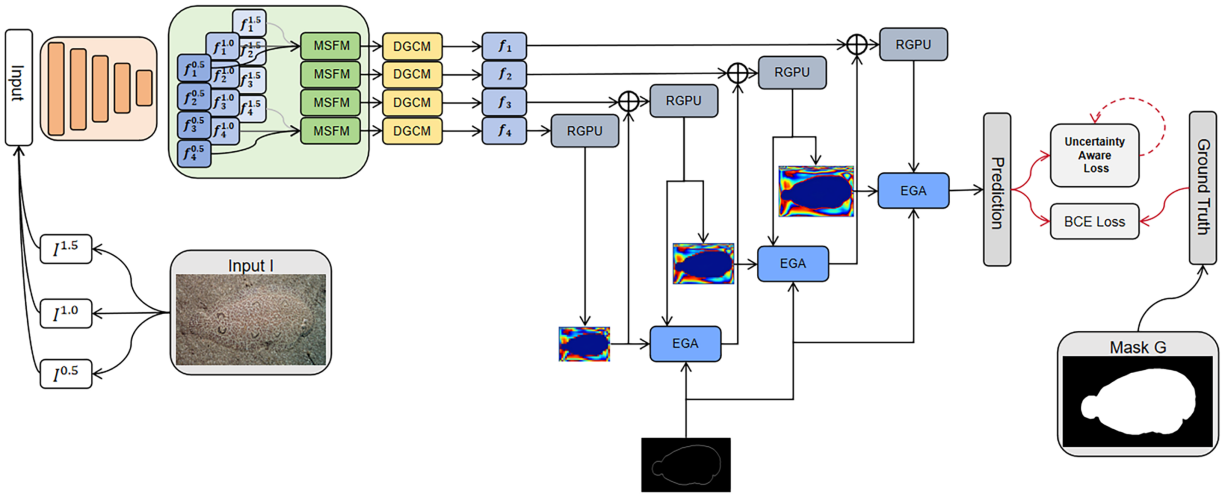


**Figure 2:** The proposed EGMFNet framework leverages a shared triple-feature encoder to extract multi-level features corresponding to different input "scaling" levels. Within the scale merging subnetwork, the Multi-scale Fusion Module (MSFM) is employed to filter and aggregate key information from different scales. Next, a Dual-Branch Global Context Module (DGCM) is used to extract rich contextual information from the fused features. Subsequently, the Edge-Guided Attention (EGA) module interacts with the decoder, performing fusion operations on the edge feature map, the decoder's feature map, and the prediction mask from the lower-level decoder. This interaction emphasizes complete edge details and boundaries across various scales. Finally, the framework generates a probability map that corresponds to the camouflaged object in the input image or frame, delivering precise and robust segmentation results

### *3.2 Triplet Feature Encoder*

Drawing motivation from [23,24], our architecture employs a shared triplet encoder to distill deep features across three distinct scales. This encoder bifurcates into two synergistic submodules: a primary feature extraction backbone and a subsequent channel compression stage. We instantiate the extraction backbone using PVTv2 [25], strategically truncating its classification head to facilitate comprehensive multi-scale representation learning from the input imagery. The compression submodule cascades thereafter, yielding computationally-efficient yet semantically-retentive feature embeddings through dimensionality reduction. Empirical calibration establishes the primary scale at 1.0×, complemented by auxiliary scales of 0.5× and 1.5×, striking an optimal trade-off between representational capacity and computational overhead. This configuration yields a quintet of 64-channel feature maps per scale, denoted as $\{f_i^k\}_{i=1}^5$ where

$k \in \{0.5, 1.0, 1.5\}$, capturing multi-granular visual cues. These multi-scale features then feed forward into our multi-head scale merging subnetwork, subsequently propagating through the hierarchical difference decoder for progressive refinement.

### *3.3 Scale Merging Subnetwork*

The scale fusion sub-network serves as a key component of our framework. Motivated by recent methods proposed for multi-scale feature fusion and cross-level interaction [24,26–29], we design the Multi-Scale Fusion Module (MSFM) to effectively integrate information across different scales. It is designed to effectively filter (weight) and combine features from different scales to capture both fine-grained details and global semantics. While multi-scale fusion has been widely studied in previous COD models, the proposed MSFM differs by explicitly decoupling spatial-scale representation from semantic aggregation. Instead of fusing multi-scale features through simple summation or concatenation, MSFM employs a hierarchical gating strategy that preserves fine-grained details before global reasoning, thereby preventing early feature mixing that may obscure subtle camouflage cues. As illustrated in the Fig. 3, several such units form the scale merging subnetwork. Through filtering and aggregation, it adaptively highlights the representations from different scales. The detailed description of this component is provided below.

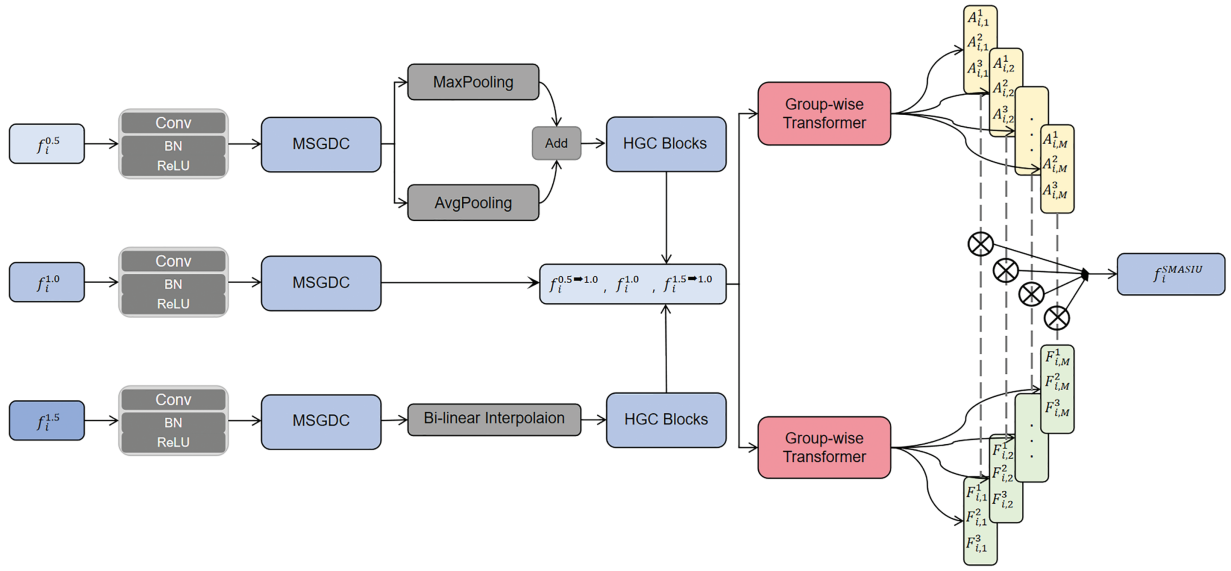


**Figure 3:** Illustration of the multi-scale fusion module (MSFM). ⊗ is the element-wise multiplication. $\phi$ and $\gamma$ are the parameters of the two separated group-wise transformation layers. More details can be found in Section 3.3

#### *3.3.1 Multi-Scale Grouped Dilated Convolution (MSGDC)*

To further and more efficiently capture rich multi-scale information during feature encoding—thereby addressing the challenge posed by camouflage targets of varying sizes—we introduce a Multi-Scale Grouped Dilated Convolution (MSGDC) module for preliminary processing of the input features. As illustrated in Fig. 4, MSGDC employs three parallel grouped convolutions, each with a distinct dilation rate, to extract localized multi-scale features within every scale level. Motivated by the hybrid dilated convolution (HDC) principle [30]: using rates with small/common-free factors mitigates the gridding artifact of dilated sampling, while preserving dense coverage and boundary details, we use three grouped $3 \times 3$ atrous branches with dilation rates $d \in \{1, 3, 5\}$. This design deeply mines the abundant local multi-scale cues hidden in each hierarchy. Compared to standard convolutions and self-attention mechanisms, grouped

convolutions markedly reduce both model parameters and computational complexity, while still preserving rich representational power. For the input feature $\mathbf{X}_{l-1} \in \mathbb{R}^{H\times W\times C}$, the embedding process of the MSGDC module is defined as

$$\begin{aligned} F_{\text{branch}}^{(n)} &= \text{ReLU}\left(\text{BN}\left(\text{Conv}_{3\times3,g}^{\text{dil}=2n-1}(F_{\text{in}})\right)\right), \quad n \in \{1,2,3\}, \\ F_{\text{out}} &= F_{\text{in}} + \text{BN}\left(\text{Conv}_{1\times1}\left(\sum_{n=1}^{3} F_{\text{branch}}^{(n)}\right)\right), \end{aligned} \tag{1}$$

where $F_{\text{in}}$ is the input feature map. $\text{Conv}_{3\times3,g}^{\text{dil}=2n-1}$ denotes a $3\times3$ grouped dilated convolution with a group number of $g$ and a dilation rate of $2n-1$, where $n \in 1,2,3$. BN represents Batch Normalization, and ReLU is the Rectified Linear Unit activation function. $F_{\text{branch}}^{(1)}$, $F_{\text{branch}}^{(2)}$ and $F_{\text{branch}}^{(3)}$ are the outputs of each grouped convolu-tion layers.

