ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887964

Camouflaged Object Detection with CNN-Transformer Harmonization and Calibration

Yilin Zhao
College of Information Engineering
Shanghai Institute of Technology
Shanghai, China
zhaoyilin0726@163.com

Qing Zhang*
College of Information Engineering
Shanghai Institute of Technology
Shanghai, China
zhangqing0329@gmail.com

Yuetong Li
College of Information Engineering
Shanghai Institute of Technology
Shanghai, China
iuueong@gmail.com

***Abstract*—Camouflaged object detection (COD) aims to segment objects that visually blend into their surroundings. However, the subtle differences between camouflaged objects and the background make this task highly challenging. Therefore, how to represent and learn local details and global contexts is crucial for improving detection performance. In this paper, we propose a novel COD network which synergistically leverages the distinct but complementary local and global knowledge to capture the camouflaged objects and identify imperceptible boundaries. Specifically, we design a Feature Coherence Harmonization module to integrate intra-layer features by bridging the knowledge gap between convolutional neural network (CNN) features, which focus on local patterns, and Transformer features, which capture global relationships. Furthermore, we propose a Cross-layer Feature Calibration Module that adaptively aligns inter-layer features, progressively aggregating diverse information to achieve an accurate prediction. Experimental results on COD benchmark datasets demonstrate that the proposed network significantly outperforms state-of-the-art approaches.**

***Index Terms*—Camouflaged object detection, intra-layer integration, inter-layer aggregation.**

I. Introduction

In nature, many animals blend seamlessly into their surroundings by altering their shapes, colors, or patterns to avoid predators. Camouflaged object detection (COD) aims to automatically identify these objects within their background. COD is essential for various practical applications, such as medical image segmentation [10], industrial defect detection [2], and pest monitoring [20] in agriculture.

Due to the high similarity between camouflaged objects and their backgrounds, COD is significantly more challenging than other generic object detection, such as salient object detection and semantic segmentation. Compared to the traditional approaches using handcrafted features, CNN-based models for COD have achieved promising results. For instance, Zhai *et al.* [24] decouple the image into two task-specific feature maps for coarse localization and boundary detection, utilizing a graph structure for mutual learning. Mei *et al.* [17] introduce a positioning module and a focusing module, which refine the target features with the help of coarse prediction features. He *et al.* [11] propose learning an auxiliary edge reconstruction task to facilitate precise segmentation. Despite their strong performance in capturing local features, these methods struggle with handling global contexts and long-range dependencies, which makes it difficult to effectively tackle COD tasks in complex backgrounds.

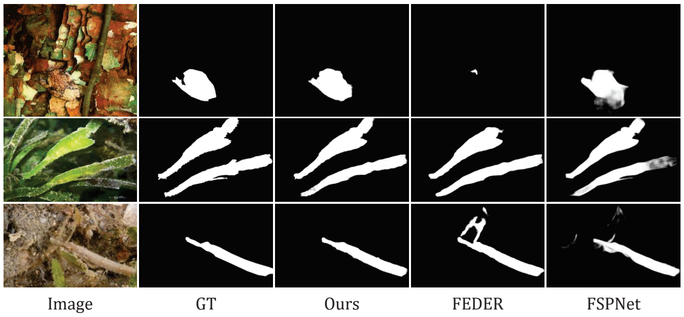

Fig. 1: Visual predictions of different methods. The Transformer-based method FSPNet [12] struggles to identify fine details, while the CNN-based method FEDER [11] face challenges in capturing global semantic, leading to inaccurate or incomplete segmentation.

To address this issue, Transformer [5], which can model long-range dependencies through self-attention mechanisms, have been introduced to improve COD performance. For instance, Cong *et al.* [4] propose a frequency-domain perception network that achieves both coarse localization and detailed correction of camouflaged objects through a two-stage model. Huang *et al.* [12] enhance neighboring feature interactions through feature shrinkage and graph-based relationship modeling. Although Transformer-based approaches have made remarkable progress, they exhibit limitations in capturing fine-grained local features, leading to challenges in accurately determining the subtle boundary details of camouflaged objects, as demonstrated in Fig.1.

Inspired by the above observations, we propose a novel Feature Harmonization and Calibration network, called FHCNet, to exploit the complementary advantages of CNNs and Transformers to compensate each other for accurate predictions. Specifically, we propose the Feature Coherence Harmonization (FCH) module, which aligns intra-layer CNN features and Transformer features, both exhibiting distinct properties. This allows the network to reduce semantic gaps, facilitating effective integration. Subsequently, we design the Cross-layer

This work is supported by Natural Science Foundation of Shanghai under Grant No. 21ZR1462600.

