HDPNet: Hourglass Vision Transformer with Dual-Path Feature Pyramid for Camouflaged Object Detection

Jinpeng He[1], Biyuan Liu[1], Huaixin Chen[1,*]

[1]University of Electronic Science and Technology of China, Chengdu, China

jphe@std.uestc.edu.cn, byliu90@outlook.com, huaixinchen@uestc.edu.cn

Abstract

Existing camouflaged object detection methods often struggle with detecting small objects and fine object boundaries. To alleviate these issues, we propose a novel hourglass vision Transformer with Dual-path Feature Pyramid (HDPNet). Specifically, we construct an hourglass Transformer encoder that effectively captures the global semantic cues while extracting detailed feature maps at various scales, preserving the spatial details and fine-grained boundaries of the camouflaged object. To ensure the preservation of essential cues of hourglass features, we introduce a dual-path feature pyramid decoder (DPFD). This decoder performs coarse-to-fine feature fusion laterally, mitigating the dilution of essential feature cues caused by the semantic gaps. In addition, to further facilitate the local feature modeling in the encoder to mine the correlation between local features and global semantic cues from the camouflaged region, we design a feature interaction enhancement module (FIEM). This module adopts a symmetric structure enables detailed appearance features and global semantic features to complement each other, enhancing the model's ability to capture a wide range of fine-grained details. Extensive quantitative and qualitative experiments demonstrate that the proposed model significantly outperforms 25 existing methods across three challenging COD benchmark datasets, particularly excelling in the detection of small objects and fine boundaries. The code is available at https://github.com/LittleGrey-hjp/HDPNet.

1. Introduction

Camouflage is a vital survival skill evolved by organisms in nature. By changing their colors, textures, and other characteristics, organisms can blend into the environment to deceive predators or capture prey more effectively. Such ability is not only prevalent in the biological world, but is also widely used in military camouflage and stealth techniques in human society. Camouflage object detection [17] aims to identify objects that closely resemble their surroundings and has been widely applied in various fields, such as polyp segmentation [3], industrial defect detection [24, 39], and military object detection [19].

The high similarity between camouflaged objects and their surroundings makes accurate detection challenging, particularly when dealing with small objects or those with fine boundaries, which poses a severe challenge for camouflaged target detection (COD). Recently, the development of deep learning has significantly improved COD performance, particularly following the introduction of large-scale datasets like COD10K [8] and NC4K [25], leading to the emergence of numerous CNN- and Transformer-based methods. However, as shown in Fig. 1, even the recently proposed state-of-the-art methods (CNN-based DGNet [12] and Transformer-based FSPNet [11]) still face challenges in accurately segmenting small objects and the precise shapes of targets in complex scenes.

Unfortunately, CNN-based methods are constrained by the inherent limitations of CNNs in extracting features within a local receptive field, making it difficult to capture long-range dependencies and global feature relations of camouflaged objects [4, 11, 36]. Consequently, they often provide predictions of incomplete object regions, especially in cases involving multiple objects, large objects, and occlusions [41]. Additionally, some CNN-based methods [7, 43], tend to emphasize high-level features in deep layers while overlooking the importance of low-level features in shallow layers on the segmentation, resulting in poor performance on small objects and camouflaged objects with fine structures. In contrast to CNNs, Vision Transformer (ViT) [4] utilize self-attention mechanisms to efficiently model long-range dependencies, and thus several Transformer-based methods [21, 36, 47] have shown superior performance. However, ViT divides images into uniform patches with a relatively low-resolution single output, and its attention mechanism focuses primarily on global information, leading to inadequate perception of local details. For instance, although FSPNet employs adjacent interactions during the decoding stage to progressively restore fea-

Figure 1. Visual comparison between our HPDNet and recent state-of-the-art (SOTA) camouflaged object detection methods (DGNet [12] and FSPNet [11]) for small object and object delicate boundary detection. Our HDPNet generates highly detailed object structures.

ture map sizes, it still struggles to recover and prevent the loss of nuanced information caused by downsampling in the encoding process, which leads to poor segmentation performance on small objects and minute details.

In response to the aforementioned analysis, we propose a novel dual-path feature pyramid network based on an advanced hourglass vision Transformer, named HDPNet. Specifically, to efficiently capture the global information of an image while extracting fine feature maps at different scales, we propose an hourglass vision Transformer encoder, a structure that helps to capture and consolidate information across multiple scales of the image, preserving the detailed spatial information of the objects in deep layers enriched with global semantic cues. For the hourglass structure, we designed a dual-path feature pyramid decoder (DFPD) that initiates multiscale feature fusion from the intermediate layers of the backbone, performing top-down and bottom-up fusion for shallow and deep features, respectively. This ensures information is sufficiently leveraged from both low-level and high-level features, avoiding the introduction of noise and the dilution of critical feature cues caused by the semantic gap. Furthermore, to explore the high-order relations between local features and global semantic cues in camouflage regions to bridge semantic gaps, we introduce a feature interaction enhancement module (FIEM). It is equipped with a non-local and graph convolutional network (GCN) in a symmetric structure, allowing detailed appearance features and global semantic cues to complement each other and produce more refined detection results. In summary, our main contributions are as follows:

- We propose a novel hourglass vision Transformer (HVT) that effectively captures global semantic cues while preserving refined local features, greatly improving the discriminative feature learning of small and finely detailed camouflaged objects in disturbing background.

- We present a dual-path pyramid decoder that sufficiently harness both low-level and high-level features from the image by separately aggregating shallow and deep features laterally, mitigating the dilution of essen-

tial feature cues.