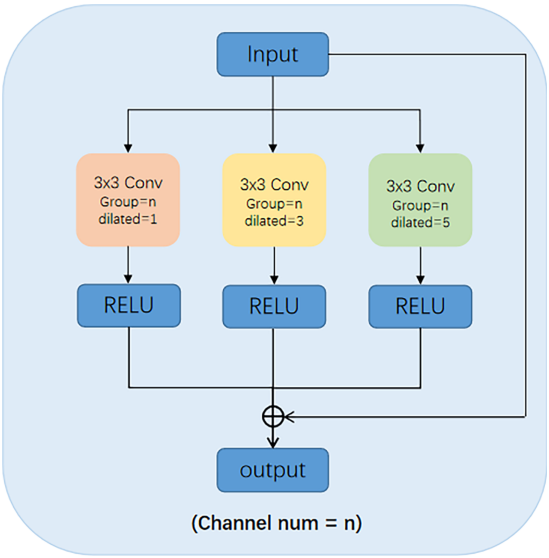


**Figure 4:** Illustration of the binary multi-scale grouped dilated convolution module

### *3.3.2 Scale Alignment*

Prior to scale integration, resolution homogenization is performed to align $f_i^{1.5}$ and $f_i^{0.5}$ with the spatial dimensions of the main scale feature $f_i^{1.0}$. For the high-resolution branch $f_i^{1.5}$, hybrid downsampling is implemented through a combined max-pooling and average-pooling architecture, this dual-stream design preserves fine-grained characteristics critical for camouflaged object representation while fostering cross-scale feature diversity. Conversely, the low-resolution branch $f_i^{0.5}$ undergoes direct upsampling via bilinear interpolation. Following this resolution standardization, the harmonized features advance to the subsequent transformation layers.

### *3.3.3 Hypergraph Convolution Block (HGCBlock)*

Inspired by [31,32], we transcend the inherent limitations of traditional convolutional and self-attention mechanisms that are constrained to pairwise relationship modeling. We introduce an adaptive hypergraph computation paradigm and design a novel Hypergraph Convolution Block (HGCBlock). This module constructs dynamic hypergraph structures to achieve a paradigmatic shift from "feature re-weighting" to "high-order relational reasoning." It adaptively learns vertex participation degrees across different

hyperedges, models high-order correlations within visual features, and effectively aggregates spatially-dispersed yet semantically-related feature regions. This many-to-many relational modeling mechanism endows HGCBlock with enhanced robustness in complex visual scenarios, particularly demonstrating unique advantages in camouflaged object detection and other tasks requiring holistic perception. Different from conventional self-attention or non-local mechanisms that model pairwise dependencies, the proposed HGCBlock operates on a hypergraph structure, capturing k-wise (high-order) relations among feature nodes. This design enables region-level consistency across spatially separated but semantically related fragments—an ability particularly beneficial for camouflage scenes, where object textures are discontinuous and boundaries are ambiguous. In addition, the hypergraph formulation introduces a group-level constraint that naturally filters out noisy pairwise connections, offering stronger robustness than self-attention under low-contrast conditions. For a visual representation of the HGCBlock module, please refer to Fig. 5.

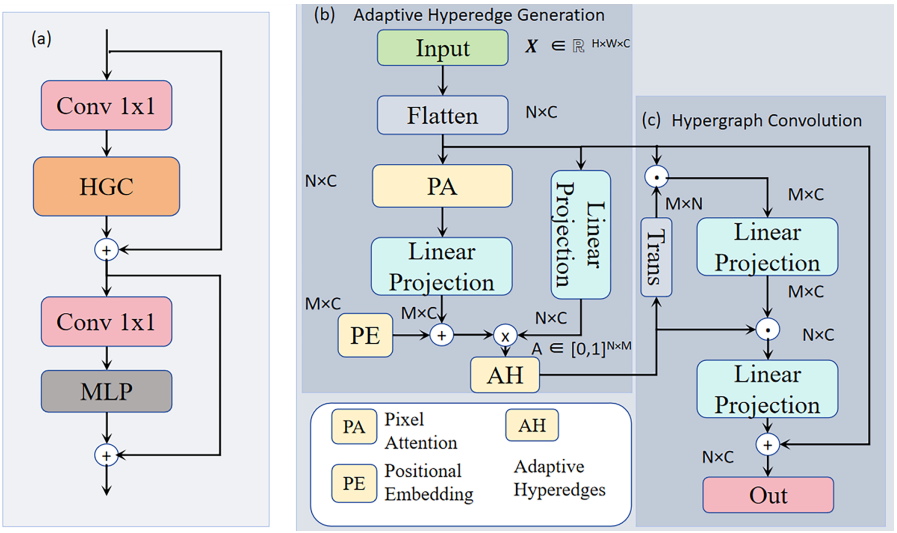


**Figure 5:** An overview of the HGCBlock architecture. (**a**) The HGC block, (**b**) The adaptive hyperedge generation, and (**c**) The hypergraph convolution

Operationally, camouflaged objects frequently present as low-saliency regions with weak textural cues or partial occlusion, causing their discriminative patterns to be marginalized by local convolutional feature extractors. Adaptive hypergraph modeling [33] addresses this deficiency by establishing high-order correlations that surface implicit long-range dependencies from the global contextual field.

This approach formalizes the adaptive hypergraph as $G = \{\mathcal{V}, \mathcal{A}\}$, comprising vertex set $\mathcal{V}$ and adaptive hyperedge set $\mathcal{A}$. The framework implements full connectivity where each hyperedge links to all vertices, with vertex contributions weighted by a continuous, differentiable participation degree. Such interactions are encoded in the continuous participation matrix $A \in [0,1]^{N\times M}$, where dimensions $N$ and $M$ denote vertex and hyperedge quantities, respectively. Individual entries $A_{i,m}$ specify the participation degree of vertex $i$ relative to hyperedge $m$. By enabling flexible modeling of weakly salient object associations within the global semantic context, this mechanism substantially advances camouflage object discovery capabilities.The proposed framework decomposes into two pivotal mechanisms: adaptive hyperedge generation and hypergraph convolution. Within the hyperedge generation stage, contextual vectors are first procured via the Pixel Attention (PA) [34] scheme. A dedicated mapping layer then transforms this attention-derived context into a global offset, which subsequently fuses with learnable positional embeddings (PE) to constitute dynamic hyperedge prototypes. These prototypes distill latent scene-level relationships—a

capability paramount for supplementing deficient salient texture cues in camouflaged instances, notably those of diminutive scale. Following prototype formulation, each input vector $x_i$ traverses a mapping layer to generate its corresponding latent representation $z_i$.

$$z_i = W_{\text{pre}} x_i \in \mathbb{R}^c \tag{2}$$

To further enhance feature diversity, we use a multi-head mechanism that partitions $z_i$ along the feature dimension into $h$ subspaces $\left\{\hat{\mathbf{z}}_i^\tau \in \mathbb{R}^{d_h}\right\}_{\tau=1}^{h}$, where $d_h = C/h$. Similarly, each hyperedge prototype is divided into $h$ corresponding subspaces $\left\{\hat{\mathbf{p}}_m^\tau \in \mathbb{R}^{d_h}\right\}_{\tau=1}^{h}$. Subsequently, within the $\tau$-th subspace, we compute the similarity between the $i$-th vertex query vector and the $m$-th prototype, thereby generating the continuous participation matrix as follows:

$$s_{i,m}^\tau = \frac{\langle \hat{\mathbf{z}}_i^\tau, \hat{\mathbf{p}}_m^\tau \rangle}{\sqrt{d_h}}. \tag{3}$$

Therefore, the overall similarity is defined as the average of the similarities across all subspaces, i.e., $\bar{s}_{i,m} = \frac{1}{h}\sum_{\tau=1}^{h} s_{i,m}^\tau$. Finally, we regard the similarity between the vertex query vector and the prototype as the contribution of the vertex to the hyperedge, and normalize over the vertices to obtain the continuous participation matrix A. The formulation is as follows:

$$A_{i,m} = \frac{\exp(\bar{s}_{i,m})}{\sum_{j=1}^{N} \exp(\bar{s}_{j,m})}. \tag{4}$$