- We design a feature interaction enhancement module (FIEM) to reduce the redundancies and bridge semantic gap between detailed appearance features and global semantic cues, which adopts a symmetric structure with self-attention and GCN, preserving the task-independent interrelationships.

- Comprehensive experiments demonstrate that our proposed HDPNet outperforms 15 CNN- and 10 Transformer-based advanced methods across three COD benchmark datasets. Additionally, the qualitative results highlight HDPNet's superiority in detecting small objects and fine object structures.

2. Related work

2.1. CNN-based Camouflaged Object Detection

Recently, CNNs have excelled in various computer vision tasks and have also made significant strides in camouflaged object detection. Some studies [7, 8, 27] draw inspiration from predator behavior in nature, achieving accurate prediction by mimicking the "search-and-recognize" process of predators. Additionally, multitask learning frameworks are commonly employed in COD to enhance the performance of camouflaged object detection by collaboratively addressing different tasks to extract richer information. These methods usually incorporate some auxiliary tasks such as localization [25], edge detection [34, 42], texture analysis [32], and gradient generation [12]. Some studies, such as SgeMar [13] and HitNet [10], have introduced iterative methods into the COD task, refining the network through multiple iterations to capture minute details and achieve precise segmentation of camouflaged objects with complex structures. Additionally, unlike methods that rely solely on RGB domain information for camouflaged object detection, some researchers have begun combining depth information [38] or frequency domain information [44] with RGB information to achieve more precise detection of camouflaged objects by leveraging multi-source information. Despite these advances, CNN-based methods still face chal-

lenges in modeling long-range dependencies and capturing the precise shapes of targets in complex scenes.

2.2. Transformer-based Camouflaged Object Detection

Compared to CNNs, Transformers offer enhanced parallel processing and better global context encoding, broadening their application across various vision tasks. Recently, inspired by the success of ViT [4], Transformer-based models are becoming a new trend in COD. UGTR [40] combines the advantages of Bayesian learning and Transformer-based inference, utilizing a probabilistic representation model within the Transformer framework to learn the uncertainty of camouflaged objects, which allows the model to focus more effectively on uncertain regions. DTINet [23] introduced a dual-task interactive transformer that enables the model to share information across different tasks. This allows for the simultaneous segmentation of camouflaged objects and their fine boundaries, improving the detection and segmentation performance of model. CamoFormer [41] utilized different attention heads to separately process foreground and background regions to improve segmentation accuracy, while retaining a set of standard attention head to establish global interaction. TPR-Net [43] proposed a transformer-induced progressive refinement mechanism to refine the feature maps layer by layer, harnessing the semantic information of high-level features to guide the detection of camouflaged objects. To complement the modeling of local characteristics in the transformer encoder, FSPNet [11] adopted a non-local mechanism to interact with adjacent similar tokens and explores graph-based high-level relations within the tokens to enhance local representations. Despite their advantages, Transformer-based methods still face the drawback of inadequate perception of local details and the loss of nuanced information due to low-resolution feature map, resulting in poor segmentation performance for small objects and fine boundaries. To this end, we propose an hourglass vision Transformer that is able to obtain high-level semantic while preserving refined object details.

3. Overview

Fig. 2 illustrates the overall architecture of our proposed HDPNet model, which comprises the hourglass vision Transformer encoder, the dual-path feature pyramid decoder, and the feature interaction enhancement module. Given an input image, the Hourglass Transformer encoder employs multi-head self-attention mechanism to build global context model and extract fine feature maps at various scales. In the decoding process, we intentionally designed a dual-path feature pyramid decoder to aggregate and decode low-level edge features and high-level semantic features extracted by the Hourglass Transformer model,

respectively. To better explore and utilize the higher-order relationships between high-level cues and local features, we designed a feature interaction enhancement module, which can reduce the redundancies and bridge semantic gap.

3.1. Hourglass transformer encoder

ViT splits the image into patches and processes them sequentially at a single resolution, potentially failing to capture sufficient feature information for small objects. Due to the lack of necessary contexts and details, it is ineffective in identifying and localizing small targets. Inspired by Hourglass [28], we design an hourglass vision Transformer (HVT) that controls the scale of feature maps through a patch embedding layer, enabling it to generate multi-scale feature maps for targets of different scales and gradually recover the spatial information lost during downsampling in forward propagation. Fig. 2 presents an overview of HVT and it consists of six stages, each with a similar structure containing a patch embedding layer and N transformer layers. Specifically, for the input feature map $R^{C \times W \times H}$, in the patch embedding layer, we employ convolution operation to divide it into HW/k² patches to reduce the spatial resolution of the feature map, or adopt pixel-shuffle to restore the feature map resolution. Then the feature map patches are linearly projected into a 1D sequence of token embeddings $T \in R^{HW/k^2 \times C1}$ (downsampling) or $T \in R^{HWk^2 \times C2}$ (upsampling), where C, H, and W represent the channel size, height, and width of the feature map respectively and k denotes the scale factor. The embedded patches along with positional embeddings are passed through the transformer layers, where each layer contains a multi-head self-attention (MSA) and a multi-layer perceptron (MLP) block, and the output is reshaped into a feature map of size H/K × W/K × C1 or HK × WK × C2. Additionally, residual connections are applied to incorporate features from the corresponding scale of the encoding layer to retain local details and spatial information, supplement the information loss for the upsampling, and promote the reconstruction capability of the model, as shown in the red dashed line in Fig. 2. Finally, the outputs of the encoder are provided by the last two Transform layers of each stage. Thus, the feature maps from each stage of the backbone can be denoted as $\{F_{1,1}, F_{1,2}, ..., F_{6,1}, F_{6,2}\}$, where $\{F_{1,1}, ..., F_{3,2}\}$ and $\{F_{4,1}, ..., F_{6,2}\}$ are input into the low- and high-level decoding paths, respectively. The topology of the hourglass is symmetric.