Within the hypergraph convolution stage, feature aggregation commences as each hyperedge synthesizes vertex information into a comprehensive hyperedge representation. This representation subsequently undergoes back-propagation to refresh vertex node features. Through integrating global dependency modeling with bidirectional propagation pathways, the mechanism amplifies weak local cues—including diminutive object boundaries and occluded texture patterns, through contextual guidance. Formally, this process can be defined as:

$$\begin{aligned} f_m &= \sigma\left(W_e \sum_{i=1}^{N} A_{i,m} x_i\right) \\ \tilde{x}_i &= \sigma\left(W_v \sum_{m=1}^{M} A_{i,m} f_m\right) \end{aligned} \tag{5}$$

Summarizing the HGCBlock pipeline, input features initially undergo a $1 \times 1$ convolution for preliminary transformation and dimensionality tuning. The resulting feature maps then enter our HGC (Hypergraph Convolution) module, substituting conventional multi-head self-attention to model high-order semantic correlations across spatially distributed pixel regions. Subsequently, a Multi-Layer Perceptron (MLP) polishes the context-enhanced features generated by HGC. Inter-stage residual connections maintain training stability and facilitate gradient flow throughout the architecture. Mathematically, the HGCBlock operations can be expressed as:

$$\begin{aligned} X'_{i+1} &= HGC(Conv(X_i)) + X_i \\ X_{i+1} &= MLP(Conv(X'i+1)) + X'i+1 \end{aligned} \tag{6}$$

where $X_i$ represents the input features to the $i$-th block, $X'_{i+1}$ and $X_{i+1}$ are the outputs of the HGC and MLP modules, respectively, and $Conv$ denotes the $1 \times 1$ convolution operation.

#### 3.3.4 Multi-Head Spatial Interaction

Drawing upon the multi-head architectural principles established in Transformer [35] and ZoomNext [24], the input feature map's channels are segregated into $M$ independent groups. These segmented groups are distributed across $M$ parallel computational branches, where each branch autonomously processes its exclusive feature subset. Such structural design substantially expands the model's capability to acquire varied fine-grained spatial attention configurations, consequently augmenting the feature space's expressive potential. At the branch level, the exclusive feature group initially passes through multiple convolutional layers to derive tri-channel intermediate representations tailored for attention calculation. A sequential softmax activation is then imposed on these intermediate maps, producing spatial attention weight maps designated as $A_m^k$, aligned with each scale inside the group. Through weighted fusion integration, these attention weights merge with the source feature maps to constitute the ultimate weighted feature representation. In this context, $k \in \{1, 2, 3\}$ indexes the scale groups while $m \in \{1, 2, \ldots, M\}$ indexes the attention groups. The operational flow unfolds as:

$$
\begin{aligned}
F_i &= \mathcal{U}(f_i^{0.5}, f_i^{1.0}, \mathcal{D}(f_i^{1.5})), \\
\hat{F}_i &= \{\mathrm{trans}(F_i, \phi^m)\}_{m=1}^{M}, \\
A_i &= \{\mathrm{softmax}(\hat{F}_{i,m})\}_{m=1}^{M}, \\
\tilde{F}_i &= \{\mathrm{trans}(F_i, \gamma^m)\}_{m=1}^{M}, \\
f_i^{\mathrm{MSFM}} &= \left\{A_i^1 \otimes \tilde{F}_i^1 + A_i^2 \otimes \tilde{F}_i^2 + A_i^3 \otimes \tilde{F}_i^3\right\}_{m=1}^{M},
\end{aligned}
\tag{7}
$$

where $[\star]$ represents the concatenation operation. $\mathcal{U}$ and $\mathcal{D}$ refer to the bi-linear interpolation and hybrid adaptive pooling operations mentioned above, respectively. $\mathrm{trans}(\star, \phi)$ and $\mathrm{trans}(\star, \gamma)$ indicate the several simple linear transformation layers with the parameters $\phi$ and $\gamma$ in the attention generator. And $\otimes$ is the element-wise multiplication operation. The enhanced features outputted by all M branches are concatenated along the channel dimension to form a unified feature map. These designs aim to adaptively and selectively aggregate the scale-specific information to explore subtle but critical semantic cues at different scales, boosting the feature representation.

### 3.4 Dual-Branch Global Context Module (DGCM)

We adopt the proposed MSFM to fuse multi-scale features by introducing a synergistic multi-attention mechanism to obtain information-rich attention-based fused features. Furthermore, global contextual information is critical for improving the performance of camouflaged object detection (COD). Therefore, Inspired by [7,36,37], We propose a dual-branch global context module (DGCM) to extract rich global contextual information from the fused features. The local branch primarily focuses on extracting fine-grained features, which play a critical role in detecting camouflaged objects with blurry boundaries or small sizes. The global branch, on the other hand, uses average pooling to capture global semantic information from the image, which helps distinguish camouflaged objects from their surrounding environment, especially in complex backgrounds. These two branches complement each other through the Multi-Scale Contextual Attention (MSCA) mechanism. MSCA assigns different attention coefficients to each branch, adaptively adjusting the weight between local and global features to ensure they collectively generate more accurate feature representations. Specifically, the output $F \in \mathbb{R}^{W \times H \times C}$ of MSFM is fed into two branches through convolutional operations and average pooling, respectively. This results in sub-features $F_c \in \mathbb{R}^{W \times H \times C}$ and

$F_p \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$. Next, $F_c$ and $F_p$ are passed to the MSCA component, where each MSCA module generates a weighting coefficient that represents the attention weights for different scales and channel information. Element-wise multiplication is applied to fuse the outputs of MSCA with the corresponding features ($F_C$ or $F_p$), resulting in $F_cm \in \mathbb{R}^{W\times H\times C}$ and $F_pm \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$. Subsequently, the features from the two branches are fused using an addition operation to obtain $F_cpm$. Finally, a residual structure is utilized to combine $f$ with $F_cpm$ to produce $f_0$. The above process can be described as follows:

$$\begin{aligned} &F_c = \mathcal{C}(F), F_{cm} = F_c \otimes \mathcal{M}(F_c), \\ &F_p = \mathcal{C}(\mathcal{P}(F)), F_{pm} = F_p \otimes \mathcal{M}(F_p), \\ &F' = \mathcal{C}\big(F \oplus \mathcal{C}(F_{cm} \oplus \mathcal{U}(F_{pm}))\big), \end{aligned} \tag{8}$$

where $\mathcal{C}$, $\mathcal{P}$, $\mathcal{M}$, and $\mathcal{U}$ represent the convolution, average pooling, MSCA, and upsampling operations, respectively. It is worth noting that the proposed DGCM is used to enhance the fusion features of MSFM, which makes our model adaptively extract multi-scale information from a specific level during the training phase. The detailed structure is shown in Fig. 6.

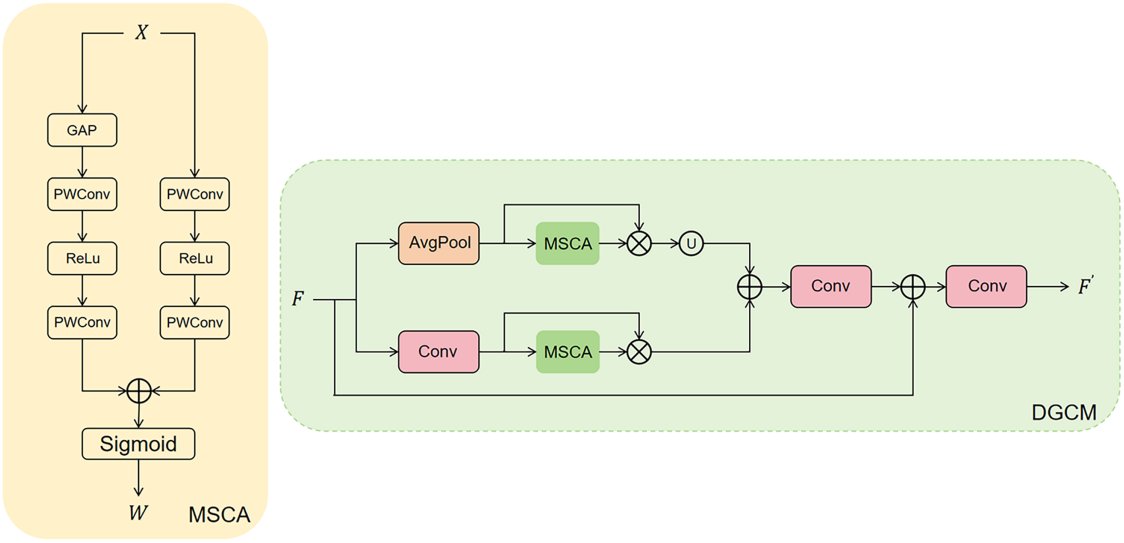


**Figure 6:** Illustration of the dual-branch global context module. U denotes upsampling

### *3.5 Edge-Guided Attention Module (EGA)*

To ensure that the network pays greater attention to edge details when processing images, we integrate edge semantic information into the network to guide learning and fuse features at different levels. We designed an Edge-Guided Attention (EGA) module, which is meticulously designed to integrate and leverage features from various hierarchical levels, effectively addressing the weak boundary issue in camouflaged object detection. At the $i$-th layer, the EGA module takes three inputs: embedded features from the decoder $f_i^d$, edge information features $f_i^e$, and higher-level predicted features generated by the decoder $\hat{f}_{i-1}^d$. The EGA module processes these inputs and produces an output feature map denote as $f_i^{de}$. The specific operations performed by the EGA module are detailed below. For a visual representation of the EGA module, please refer to Fig. 7.