3.2. Dual-path feature pyramid decoder

Typically, multi-scale methods [7, 18, 43] favor transferring the semantic information from deep features to shallow features, ignoring spatial information to some extent. To this end, considering the unique structure of the Hourglass Transformer model, we designed a dual-path feature pyramid decoder (DFPD) to stepwise aggregate low-level
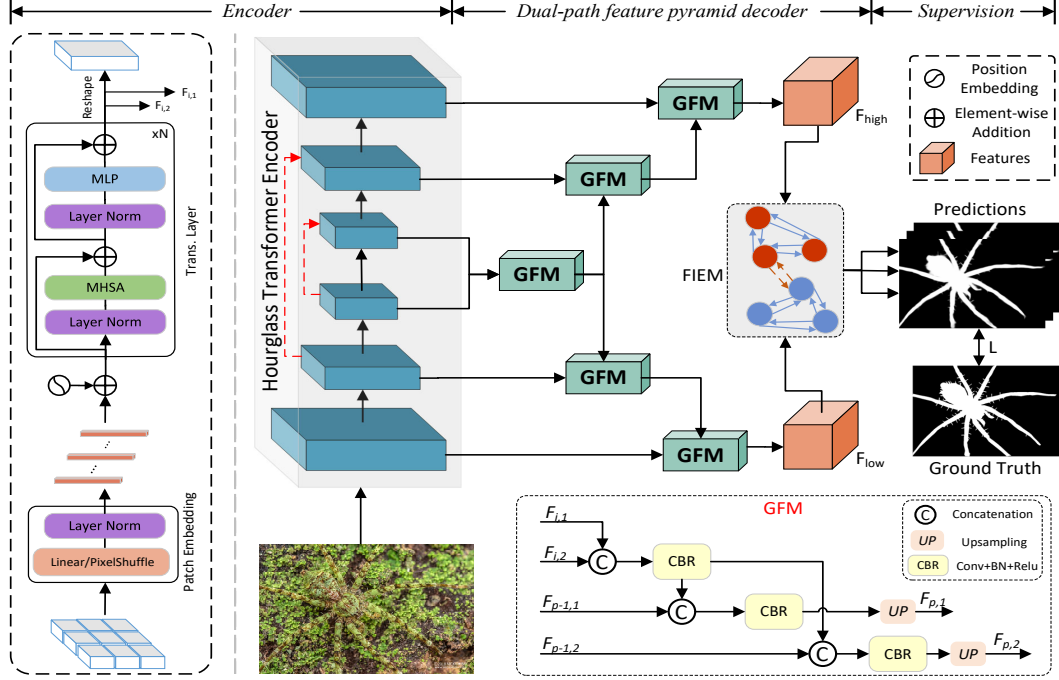
Figure 2. Overall architecture of the proposed HDPNet. It consists of three key components: hourglass vision Transformer encoder (HVT), dual-path feature pyramid decoder (DFPD) and feature interaction enhancement module (FIEM). MHSA denotes Multi-Head Self-Attention, and MLP denotes Multi-Layer Perceptron.

features with details and high-level features with semantics, respectively, better preserving important cues and avoiding the noise introduced by direct feature fusion and the dilution of weak feature cues. In our DFPD decoder, we propose a grouped fusion module (GFM) that performs multi-level fusion of features from the backbone at the current stage and from the GFM of the previous stage (if present), and then passes the aggregated features to the next GFM module. Specifically, as shown in Fig. 2, the decoder builds top-down and bottom-up feature flows from the intermediate layers of the backbone, facilitating dual-path feature fusion. Taking bottom-up feature fusion as an example, the decoder initially receives feature maps ($\{F_{3,1}, F_{3,2}, F_{4,1}, F_{4,2}\}$) from different layers of the encoder as input to the GFM. After integrating these feature maps according to specific rules, the GFM guides them towards higher resolution and passes the outputs to the next-stage GFM. Simultaneously, the next stage GFM receives feature maps from the same scale of the encoder ($\{F_{5,1}, F_{5,2}\}$). Eventually, feature maps at half the original size are obtained after three stages of GFM processing. Notably, the top-down and bottom-up processes are consistent. The details of GFM is shown in Fig. 2 and the process can be expressed as follows:

$$F_{p,1} = \text{Up}(\text{CBR}(\text{cat}(\text{CBR}(\text{cat}(F_{i,1}, F_{i,2})), F_{p-1,1}))) \quad (1)$$

$$F_{p,2} = \text{Up}(\text{CBR}(\text{cat}(\text{CBR}(\text{cat}(F_{i,1}, F_{i,2})), F_{p-1,2}))), \quad (2)$$

Where cat represents channel concatenation, CBR denotes Convolution, Batch Normalization, and ReLU activation, and Up represents 2x upsampling. $F_{i,1}$ and $F_{i,2}$ represent the feature pairs at the current stage, $F_{p-1,1}$ and $F_{p-1,2}$ are the the output aggregated feature from the previous GFM or backbone (in first stage), and $F_{p,1}$ and $F_{p,2}$ are the aggregated features output by the current GFM.

3.3. Feature interaction enhancement module

To more effectively explore and utilize the complex relationship between low- and high-level features, we design a feature interaction enhancement module (FIEM), drawing inspiration from [3, 35]. Considering that both high-level and low-level features are definitely crucial for precise localization and segmentation, the FIEM adopts a symmetric structure that maintains equal attention to both deep and shallow features, as shown in Fig. 3. To inject detailed appearance features into high-level semantic features, the high-level features are passed through two linear mapping functions with trainable weights φ_q and φ_k, producing dimension-reduced feature sequences Q and K. The low-level features F_l are processed through a Window Multi-headed Self-Attention module (W-MSA) [22], generating local attention vectors A to refine the model's emphasis on local features. This process can be expressed as $Q = \varphi_q(F_{high})$, $K = \varphi_k(F_{high})$ and $A = \text{W-MSA}(F_{low})$.
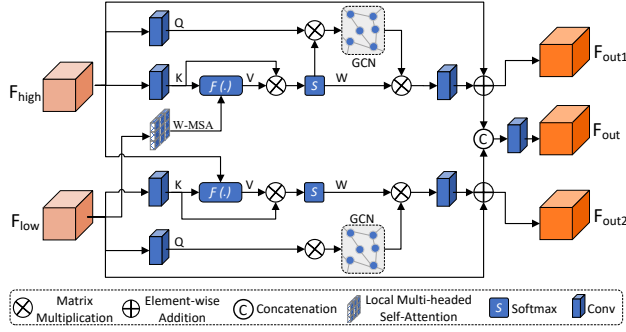
Figure 3. Details of the presented FIEM. It is a symmetrical structure with upper and lower parts, where the upper part extends pixel features of the camouflaged area with high-level semantic cues across the entire region, and the lower part integrates spatial information containing rich details into the overall object structure.

Next, the softmax function is applied to A to generate the weight map and compute the Hadamard product with K to emphasize the edge pixels, and then an adaptive average pooling operation is performed to obtain the feature mapping V. This operation, denoted as $F(.)$ in the Fig. 3, can be expressed as follows:

$$F(\cdot) = \mathcal{P}_{avg}(K \odot \text{softmax}(A)). \quad (3)$$

The matrix product is then applied to K and V to explore the correlation between them, followed by the softmax operation to generate a correlation attention map W:

$$W = \text{softmax}(K^T \otimes V). \quad (4)$$

After obtaining the correlation attention map W, it is multiplied by the feature map Q. The resulting features are then fed into the graph convolutional network (GCN) [35] to learn high-order semantic relationships between regions (sets of pixels with similar features). To reconstruct the graph domain features into the original structural features, the inner product between $GCN(.)$ and G is computed and converted back to the 2D image features of the same dimensions as the original features through a linear mapping function φ_f, which is then combined with the feature F_{high} to obtain the final output of the FIEM. In particular, the number of nodes of GCN is 16. The aforementioned operations correspond to the upper part of the Fig. 3.

$$F_{out1} = \varphi_f(W^T \otimes \text{GCN}(W^T \otimes Q)) + F_{high}. \quad (5)$$

Similarly, to inject contextual information into the shallow detail features, we swap F_{high} and F_{low} and perform the aforementioned operations to obtain F_{out2}. Specifically, F_{high} does not pass through the W-MSA module. These correspond to the lower part of the Fig. 3. Finally, F_{out1} and F_{out2} are further fused to obtain the final output.

$$F_{out} = Conv_{3\times3}(Conv_{1\times1}(cat(F_{out1}, F_{out2}))), \quad (6)$$

where $Conv_{1\times1}$ denotes the convolution with a kernel size of 1, used to adjust the number of channels so that F_{out}, F_{out1}, and F_{out2} have the same number of channels, and $Conv_{3\times3}$ denotes the convolution with a kernel size of 3.

3.4. Loss

We use binary cross-entropy loss (L_{BCE}) to supervise the three layers of output prediction (F_{out}, F_{out1}, and F_{out2}) of FIEM with the following overall loss function:

$$L = \frac{1}{4}\sum_{i=1}^{2} L_{BCE}(F_{outi}, G) + L_{BCE}(F_{out}, G), \quad (7)$$

where G is the ground truth annotation.

4. Experiments and Results

4.1. Experiment Settings

Implementation Details. We use PVTv2 [37], pre-trained on the ImageNet dataset [14], as the prototype for the hourglass transformer encoder, and the other modules are randomly initialized. The transformer layers contained in each stage of the HVT are $\{3, 6, 20, 20, 6, 3\}$, with corresponding channel numbers of $\{64, 128, 320, 320, 128, 64\}$ respectively. For training and testing, all input images are resized to 384×384, and random flipping is applied to augment the training data. We use Adam as the optimizer, with the learning rate initialized to 1e-4 and then scaled down by 10 every 50 epochs. Our model is trained end-to-end on an NVIDIA A100-SXM for 150 epochs with a batch size of 2.
Datasets. We evaluate the proposed method on three widely-used COD datasets i.e., CAMO [15], COD10K [8] and NC4K [25]. CAMO is the first dataset for COD and contains 1000 training images and 250 test images. COD10K is currently the largest and most challenging COD dataset, consisting of 10 superclasses and 78 subclasses, with 3,040 training images and 2,026 testing images. NC4K is a large-scale COD dataset containing 4,121 images for testing, which mainly focus on the natural and artificial camouflage. Following the previous training setup [7, 11], our training set consists of 1000 images from CAMO and 3040 images from COD10k, while other images are used for testing.
Evaluation Metrics. Following [11,17], We adopt six well-known evaluation metrics, including S-measure [5] (S_m), weighted F-measure [26] (F_w), mean F-measure [1] (F_m), mean E-measure [6] (E_m), max E-measure (E_x), and mean absolute error [31] ($\mathcal{M}$). In addition, we plot the precision-recall (PR), F_m-threshold and E_m-threshold curves, which are available in *Supplementary Materials*.