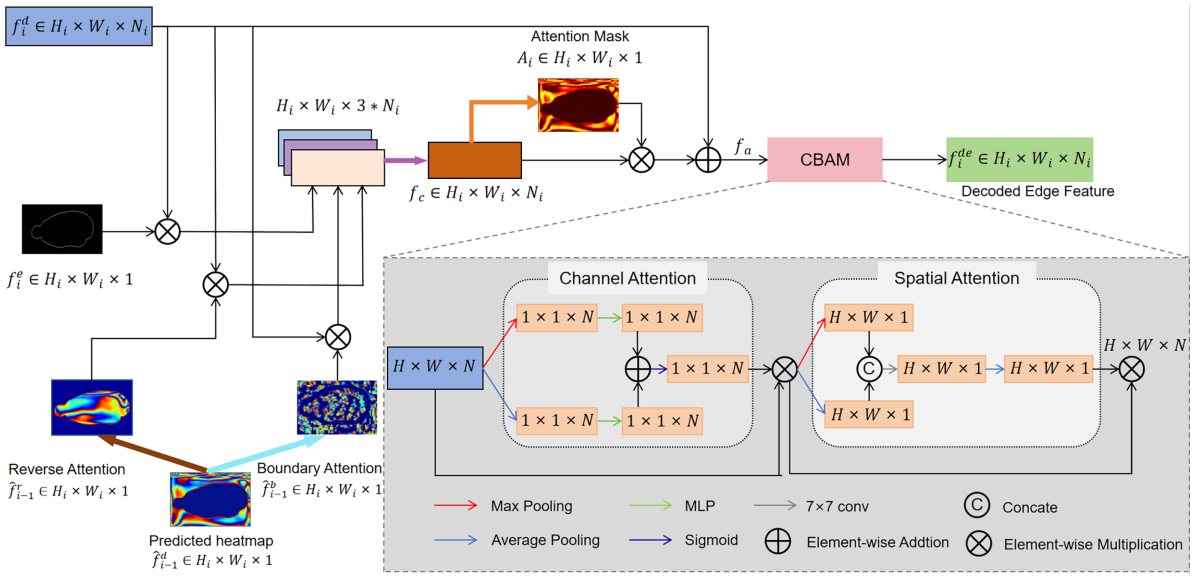


**Figure 7:** The architecture of our EGA block, which takes features from the current layer of the decoder $f_i^d \in \mathbb{R}^{H_i \times W_i \times 1}$, the prediction feature maps from the lower layers of the decoder $f_i^e \in \mathbb{R}^{H_i \times W_i \times 1}$, and boundary information features $f_{i-1}^d \in \mathbb{R}^{H_i \times W_i \times 1}$ as its input. $H_i$, $W_i$, $N_i$ denote height, width, and the number of channels of the input at the layer $i$th

We integrate the EGA module with the encoder. At the $i$-th layer, the EGA module receives multiple inputs, including the visual features from the current layer of the decoder $f_i^d \in \mathbb{R}^{H_i \times W_i \times 1}$, the prediction feature maps from the lower layers of the decoder $\hat{f}_{i-1}^d \in \mathbb{R}^{H_i \times W_i \times 1}$, and boundary information features $f_i^e \in \mathbb{R}^{H_i \times W_i \times 1}$. Inspired by [38], we decompose the low-level prediction maps and generate two distinct attention maps: i) a reverse attention map $\hat{f}_{i-1}^r$, calculated as $\hat{f}_{i-1}^r = 1 - f_{i-1}^d$ computed to reassess and refine imprecise prediction maps from higher layers; ii) a boundary attention map $\hat{f}_{i-1}^b$, generated by applying the Laplace operator, i.e., $\hat{f}_{i-1}^b = L_0(\hat{f}_{i-1}^d)$. Subsequently, we perform element-wise multiplication between the current decoder features and the three attention maps namely $f_i^e$, $\hat{f}_{i-1}^r$, $\hat{f}_{i-1}^b$. This process culminates in the creation of a combined feature map, which is characterized as follows:

$$f_i^c = \mathrm{Conv}\left(\left[f_i^e \otimes f_i^d\right], \left(\hat{f}_{i-1}^b \otimes f_i^d\right), \left(\hat{f}_{i-1}^r \otimes f_i^d\right)\right) \tag{9}$$

where [.] denotes concatenation. we introduce an attention mask denoted as $A_i$ at level $i$. This mask serves the purpose of guiding the model's attention toward vital regions while simultaneously suppressing background noise and redundant information. The attention feature map at the $i$-th layer, $f_i^a$, is defined as follows:

$$f_i^a = f_i^d + (f_i^c \otimes A_i), \quad \text{where} \quad A_i = \sigma(\mathrm{Conv}(f_i^c)) \tag{10}$$

In this context, the symbol $\sigma$ signifies the sigmoid function. As depicted in Fig. 7, the attention masks $A_i$ exhibit markedly elevated values precisely at pixels located along the edges of the polyp. In simpler terms, the fusion of deep features and Laplacian features empowers the model to prioritize the polyp's edge accurately. Building upon insights from [39], we subject the attention feature $f_i^a$ to a CBAM (Convolutional Block Attention Module) for recalibration purposes. This step facilitates the capture of feature correlations between the boundary and the background region. The CBAM comprises two consecutive blocks: channel attention, concentrating on the channel dimension, and spatial attention, centering on the spatial dimension. In the channel attention, the attention feature map $f_i^a$ is refined by convolution with kernels $1 \times 1 \times N_i$. In the spatial attention, spatial kernels of $H_i \times W_i \times 1$ are used. The configuration of this module is visually depicted

in Fig. 7. This figure uses $H$, $W$, $N$ for a general case. Consequently, this process yields a refined decoding feature $f_i^{de}$, i.e., $f_i^{de} = \text{CBAM}(f_i^a)$.

## 4 Experiments

### *4.1 Experiment Setup*

#### *4.1.1 Dataset*

We evaluated EGMFNet on four benchmark datasets: CAMO [40], CHAMELEON [41], COD10K [2], and NC4K [42]. CAMO is a widely used COD dataset that consists of 1250 camouflaged images and 1250 non-camouflaged images. CHAMELEON contains 76 manually annotated images. COD10K, the largest COD dataset to date, includes 5066 camouflaged images, 3000 background images, and 1934 non-camouflaged images. NC4K is the largest testing dataset, comprising 4121 samples. Following the dataset splits defined in [2,43], we used all images containing camouflaged objects in our experiments. Specifically, 3040 images from COD10K and 1000 images from CAMO were used for training, while the remaining images were reserved for testing.

#### *4.1.2 Implementation Details*

We implement EGMFNet using PyTorch [44], with the training configuration following recent best practices [7,24,45]. The encoder of our network is sed on PVTv2 [25], pre-trained on the ImageNet dataset. Additionally, we report results for alternative backbones, including CNN-based ResNet [46] and EfficientNet [47].

We use the Adam optimizer with $\beta = (0.9, 0.999)$ for model parameter updates. The learning rate is initialized at 0.0001 and follows a step decay strategy. The model is trained end-to-end for 100 epochs on an NVIDIA A6000 GPU, with a batch size of 8. During training, the primary input $I_{1.0}$ and the ground truth $G$ are bilinearly interpolated to a resolution of $384 \times 384$. For inference, $I_{1.0}$ and the predictions are interpolated to $384 \times 384$, while $G$ is resized accordingly. Considering that some methods use a smaller resolution of $352 \times 352$, we present the corresponding results for different resolutions and backbones in Table 1. The training data is augmented using random flipping, rotation, and color jittering, while no test-time augmentation is applied during inference.