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

To demonstrate the effectiveness of our proposed method, we compared HDPNet with 15 CNN-based SOTA

COD models, including SINet [8], Rank-Net [25], JCOD [16], PFNet [27], R-MGL [42], C2FNet [33], BSA-Net [46], SegMaR [13], BGNet [34], ZoomNet [29], SINetv2 [7], FEDER [9], DINet [45], ZoomNext [30] and DGNet [12], as well as 10 Transformer-based COD methods, including COST [36], VST [20], UGTR [40], ICON-P [47], DTINet [23], TPRNet [43], CamoFormer [41], FPNet [2], EVP [21] and FSPNet [11]. For a fair comparison, the predictions of competitors are either directly provided by the authors or generated by their well-pretrained models based on open source code. The quantitative comparisons of all methods are summarized in Tab. 1 and the qualitative performance is shown in Figs. 4 and 5. More experimental results are provided in the *Supplementary Material*.

Quantitative Results. Tab. 1 summarizes the quantitative results of our proposed method on three challenging COD benchmark datasets versus 25 competitors under six commonly used assessment metrics. As can be seen from the results, compared to the SOTA CNN-based methods, our method achieves average performance gains of 6.57%, 9.38%, 11.09%, 3.14%, 3.64% and 33.39% over DGNet and 4.51%, 5.83%, 4.11%, 3.71%, 3.86% and 27.72% over ZoomNext in terms of S_m, F_ω, F_m, E_m, E_x and $\mathcal{M}$ on these three datasets, respectively. Meanwhile, compared to the cutting-edge transformer-based methods, our method shows significant performance improvements of 4.66%, 8.48%, 6.87%, 2.81%, 2.64% and 29.63% over EVP and 3.76%, 6.23%, 4.92%, 3.11%, 1.93% and 20.07% over FSPNet on average in terms of S_m, F_ω, F_m, E_m, E_x and $\mathcal{M}$, respectively. In all, the proposed HDPNet model consistently outperforms the competitors in detection performance on the experimental datasets, particularly on the COD10K and NC4K datasets, which contain numerous small objects and subtle object boundaries. The superiority in performance is attributed to the preservation the of local features by the hourglass vision Transformer encoder, as well as the dual-path feature decoding and the mutual enhancement and complementarity between features.

Qualitative results. Fig. 4 provides an intuitive comparison of our proposed SOTA method with six top-performing SOTA methods (three CNN- and three Transformer-based) in several typical complex scenarios, including occluded, small, multiple objects, and object border. It can be observed that the compared methods are prone to provide inaccurate object localization, incomplete object regions, or missing objects, resulting in inferior segmentation of camouflaged objects, particularly with weak small targets and subtle boundaries. In contrast, our method is able to capture the entire object area, recover and preserve edge details, providing more accurate, complete, vivid, and exact high-quality segmentation maps of camouflaged objects with excellent robustness.

Comparison of Small Object and Object Border Qual-

ity. As shown in Fig. 4, prevailing COD methods often perform unsatisfactorily when tackling small objects and intricate target boundaries. To further demonstrate the performance of HDPNet on such cases, we present some prediction results in Fig. 5. Compared to the top-performing models DGNet and FSPNet, our model can accurately and completely predict objects and their minute details. In contrast, DGNet and FSPNet exhibit significant deviations in their predictions due to insufficient attention to local and surrounding context information, making them particularly sensitive to small objects and fine structures. These visualizations demonstrate that our model can capture fine-grained cues and segment camouflaged objects with superior accuracy.

4.3. Ablation Study

Overall results. To validate the effectiveness of the hourglass structure and modules proposed in this paper, we conducted an incremental ablation study on benchmark camouflage datasets, and the results are presented in Tab. 2. Our hourglass structure (denoted as H) is based on PVTv2 (denoted as P). The baseline models (P and H) employ only the encoder and convolution for prediction. Then, we incrementally add the proposed DFPD decoder and FIEM module to the baseline model H. Additionally, we incorporated the proposed modules into VIT-Base (denoted as V) to further validate their effectiveness. The overall results of the ablation experiments demonstrate that the hourglass model outperforms the other two baseline models. Furthermore, the proposed DFPD decoder and FIEM module both make significant contributions to the performance of the network.

Dual-path feature pyramid decoder. Next, we analyze the importance of the dual-path feature pyramid decoder (denoted as D). In the table, "H+D" and "V+D" present the results of our decoder when integrated with the HVT backbone and the VIT backbone, respectively. It is evident that the decoder significantly enhances the performance of the baseline model, demonstrating its ability to effectively aggregate and preserve key features from different layers, resulting in more accurate predictions. Specifically, with the addition of the DPFD decoder, the HVT and VIT achieve average performance gains of 3.14%, 3.90%, 3.25%, 4.06%, 3.02%, 22.75% and 3.09%, 6.08%, 5.11%, 3.75%, 2.62%, 25.18% in terms of S_m, F_ω, F_m, E_m, E_x and $\mathcal{M}$ on these three datasets, respectively. Additionally, as shown in the ablation curves and progressive visualizations of modules in Figs. 6 and 7, high- and low-level features focus on different aspects of the target structure. The retention of basic clues of hourglass features by the DFPD mitigates the dilution of fundamental feature clues by the semantic gap, significantly contributing to the improvement of model performance.