#### *4.1.3 Evaluation Metrics*

We use five common metrics for evaluation based on [48,49], including the S-measure ($S_m$), weighted F-measure ($F_\beta^\omega$), mean absolute error (MAE), F-measure ($F_\beta$), E-measure ($E_m$). In general, a better COD method will present higher $S_m$, $F_\beta^\omega$, $F_\beta$, $E_m$ and lower M.

**Table 1:** Performance comparison on CAMO, CHAMELEON, COD10K, and NC4K datasets

| Model | Param. | CAMO | | | | | CHAMELEON | | | | | COD10K | | | | | NC4K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | MAE ↓ | $F_\beta \uparrow$ | $E_m$ | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | MAE ↓ | $F_\beta \uparrow$ | $E_m$ | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | MAE ↓ | $F_\beta \uparrow$ | $E_m$ | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | MAE ↓ | $F_\beta \uparrow$ | $E_m$ |
| **Convolutional Neural Network Based Methods** | | | | | | | | | | | | | | | | | | | | | |
| SINet | 48.947M | 0.745 | 0.644 | 0.091 | 0.702 | 0.829 | 0.872 | 0.806 | 0.034 | 0.827 | 0.946 | 0.776 | 0.631 | 0.043 | 0.679 | 0.874 | 0.808 | 0.723 | 0.058 | 0.769 | 0.883 |
| $C^2$FNet | 28.411M | 0.796 | 0.719 | 0.080 | 0.762 | 0.864 | 0.888 | 0.829 | 0.032 | 0.844 | 0.946 | 0.813 | 0.686 | 0.036 | 0.723 | 0.901 | 0.838 | 0.762 | 0.049 | 0.794 | 0.906 |
| SINetV2 | 26.976M | 0.820 | 0.743 | 0.070 | 0.782 | 0.895 | 0.888 | 0.816 | 0.030 | 0.835 | 0.961 | 0.815 | 0.680 | 0.037 | 0.718 | 0.906 | 0.847 | 0.770 | 0.048 | 0.805 | 0.914 |
| SLSR | 50.935M | 0.786 | 0.700 | 0.082 | 0.746 | 0.857 | 0.891 | 0.822 | 0.034 | 0.841 | 0.948 | 0.802 | 0.674 | 0.037 | 0.716 | 0.899 | 0.840 | 0.765 | 0.048 | 0.804 | 0.907 |
| MGL-R | 63.595M | 0.775 | 0.673 | 0.088 | 0.726 | 0.842 | 0.894 | 0.806 | 0.028 | 0.829 | 0.943 | 0.818 | 0.661 | 0.036 | 0.712 | 0.892 | 0.836 | 0.741 | 0.055 | 0.782 | 0.893 |
| PFNet | 46.498M | 0.782 | 0.695 | 0.082 | 0.712 | 0.856 | 0.882 | 0.810 | 0.033 | 0.829 | 0.945 | 0.805 | 0.660 | 0.041 | 0.712 | 0.890 | 0.829 | 0.746 | 0.051 | 0.789 | 0.898 |
| UJSC | 217.982M | 0.800 | 0.728 | 0.073 | 0.772 | 0.882 | 0.890 | 0.819 | 0.030 | 0.842 | 0.956 | 0.809 | 0.684 | 0.036 | 0.721 | 0.891 | 0.842 | 0.772 | 0.047 | 0.806 | 0.907 |
| SegMaR | 56.215M | 0.815 | 0.753 | 0.071 | 0.795 | 0.884 | 0.906 | 0.860 | 0.025 | 0.872 | 0.959 | 0.833 | 0.724 | 0.034 | 0.757 | 0.906 | 0.841 | 0.781 | 0.046 | 0.821 | 0.907 |
| CamoFormer-R | 71.403M | 0.817 | 0.752 | 0.067 | 0.792 | 0.885 | 0.898 | 0.847 | 0.025 | 0.867 | 0.956 | 0.838 | 0.726 | 0.029 | 0.753 | 0.930 | 0.855 | 0.808 | 0.038 | 0.821 | 0.914 |
| BGNet | 79.853M | 0.812 | 0.742 | 0.076 | 0.789 | 0.882 | 0.906 | 0.852 | 0.027 | 0.860 | 0.958 | 0.831 | 0.722 | 0.033 | 0.753 | 0.911 | 0.851 | 0.788 | 0.045 | 0.820 | 0.916 |
| BSA-Net | 32.585M | 0.794 | 0.717 | 0.079 | 0.763 | 0.867 | 0.895 | 0.841 | 0.027 | 0.858 | 0.957 | 0.819 | 0.699 | 0.034 | 0.738 | 0.901 | 0.842 | 0.771 | 0.048 | 0.808 | 0.907 |
| DGNet | 18.113M | 0.838 | 0.768 | 0.057 | 0.805 | 0.914 | 0.890 | 0.816 | 0.029 | 0.834 | 0.956 | 0.822 | 0.692 | 0.033 | 0.727 | 0.911 | 0.857 | 0.783 | 0.042 | 0.813 | 0.922 |
| EGMFNet-R(Ours) | **27.986M** | 0.862 | 0.830 | 0.046 | 0.832 | 0.892 | **0.942** | **0.921** | **0.012** | 0.916 | **0.975** | 0.895 | 0.820 | 0.020 | 0.842 | 0.937 | 0.902 | 0.859 | 0.038 | 0.876 | 0.928 |
| EGMFNet-E(Ours) | **20.892M** | **0.912** | **0.884** | **0.032** | **0.902** | **0.936** | 0.928 | 0.916 | **0.012** | **0.920** | 0.974 | **0.916** | **0.861** | **0.016** | **0.876** | **0.951** | **0.912** | **0.895** | **0.030** | **0.902** | **0.942** |
| **Vision Transformer Based Methods** | | | | | | | | | | | | | | | | | | | | | |
| CamoFormer-P | 71.403M | 0.872 | 0.831 | 0.046 | 0.854 | 0.938 | 0.910 | 0.865 | 0.022 | 0.882 | 0.966 | 0.869 | 0.786 | 0.023 | 0.811 | 0.939 | 0.892 | 0.847 | 0.030 | 0.868 | 0.946 |
| MSCAF-Net | 30.364M | 0.873 | 0.828 | 0.046 | 0.852 | 0.937 | 0.912 | 0.865 | 0.022 | 0.876 | 0.970 | 0.887 | 0.775 | 0.024 | 0.798 | 0.936 | 0.887 | 0.838 | 0.032 | 0.860 | 0.942 |
| HitNet | 25.727M | 0.849 | 0.809 | 0.055 | 0.831 | 0.910 | 0.921 | 0.897 | 0.019 | 0.900 | 0.972 | 0.871 | 0.806 | 0.023 | 0.823 | 0.938 | 0.875 | 0.834 | 0.037 | 0.854 | 0.929 |
| DTINet | 36.663M | 0.857 | 0.796 | 0.050 | 0.812 | 0.912 | 0.883 | 0.813 | 0.033 | 0.820 | 0.928 | 0.825 | 0.695 | 0.034 | 0.712 | 0.893 | 0.863 | 0.792 | 0.041 | 0.811 | 0.915 |
| TPRNet | 52.845M | 0.814 | 0.781 | 0.076 | 0.816 | 0.872 | 0.890 | 0.816 | 0.031 | 0.826 | 0.936 | 0.830 | 0.725 | 0.034 | 0.779 | 0.893 | 0.857 | 0.790 | 0.048 | 0.816 | 0.906 |
| FSPNet | 69.787M | 0.856 | 0.799 | 0.050 | 0.831 | 0.928 | 0.908 | 0.851 | 0.023 | 0.867 | 0.965 | 0.878 | 0.816 | 0.035 | 0.843 | 0.937 | 0.878 | 0.816 | 0.035 | 0.843 | 0.937 |
| SARNet | 47.477M | 0.874 | 0.844 | 0.046 | 0.866 | 0.935 | 0.933 | 0.909 | 0.017 | 0.915 | 0.978 | 0.885 | 0.820 | 0.021 | 0.839 | 0.947 | 0.889 | 0.851 | 0.030 | 0.872 | 0.941 |
| ZoomNeXt | 84.774M | 0.889 | 0.857 | 0.041 | 0.875 | 0.945 | 0.924 | 0.885 | 0.018 | 0.896 | 0.975 | 0.898 | 0.827 | 0.018 | 0.848 | 0.956 | 0.903 | 0.862 | 0.028 | 0.884 | 0.951 |
| EGMFNet-P(Ours) | **83.981M** | **0.930** | **0.912** | **0.027** | **0.920** | **0.957** | **0.961** | **0.942** | **0.010** | **0.941** | **0.981** | **0.937** | **0.893** | **0.012** | **0.896** | **0.961** | **0.930** | **0.913** | **0.027** | **0.920** | **0.958** |

Note: The best results are highlighted in black.