Feature Interaction Enhancement Module. We further

Table 1. Quantitative comparison of the proposed method with 25 state-of-the-art methods on three benchmark COD datasets. The best three results are highlighted in red, green and blue. "-": unavailable. "↑": the higher the better, "↓": the lower the better.

Method	CAMO (250)						COD10K (2026)						NC4K(4121)					
	$S_m\uparrow$	$F_\omega\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\omega\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\omega\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$E_x\uparrow$	$\mathcal{M}\downarrow$
CNN-Based Methods																		
SINet[20] [8]	0.751	0.606	0.675	0.771	0.831	0.100	0.771	0.551	0.634	0.806	0.868	0.051	0.808	0.723	0.769	0.871	0.883	0.058
Rank-Net[21] [25]	0.787	0.696	0.744	0.838	0.854	0.080	0.804	0.673	0.715	0.880	0.892	0.037	0.840	0.766	0.804	0.895	0.907	0.048
JCOD[21] [16]	0.800	0.727	0.771	0.856	0.873	0.073	0.808	0.683	0.721	0.884	0.891	0.035	0.840	0.771	0.806	0.898	0.907	0.047
PFNet[21] [27]	0.782	0.695	0.746	0.842	0.855	0.085	0.800	0.660	0.701	0.877	0.890	0.040	0.829	0.745	0.784	0.888	0.898	0.053
R-MGL[21] [42]	0.775	0.673	0.726	0.812	0.842	0.088	0.814	0.666	0.711	0.852	0.890	0.035	0.833	0.740	0.782	0.867	0.893	0.052
C2FNet[21] [33]	0.796	0.719	0.762	0.854	0.864	0.080	0.813	0.686	0.723	0.890	0.900	0.036	0.838	0.762	0.795	0.897	0.904	0.049
BSA-Net[22] [46]	0.794	0.717	0.763	0.851	0.867	0.079	0.818	0.699	0.738	0.891	0.901	0.034	0.841	0.771	0.808	0.897	0.907	0.048
SegMaR[22] [13]	0.815	0.753	0.795	0.874	0.884	0.071	0.833	0.724	0.757	0.899	0.906	0.034	0.841	0.781	0.820	0.896	0.907	0.046
BGNet[22] [34]	0.812	0.749	0.789	0.870	0.882	0.073	0.831	0.722	0.753	0.901	0.911	0.033	0.851	0.788	0.820	0.907	0.916	0.044
ZoomNet[22] [29]	0.820	0.751	0.792	0.876	0.891	0.066	0.838	0.727	0.764	0.886	0.909	0.029	0.853	0.783	0.816	0.895	0.912	0.044
SINetv2[22] [7]	0.820	0.743	0.782	0.882	0.895	0.070	0.815	0.680	0.718	0.887	0.906	0.037	0.847	0.770	0.805	0.903	0.914	0.048
FEDER[23] [9]	0.802	0.738	0.781	0.867	0.870	0.071	0.822	0.716	0.751	0.899	0.900	0.032	0.847	0.789	0.824	0.907	0.915	0.044
DINet[23] [45]	0.821	-	0.790	0.874	0.886	0.068	0.832	-	0.761	0.903	0.914	0.031	0.856	-	0.824	0.909	0.919	0.043
ZoomNeXt[23] [30]	0.833	0.774	0.813	0.875	0.891	0.065	0.861	0.768	0.801	0.906	0.925	0.026	0.874	0.816	0.846	0.913	0.928	0.037
DGNet[23] [12]	0.839	0.769	0.806	0.901	0.916	0.057	0.822	0.728	0.693	0.896	0.911	0.033	0.857	0.784	0.814	0.911	0.922	0.042
Transformer-Based Methods																		
COST[21] [36]	0.813	0.776	-	0.896	-	0.060	0.790	0.693	-	0.901	-	0.035	0.825	0.693	-	0.891	-	0.055
VST[21] [20]	0.787	0.691	0.738	0.838	0.866	0.076	0.781	0.604	0.653	0.837	0.877	0.042	0.831	0.732	0.771	0.877	0.901	0.050
UGTR[21] [40]	0.784	0.684	0.735	0.822	0.851	0.086	0.817	0.666	0.712	0.853	0.890	0.036	0.839	0.747	0.787	0.875	0.899	0.052
ICON-P[22] [47]	0.818	0.630	0.682	0.774	0.797	0.111	0.779	0.641	0.685	0.834	0.846	0.050	0.826	0.742	0.778	0.866	0.878	0.056
DTINet[22] [23]	0.856	0.796	0.823	0.915	0.926	0.051	0.824	0.695	0.726	0.896	0.910	0.034	0.863	0.792	0.818	0.917	0.926	0.406
TPRNet[23] [43]	0.817	0.736	0.778	0.869	0.885	0.074	0.818	0.682	0.723	0.884	0.904	0.037	0.850	0.772	0.807	0.901	0.913	0.048
CamoFormer[23] [41]	0.817	0.759	0.792	0.866	0.885	0.067	0.838	0.721	0.753	0.916	0.930	0.029	0.855	0.788	0.821	0.900	0.914	0.042
FPNet[23] [2]	0.851	0.802	0.836	0.905	0.916	0.057	0.850	0.754	0.781	0.902	0.920	0.029	-	-	-	-	-	-
EVP[23] [21]	0.846	0.771	0.803	0.895	0.915	0.059	0.844	0.728	0.764	0.906	0.927	0.029	0.874	0.802	0.830	0.916	0.934	0.039
FSPNet[23] [11]	0.856	0.799	0.830	0.899	0.928	0.050	0.851	0.735	0.769	0.895	0.930	0.026	0.879	0.816	0.843	0.915	0.937	0.035
HDPNet(Ours)	0.893	0.851	0.870	0.934	0.948	0.040	0.888	0.794	0.820	0.925	0.951	0.020	0.902	0.850	0.871	0.934	0.950	0.029

Figure 4. Visual comparison with some representative SOTA models in challenging scenarios demonstrates the performance and robustness of our method.