### *4.2 Comparisons with State-of-the-Arts*

To assess the proposed method, we compared it with several state-of-the-art image-based and video-based approaches. The results of all these methods were obtained either from existing public datasets or generated by models retrained using the corresponding code.

**Quantitative Evaluation.** Table 1 summarizes the comparison of the proposed method, and Fig. 8 displays the qualitative performance of all methods. Table 1 presents a detailed comparison of the image COD (Concealed Object Detection) task. It can be observed that, without relying on any post-processing techniques, the proposed model consistently and significantly outperforms recent methods across all datasets. Compared to the best methods with different backbones (ZoomNeXt [50], DGNet [43], and SARNet [51]), which have surpassed other existing methods, our approach still demonstrates performance advantages on these datasets. Furthermore, in the CAMO (Challenge on Automatic Modulation Recognition) dataset, our EGMFNet (Enhanced Global Multi-Feature Network) with ResNet50 as the backbone network scores 10.

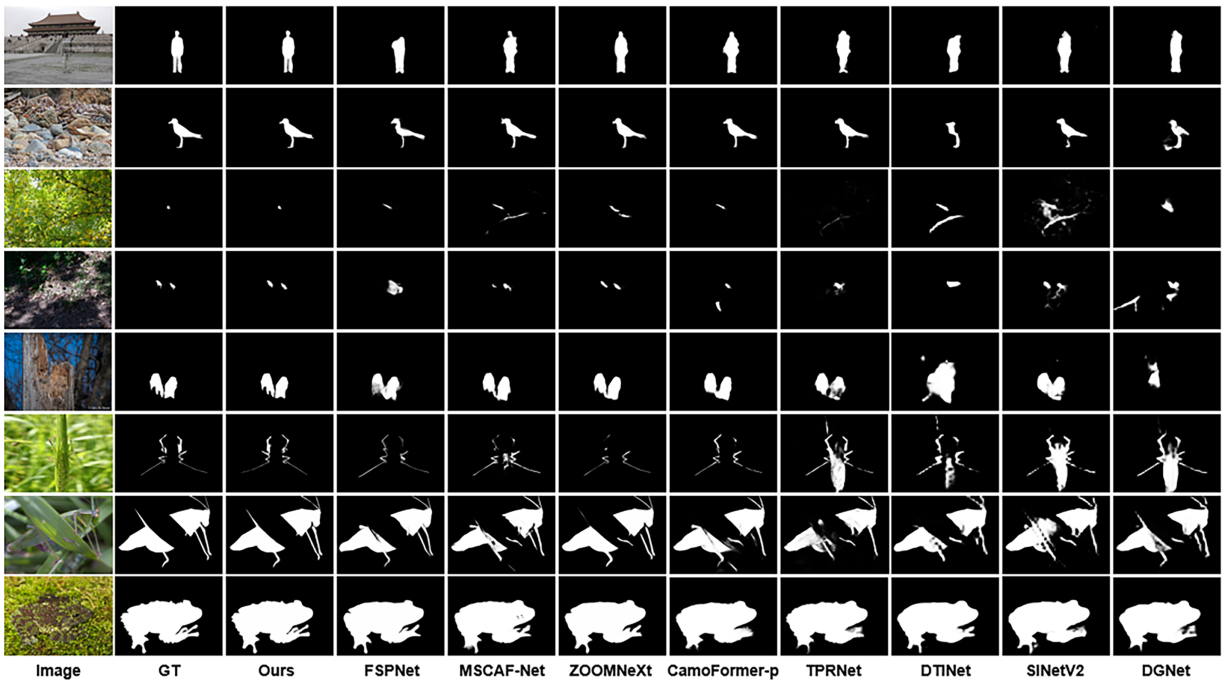


**Figure 8:** Some recent image COD methods and ours on different types of samples

A visual comparison of different methods on several typical samples is shown in Fig. 8. To better demonstrate the performance of these models, several representative samples from the COD domain, encompassing various complex scenarios, were selected. For instance, small objects (rows 3–4), medium-sized objects (rows 1–2), large objects (row 8), occlusions (rows 6–7), multiple objects (rows 4–6), and ill-defined boundaries (row 1) are depicted in Fig. 8. These results intuitively showcase the superior performance of the proposed method. Additionally, it can be observed that our predictions feature clearer and more complete object regions with sharper contours.

### *4.3 Ablation Study*

In this model, MSFM, EGA, and DGCM are critical structures. We installed them one by one on the baseline model to evaluate their performance. The results are shown in Table 2. As can be seen, our baseline

model demonstrates good performance, which may be attributed to a more reasonable network architecture. As can be seen from the table, the three proposed modules demonstrate performance improvements when compared against the baseline. Furthermore, the visual results in Fig. 9 show that the three modules can mutually benefit from each other, reducing their respective errors and achieving more precise localization and target discrimination. To validate the effectiveness of each key module, we designed nine ablation experiments, and the results are presented in Table 2. In Experiment 0 (Basic), all MSFM, EGA, and DGCM modules were removed, leaving only the baseline. In Experiment 1.1 (Basic+MSFM-Base), the MSFM-Base module was added. In Experiment 1.2 (Basic+MSFM-Base+MSGDC), the Multi-scale Grouped Dilated Convolution module was integrated to assess its impact on feature aggregation. In Experiment 1.3 (Basic+MSFM-Base+SMABlock), the synergistic multi-attention block from our original design was added for comparison. In Experiment 1.4 (Basic+MSFM-Base+HGCBlock), we replaced the attention block with our proposed Hypergraph Convolution block to verify its effectiveness in high-order relationship modeling. In Experiment 2 (Basic+MSFM), We integrated the MSGDC Module and the HGCBlock. In Experiment 3 (Basic+DGCM), the DGCM module was added. In Experiment 4 (Basic+MSFM+DGCM), both MSFM and DGCM were retained. In Experiment 5 (Basic+EGA), the EGA module was added to introduce edge information. Finally, Experiment 6 (Basic+MSFM+DGCM+EGA) represents the complete model, consistent with the structure in Fig. 2.

**Table 2:** Ablation study on the COD10K and NC4K Datasets. MSFM-Base: multi-scale fusion module (Base); MSGDC: multi-scale grouped dilated convolution; SMABLOCK: multi-collaborative attention block; HGCBlock: hypergraph convolution block; DGCM: dual-branch context module; EGA: edge-guided attention module

| No. | Ver | COD10K | | | | | NC4K | | | | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | MAE ↓ | $F_\beta \uparrow$ | $E_m$ | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | MAE ↓ | $F_\beta \uparrow$ | $E_m$ | |
| 0 | Basic | 0.826 | 0.679 | 0.032 | 0.722 | 0.901 | 0.852 | 0.761 | 0.048 | 0.801 | 0.913 | 0.00% |
| 1.1 | Basic+MSFM-Base | 0.854 | 0.719 | 0.031 | 0.762 | 0.916 | 0.868 | 0.782 | 0.041 | 0.821 | 0.921 | ↑4.22% |
| 1.2 | Basic+MSFM-Base+MSGDC (d = 1, 3, 5) | 0.861 | 0.748 | 0.031 | 0.817 | 0.928 | 0.883 | 0.829 | 0.036 | 0.861 | 0.923 | ↑7.98% |
| 1.2* | Basic+MSFM-Base+MSGDC (d = 2, 4, 6) | 0.858 | 0.742 | 0.033 | 0.810 | 0.921 | 0.879 | 0.824 | 0.038 | 0.855 | 0.920 | ↑6.78% |
| 1.3 | Basic+MSFM-Base+SMABlock | 0.860 | 0.751 | 0.032 | 0.814 | 0.925 | 0.879 | 0.835 | 0.034 | 0.863 | 0.922 | ↑8.09% |
| 1.4 | Basic+MSFM-Base+HGCBlock | 0.862 | 0.754 | 0.031 | 0.815 | 0.927 | 0.880 | 0.838 | 0.034 | 0.865 | 0.922 | ↑8.57% |
| 2 | Basic+MSFM | 0.865 | 0.756 | 0.028 | 0.826 | 0.930 | 0.885 | 0.842 | 0.031 | 0.869 | 0.925 | ↑10.59% |
| 3 | Basic+DGCM | 0.836 | 0.736 | 0.032 | 0.766 | 0.907 | 0.861 | 0.789 | 0.045 | 0.836 | 0.926 | ↑3.31% |
| 4 | Basic+MSFM+DGCM | 0.887 | 0.811 | 0.019 | 0.871 | 0.942 | 0.891 | 0.893 | 0.027 | 0.897 | 0.932 | ↑17.24% |
| 5 | Basic+EGA | 0.896 | 0.789 | 0.028 | 0.746 | 0.957 | 0.912 | 0.866 | 0.039 | 0.855 | 0.951 | ↑9.72% |
| 6 | Basic+MSFM+DGCM +EGA | 0.937 | 0.893 | 0.012 | 0.896 | 0.961 | 0.930 | 0.913 | 0.027 | 0.920 | 0.958 | ↑23.09% |

Note: The asterisk indicates the computational efficiency and detection performance when using the expansion rate combination of (2, 4, 6).

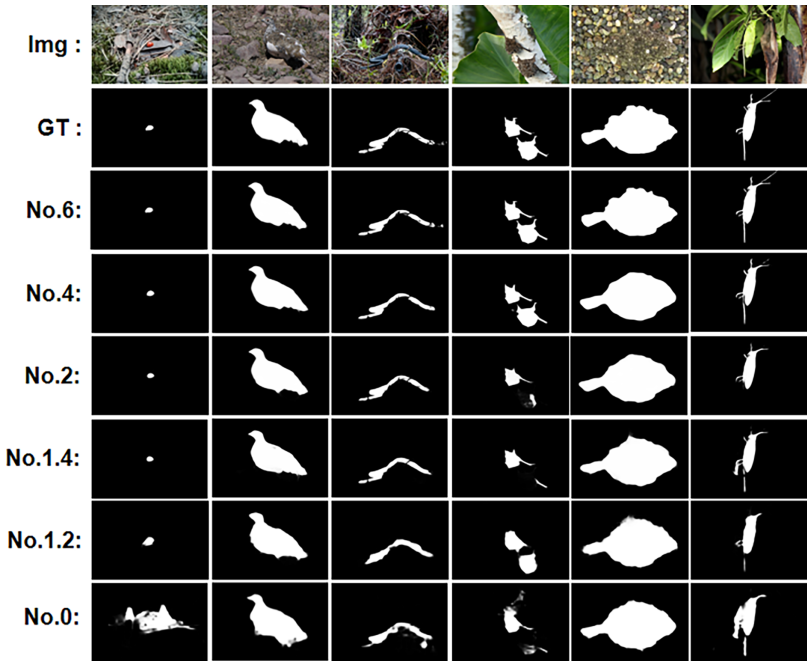


**Figure 9:** Visual comparisons for showing the effects of the proposed components. The names "No.*" are the same as those in Table 2

**Effectiveness of MSFM-Base.** We investigated the importance of MSFM-Base. From Table 2, we observe that Experiment 1.1 (Basic+MSFM-Base) outperforms Experiment 0 (Basic), which clearly indicates that the multi-scale merging unit module is necessary for improving COD performance.

**Effectiveness of MSFM.** We further examined the contribution of MSFM. We observed that Experiment 2 (Basic+MSFM) improves the performance of Experiment 0 (Basic) on all two benchmark datasets. Next, we evaluated the individual contributions of the key components within the MSFM. Adding MSGDC (Experiment 1.2) brought further improvements, validating its role in expanding the receptive field. We report the computational efficiency and detection performance of the three parallel grouped convolutions under different dilation-rate combinations in Table 2. The comparison shows that the configuration with dilation rates 1, 3, 5 delivers the best performance, and thus it is adopted for all subsequent experiments. More importantly, we compared the effects of the conventional attention block (SMABlock) and our proposed HGCBlock. While SMABlock (Experiment 1.3) enhanced performance over the baseline, the introduction of the HGCBlock in Experiment 1.4 yielded more substantial gain. This result powerfully demonstrates the superior capability of our HGC module in modeling high-order semantic relationships. By moving beyond simple feature re-weighting, the HGC aggregates globally-distributed but semantically-related features, which is critical for identifying camouflaged objects.

Experiment 2, which combines both MSGDC and our HGCBlock, represents the optimal configuration of our MSFM, achieving a remarkable balance of performance and advanced feature representation before incorporating other modules.

**Effectiveness of DGCM.** We further investigated the contribution of DGCM. We observed that Experiment 3 (Basic+DGCM) improves the performance of Experiment 0 (Basic) on all two benchmark datasets. These improvements suggest that introducing the dual-branch global context module enables our model to accurately detect objects.

**Effectiveness of MSFM&DGCM.** To evaluate the combination of MSFM and DGCM, we conducted an ablation study for Experiment 4. As shown in Table 2, the combination generally outperforms the previous four settings.

**Effectiveness of EGA.** We further investigated the contribution of EGA. We found that Experiment 5 (Basic+EGA) achieves significant performance improvement across all benchmark datasets, with average performance increases of 6.5%, 4.7% in S$\alpha$, e$\phi$, respectively.

**Effectiveness of MSFM, DGCM, and EGA.** To evaluate the full version of our model, we tested the performance of Experiment 6. From the results in Table 2, it can be seen that the full version of EGMFNet outperforms all other ablation versions.

Practically, the MSFM gains on COD10K/NC4K (e.g., $+F_\beta^\omega$ and $\downarrow$ MAE in Nos. 1.2–2) indicate better small-object sensitivity and background suppression. The DGCM increments (No. 3 and with MSFM in No. 4) reflect improved context reasoning under clutter. The EGA boosts (No. 5 and full model No. 6) translate into sharper, more complete boundaries under occlusion—an interpretable effect that aligns with the qualitative comparisons.

## 5 Conclusion

This paper proposes a novel camouflage target detection method based on EGMFNet, which integrates multi-scale features and edge features. A multi-scale feature fusion module based on a multi-scale grouped dilated convolution and hypergraph convolution (MSFM) is introduced. The paper utilizes a context-aware module (DGCM), which leverages global contextual information in the fused features. Building on this, an effective edge-guided attention module (EGA) is proposed. This fusion module, while considering the valuable fused features provided by the previous two modules, introduces edge features into the network for learning. It effectively addresses the weak boundary problem in camouflage target detection. Extensive experimental results on four benchmark datasets demonstrate that the proposed method outperforms state-of-the-art approaches. Limitations include sensitivity to very small targets in extreme clutter and runtime constraints under high-resolution inputs. Future work will explore video-level COD (temporal cues), self-/semi-supervised learning for label efficiency, and lightweight backbones for real-time deployment.

**Acknowledgement:** Thanks for the support from my teachers and friends during the writing of this thesis.

**Funding Statement:** This work was financially supported by Chongqing University of Technology Graduate Innovation Foundation (Grant No. gzlcx20253267).

**Author Contributions:** The authors confirm contribution to the paper as follows: Tianze Yu: Designing methodologies, crafting network modules, coding, thesis composition and dataset analysis, inference result uploading. Jianxun Zhang: Guides the work and analyzes the theoretical nature of the module. Hongji Chen: Review, supervision, and collect recent papers. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** Not applicable.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Fan DP, Ji GP, Cheng MM, Shao L. Concealed object detection. IEEE Trans Patt Anal Mach Intell. 2022;44(10):6024–42. doi:10.1109/tpami.2021.3085766.