Table 2. Ablation studies of HDPNet on benchmark datasets. "P" is PVT backbone, "H" is HVT backbone, "V" is VIT backbone, "D" is DPFD, "F" is FIEM. Specially, the number and size of channels in D and F are adjusted for V.

| Settings	Params.(M)	GFLOPs	CAMO (250)						COD10K (2026)						NC4K(4121)					
			$S_m\uparrow$	$F_\omega\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\omega\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\omega\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$E_x\uparrow$	$M\downarrow$
P	84.35	21.26	0.834	0.781	0.818	0.864	0.889	0.059	0.830	0.717	0.755	0.864	0.912	0.033	0.859	0.808	0.837	0.884	0.914	0.046
H	72.02	36.28	0.842	0.801	0.830	0.876	0.896	0.058	0.848	0.733	0.764	0.873	0.916	0.031	0.865	0.813	0.841	0.892	0.915	0.043
H+D	82.95	92.24	0.880	0.835	0.851	0.920	0.934	0.045	0.871	0.775	0.802	0.909	0.937	0.024	0.884	0.827	0.860	0.919	0.938	0.033
H+D+F	84.51	150.1	**0.893**	**0.851**	**0.870**	**0.934**	**0.948**	**0.040**	**0.888**	**0.794**	**0.820**	**0.925**	**0.951**	**0.020**	**0.902**	**0.850**	**0.871**	**0.934**	**0.950**	**0.029**
V	86.21	51.18	0.811	0.726	0.771	0.847	0.868	0.071	0.816	0.701	0.732	0.871	0.901	0.040	0.853	0.781	0.813	0.901	0.918	0.051
V+D	114.6	204.2	0.852	0.789	0.820	0.900	0.916	0.056	0.837	0.742	0.777	0.900	0.919	0.030	0.867	0.810	0.836	0.916	0.921	0.036
V+D+F	116.9	244.2	0.865	0.815	0.845	0.915	0.930	0.050	0.854	0.748	0.783	0.901	0.930	0.026	0.881	0.824	0.851	0.922	0.939	0.034
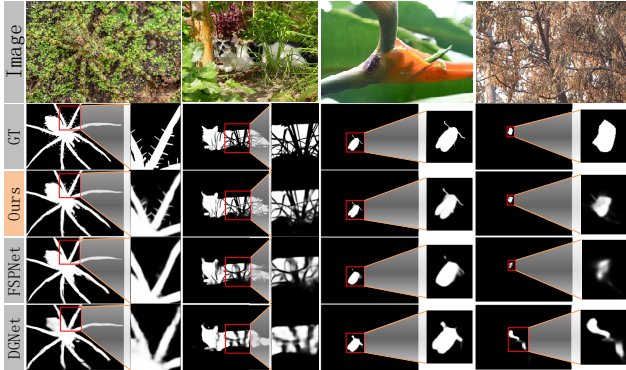
Figure 5. Visual comparison of small object and object border quality with two top-performing models highlights the performance and effectiveness of our method.

Figure 6. Module Performance Ablation Curves of S_m, F_w and $\mathcal{M}$ on the CAMO dataset. The proposed DPFD and FIEM facilitate the model performance to improve gradually.

Figure 7. Progressive visualization of predictions. Please zoom in for details. DFPD low-level output (denoted as F_{low}), DFPD high-level output (denoted as F_{high}), FIEM low-level output (denoted as F_{out2}), FIEM high-level output (denoted as F_{out1}), and FIEM fused output (denoted as F_{out}). Our method comprehensively harnesses the respective strengths of shallow and deep features, achieving complementarity.

investigate the contribution of FIEM (denoted as F). In Tab. 2, 'H+D+F' and 'V+D+F' present the quantitative results of FIEM. The FIEM is capable of mining high-order relationships between low-level and high-level features, enabling the features to complement and enhance each other. By fully leveraging low-level features that contain local details, FIEM significantly boosts detection performance. Its performance is particularly notable on the COD10K and NC4K datasets, which include numerous small objects and intricate fine boundaries.

Progressive visualization. To further intuitively demonstrate the effectiveness of each component mentioned above, we depict the performance ablation curves and visualization results of the model on the CAMO dataset with outputs from different layers. The results are shown in Figs. 6 and 7. Please note that all labels correspond to the symbols shown in Fig. 3. The progressive visualization results indicate that low-level and high-level features focus on different aspects of the target structure, which are fused separately to retain their respective important cues, and the feature interaction enhancement module further explores the correlation between features promoting feature complementarity. Overall, our proposed DPFD and FIEM facilitate the restoration of object details and the gradual generation of clear camouflage predictions, leading to a pro-gressive improvement in model performance.
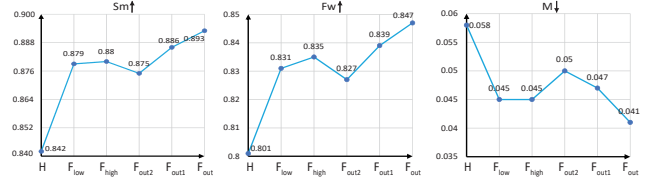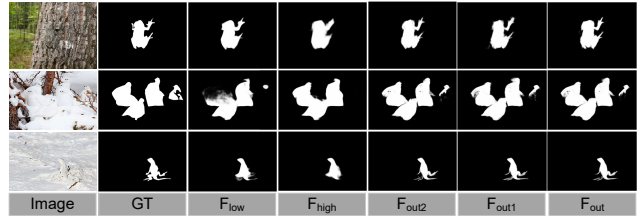
5. Conclusion