2. Fan DP, Ji GP, Sun G, Cheng MM, Shen J, Shao L. Camouflaged object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 2774–84.
3. Lindeberg T. Scale-space theory in computer vision. Vol. 2. Berlin/Heidelberg, Germany: Springer Science & Business Media; 1994.
4. Lindeberg T. Feature detection with automatic scale selection. Int J Comput Vis. 1998;2(2):2. doi:10.1023/a:1008045108935.
5. Witkin A. Scale-space filtering: a new approach to multiscale description. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 1984 Mar 19–21; San Diego, CA, USA.
6. Yan J, Le TN, Nguyen KD, Tran MT, Do TT, Nguyen TV. MirrorNet: bio-inspired camouflaged object segmentation. IEEE Access. 2021;9:43290–300. doi:10.1109/access.2021.3064443.
7. Chen G, Liu SJ, Sun YJ, Ji GP, Wu YF, Zhou T. Camouflaged object detection via context-aware cross-level fusion. IEEE Trans Circ Syst Video Technol. 2022;32(10):6981–93. doi:10.1109/tcsvt.2022.3178173.
8. Yao S, Sun H, Xiang TZ, Wang X, Cao X. Hierarchical graph interaction transformer with dynamic token clustering for camouflaged object detection. IEEE Trans Image Process. 2024;33:5936–48. doi:10.1109/tip.2024.3475219.
9. Chen Z, Zhang X, Xiang TZ, Tai Y. Adaptive guidance learning for camouflaged object detection. arXiv:2405.02824. 2024.
10. Sun Y, Wang S, Chen C, Xiang TZ. Boundary-guided camouflaged object detection. arXiv:2207.00794. 2022.
11. Zhu H, Li P, Xie H, Yan X, Liang D, Chen D, et al. I can find you! Boundary-guided separated attention network for camouflaged object detection. Proc AAAI Conf Artif Intell. 2022;36:3608–16. doi:10.1609/aaai.v36i3.20273.
12. Sun D, Jiang S, Qi L. Edge-aware mirror network for camouflaged object detection. In: Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME); 2023 Jul 10–14; Brisbane, Australia. p. 2465–70.
13. Zhang J, Zhang R, Shi Y, Cao Z, Liu N, Khan FS. Learning camouflaged object detection from noisy pseudo label. arXiv:2407.13157. 2024.
14. Lai X, Yang Z, Hu J, Zhang S, Cao L, Jiang G, et al. Camoteacher: dual-rotation consistency learning for semi-supervised camouflaged object detection. arXiv:2408.08050. 2024.
15. Luo Z, Mishra A, Achkar A, Eichel J, Li S, Jodoin PM. Non-local deep features for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 6609–17.
16. Feng M, Lu H, Ding E. Attentive feedback network for boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 1623–32.
17. Xu Y, Zhao L, Cao S, Feng S. Dual priors network for RGB-D salient object detection. In: Proceedings of the 2022 IEEE International Conference on Big Data (Big Data); 2022 Dec 17–20; Osaka, Japan. p. 4201–9.
18. Zhao JX, Liu JJ, Fan DP, Cao Y, Yang J, Cheng MM. Egnet: edge guidance network for salient object detection. In: Proceedings of the International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8779–88.
19. Zhang L, Zhang Q. Salient object detection with edge-guided learning and specific aggregation. IEEE Trans Circ Syst Video Technol. 2024;34(1):534–48. doi:10.1109/tcsvt.2023.3287167.
20. Luo H, Liang B. Semantic-edge interactive network for salient object detection in optical remote sensing images. IEEE J Select Topics Appl Earth Observat Remote Sens. 2023;16:6980–94. doi:10.1109/jstars.2023.3298512.
21. Bi H, Yang L, Zhu H, Lu D, Jiang J. STEG-Net: spatiotemporal edge guidance network for video salient object detection. IEEE Trans Cognit Develop Syst. 2022;14(3):902–15. doi:10.1109/tcds.2021.3078824.
22. Xu B, Liang H, Liang R, Chen P. Synthesize boundaries: a boundary-aware self-consistent framework for weakly supervised salient object detection. IEEE Trans Multim. 2024;26:4194–205. doi:10.1109/tmm.2023.3321393.
23. Zamani L, Azmi R. TriMAE: fashion visual search with triplet masked auto encoder vision transformer. In: Proceedings of the 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE); 2024 Nov 19–20; Mashhad, Iran, Islamic Republic of. p. 338–42.

24. Pang Y, Zhao X, Xiang TZ, Zhang L, Lu H. ZoomNeXt: a unified collaborative pyramid network for camouflaged object detection. IEEE Trans Pattern Anal Mach Intell. 2024;46(12):9205–20. doi:10.1109/tpami.2024.3417329.
25. Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. Pvt v2: improved baselines with pyramid vision transformer. Comput Vis Media. 2022;5(7):8. doi:10.1007/s41095-022-0274-8.
26. Xu G. Multi-feature fusion network for infrared and visible image fusion. In: Proceedings of the 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL); 2025 May 23–25; Ningbo, China. p. 562–5.
27. Li L, Chen Z, Dai L, Li R, Sheng B. MA-MFCNet: mixed attention-based multi-scale feature calibration network for image dehazing. IEEE Trans Emerg Topics Computat Intell. 2024;8(5):3408–21. doi:10.1109/tetci.2024.3382233.
28. Lu H, Zou Y, Zhang Z, Wang Y. Adaptive bidirectional asymptotic feature pyramid network for object detection. In: Proceedings of the 2024 4th International Conference on Robotics, Automation and Intelligent Control (ICRAIC); 2024 Dec 6–9; Changsha, China. p. 562–7.
29. Gao T, Zhang Z, Zhang Y, Liu H, Yin K, Xu C, et al. BHViT: binarized hybrid vision transformer. arXiv:2503.02394. 2025.
30. Zhou Y, Jiang Y, Yang Z, Li X, Sun W, Zhen H, et al. UAV image detection based on multi-scale spatial attention mechanism with hybrid dilated convolution. In: Proceedings of the 2024 3rd International Conference on Image Processing and Media Computing (ICIPMC); 2024 May 17–19; Hefei, China. p. 279–84.
31. Lei M, Li S, Wu Y, Hu H, Zhou Y, Zheng X, et al. YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv:2506.17733. 2025.
32. Yin N, Feng F, Luo Z, Zhang X, Wang W, Luo X, et al. Dynamic hypergraph convolutional network. In: Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE); 2022 May 9–12; Kuala Lumpur, Malaysia. p. 1621–34.
33. Xu R, Liu G, Wang Y, Zhang X, Zheng K, Zhou X. Adaptive hypergraph network for trust prediction. In: Proceedings of the 2024 IEEE 40th International Conference on Data Engineering (ICDE); 2024 May 13–16; Utrecht, The Netherlands. p. 2986–99.
34. Sun W, Liu BD. ESinGAN: enhanced single-image GAN using pixel attention mechanism for image super-resolution. In: Proceedings of the 2020 15th IEEE International Conference on Signal Processing (ICSP); 2020 Dec 6–9; Beijing, China. p. 181–6.
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2024 Dec 4–9; Long Beach, CA, USA.
36. Yu Y, Zhang Y, Zhu X, Cheng Z, Song Z, Tang C. Spatial global context attention for convolutional neural networks: an efficient method. In: Proceedings of the 2024 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC); 2025 Aug 19–22; Bali, Indonesia. p. 1–6.
37. Timbi Z, Yin Y, Yang F. MSCAM-Net: a multi-scale and context-aware model with adaptive fusion module for robust skin lesion segmentation. In: Proceedings of the 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI); 2025 Apr 14–17; Houston, TX, USA. p. 1–4.
38. Chen S, Tan X, Wang B, Hu X. Reverse attention for salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 234–50.
39. Woo S, Park J, Lee JY, Kweon IS. Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 3–19.
40. Le TN, Nguyen TV, Nie Z, Tran MT, Sugimoto A. Anabranch network for camouflaged object segmentation. Comput Vis Image Underst. 2019;184:45–56. doi:10.1016/j.cviu.2019.04.006.
41. Skurowski P, Abdulameer H, Blaszczyk J, Depta T, Kornacki A, Kozioł P. Animal camouflage analysis: chameleon database. 2017 [cited 2025 Nov 1]. Available from: https://www.polsl.pl/rau6/chameleon-database-animal-camouflage-analysis/.
42. Lyu Y, Zhang J, Dai Y, Li A, Liu B, Barnes N, et al. Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA.

43. Ji GP, Fan DP, Chou YC, Dai D, Liniger A, Van Gool L. Deep gradient learning for efficient camouflaged object detection. Mach Intell Res. 2023;20(1):92–108. doi:10.1007/s11633-022-1365-9.
44. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada.
45. Yin B, Zhang X, Fan DP, Jiao S, Cheng MM, Van Gool L, et al. CamoFormer: masked separable attention for camouflaged object detection. IEEE Trans Patt Anal Mach Intell. 2024;46(12):10362–74. doi:10.1109/tpami.2024.3438565.
46. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 7–8.
47. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 7–8.
48. Pang Y. PySODEvalToolkit: a Python-based evaluation toolbox for salient object detection and camouflaged object detection. 2020 [cited 2025 Nov 1]. Available from: https://github.com/lartpang/PySODEvalToolkit.
49. Pang Y. PySODMetrics: a simple and efficient implementation of SOD metrics. 2020 [cited 2025 Nov 1]. Available from: https://github.com/lartpang/PySODMetrics.
50. Pang Y, Zhao X, Xiang TZ, Zhang L, Lu H. Zoom in and out: a mixed-scale triplet network for camouflaged object detection. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 2150–60.
51. Xing H, Wang Y, Wei X, Tang H, Gao S, Zhang W. Go closer to see better: camouflaged object detection via object area amplification and figure-ground conversion. IEEE Trans Circuits Syst Video Technol. 2023;33(10):5444–57. doi:10.1109/tcsvt.2023.3255304.