Given the poor performance of existing COD methods in small objects with fine object boundaries, in this paper, we propose a novel hourglass vision Transformer-based dual-path feature pyramid network (HDPNet). It consists of an hourglass-shaped Transformer backbone, a dual-path feature pyramid decoder, and a feature interaction enhancement module. Extensive comparative experiments and ablation studies show that the proposed HDPNet outperforms 25 cutting-edge approaches on three widely used COD benchmark datasets, achieving superior performance, especially at small object and fine object boundaries.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009. 5

[2] Runmin Cong, Mengyao Sun, Sanyi Zhang, Xiaofei Zhou, Wei Zhang, and Yao Zhao. Frequency perception network for camouflaged object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1179–1189, 2023. 6, 7

[3] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, 2:9150015, 2023. 1, 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1, 3

[5] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps, 2017. 5

[6] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 698–704. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 5

[7] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, Oct. 2022. 1, 2, 3, 5, 6, 7

[8] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2774–2784, 2020. 1, 2, 5, 6, 7

[9] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22046–22055, 2023. 6, 7

[10] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Ying Tai, Chengjie Wang, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection, 2023. 2

[11] 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. 1, 2, 3, 5, 6, 7

[12] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, Jan. 2023. 1, 2, 6, 7

[13] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4703–4712, 2022. 2, 6, 7

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. 5

[15] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019. 5

[16] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection, 2021. 6, 7

[17] Yanhua Liang, Guihe Qin, Minghui Sun, Xinchao Wang, Jie Yan, and Zhonghan Zhang. A systematic review of image-level camouflaged object detection with deep learning. *Neurocomputing*, 566:127050, 2024. 1, 5

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 3

[19] Maozhen Liu and Xiaoguang Di. Extraordinary mhnet: Military high-level camouflage object detection network and dataset. *Neurocomputing*, 549:126466, 2023. 1

[20] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer, 2021. 6, 7

[21] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations, 2023. 1, 6, 7

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 4

[23] Zhengyi Liu, Zhili Zhang, and Wei Wu. Boosting camouflaged object detection with dual-task interactive transformer, 2022. 3, 6, 7

[24] Qiwu Luo, Ben Li, Jiaojiao Su, Chunhua Yang, Weihua Gui, Olli Silven, and Li Liu. Cddnet: Camouflaged defect detection network for steel surface. *IEEE Transactions on Instrumentation and Measurement*, 2023. 1

[25] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects, 2021. 1, 2, 5, 6, 7

[26] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014. 5

[27] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining, 2021. 2, 6, 7

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. 3

[29] 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, 2022. 6, 7

[30] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6, 7

[31] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012. 5

[32] Jingjing Ren, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Yangyang Xu, Weiming Wang, Zijun Deng, and Pheng-Ann Heng. Deep texture-aware features for camouflaged object detection, 2021. 2

[33] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection, 2021. 6, 7

[34] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection, 2022. 2, 6, 7

[35] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing, 2020. 4, 5

[36] Haiwen Wang, Xinzhou Wang, Fuchun Sun, and Yixu Song. Camouflaged object segmentation with transformer. In Fuchun Sun, Dewen Hu, Stefan Wermter, Lei Yang, Huaping Liu, and Bin Fang, editors, *Cognitive Systems and Information Processing*, pages 225–237, Singapore, 2022. Springer Nature Singapore. 1, 6, 7

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 5

[38] Mochu Xiang, Jing Zhang, Yunqiu Lv, Aixuan Li, Yiran Zhong, and Yuchao Dai. Exploring depth contribution for camouflaged object detection, 2022. 2

[39] Yu-Jie Xiong, Yong-Bin Gao, Hong Wu, and Yao Yao. Attention u-net with feature fusion module for robust defect detection. *Journal of Circuits, Systems and Computers*, 30(15):2150272, 2021. 1

[40] 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. 3, 6, 7

[41] Bowen Yin, Xuying Zhang, Qibin Hou, Bo-Yuan Sun, Deng-Ping Fan, and Luc Van Gool. Camoformer: Masked separable attention for camouflaged object detection, 2022. 1, 3, 6, 7

[42] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection, 2021. 2, 6, 7

[43] Qiao Zhang, Yanliang Ge, Cong Zhang, and Hongbo Bi. Tpr-net: camouflaged object detection via transformer-induced progressive refinement network. *The Visual Computer*, 39(10):4593–4607, Oct 2023. 1, 3, 6, 7

[44] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4494–4503, 2022. 2

[45] Xiaofei Zhou, Zhicong Wu, and Runmin Cong. Decoupling and integration network for camouflaged object detection. *IEEE Transactions on Multimedia*, 2024. 6, 7

[46] Hongwei Zhu, Peng Li, Haoran Xie, Xuefeng Yan, Dong Liang, Dapeng Chen, Mingqiang Wei, and Jing Qin. I can find you! boundary-guided separated attention network for camouflaged object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3608–3616, Jun. 2022. 6, 7

[47] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning, 2022. 1, 6, 7