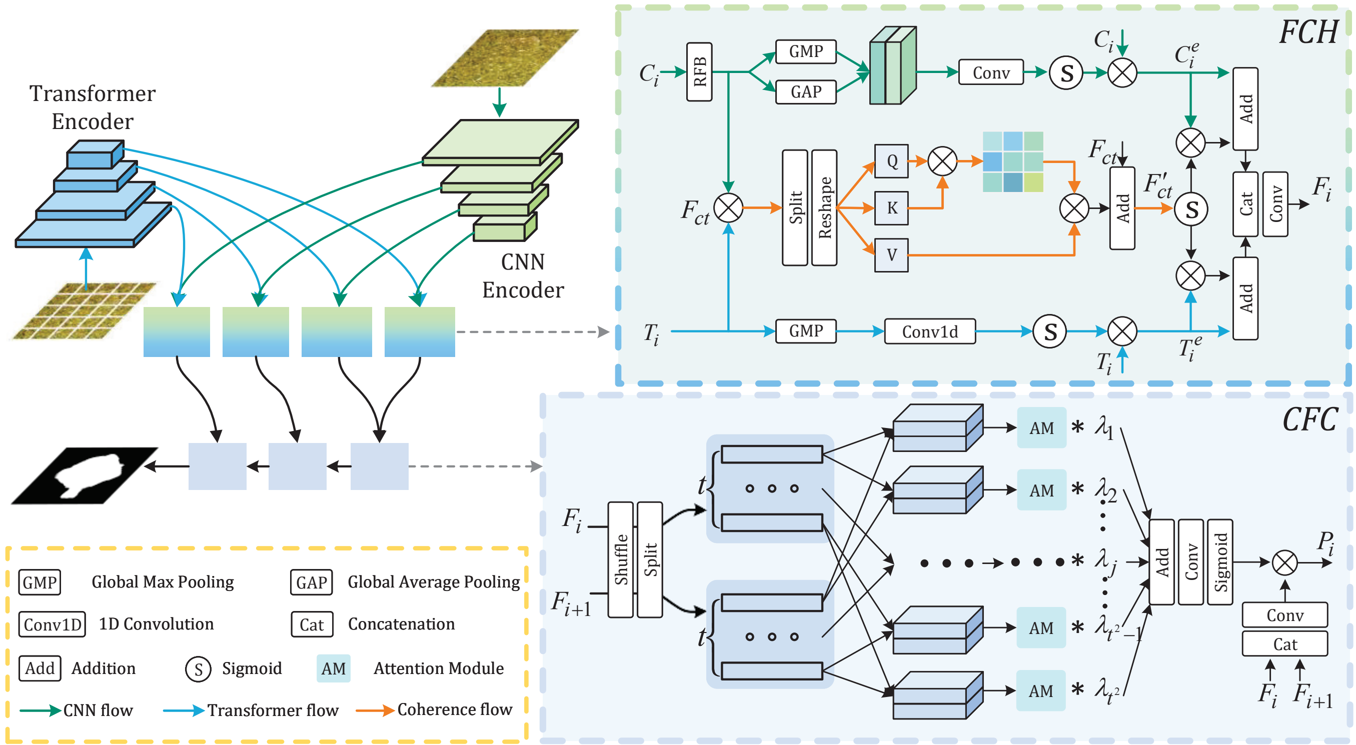

Fig. 2: The framework of our proposed FHCNet, which consists of a Feature Coherence Harmonization (FCH) module and a Cross-layer Feature Calibration (CFC) Module.

Feature Calibration (CFC) module to ensure high consistency and coordination among inter-layer features through adaptive calibration. Unlike existing works [1], [3] that combine CNNs and Transformers, our work specifically addresses the information disparity between CNN features and Transformer features by emphasizing the information harmonization and calibration to explore complementarity, thereby maximizing their strengths to localize the camouflaged objects and capture the subtle boundaries.

Our contributions can be summarized as follows:

- We propose a hybrid model that combines CNNs and Transformers, focusing on harmonization and calibration to bridge their semantic gaps, ultimately improving detection performance for COD.
- We design a Feature Coherence Harmonization module to align intra-layer features from CNN and Transformer, ensuring simultaneous exploitation and integration of both local details and global contexts.
- The Cross-layer Feature Calibration Module is proposed to integrate features from deep to shallow layer through a learnable calibration strategy, balancing deep semantic information with shallow spatial details.

II. Methodology

A. Overview

The overall architecture of the proposed network is illustrated in Fig. 2. Firstly, a RGB image $I \in \mathbb{R}^{H\times W\times 3}$ is fed into the Transformer and CNN encoders separately, extracting multi-scale features, denoted as T_i and C_i $(i = 1, 2, ..., 4)$, respectively. Then, the FCH module is responsible for coordinating and integrating the consistency between these two types of features. The harmonized coarse features are then adaptively calibrated by the CFC module from deep to shallow layers to further optimize the feature representation, ultimately generating the final prediction results.

B. Feature Coherence Harmonization (FCH) Module

Transformers excel at modeling long-range dependencies, while CNNs are particularly adept at capturing fine-grained local features. To address the potential knowledge discrepancy between CNN and Transformer features, we design an interactive guidance fusion strategy to leverage their respective advantages while mitigating feature gaps.

Specifically, we design three branches (*i.e.*, CNN branch, Transformer branch and coherence branch), as shown in Fig. 2. The CNN branch is first enhances spatial local details using the RFB [14] module to enlarge the receptive field. Subsequently, a spatial attention module is introduced to further make the network aware of where to learn:

$$C_i^e = \sigma(\mathcal{C}_7([Avg(R(C_i)), Max(R(C_i))])) \otimes C_i \qquad (1)$$

where $\otimes$ denotes the element-wise multiplication, $[\cdot]$ is the concatenation, $\sigma(\cdot)$ refers the sigmoid function, $\mathcal{C}_k(\cdot)$ is a $k\times k$ convolutional layer, $R(\cdot)$ is the RFB module, $Avg(\cdot)$ and $Max(\cdot)$ refer to the global average pooling and global max pooling, respectively.

In Transformer branch, we deploy a local enhancement to guide Transformer features towards regions with significant changes. Specifically, we utilize global max pooling to aggregate the Transformer features T_i. This allows the most prominent activations to be adaptively emphasized, thereby enhancing the detection of subtle yet critical variations across the feature map. The corresponding enhancement weights for each channel are then computed using 1D convolution:

$$T_i^e = \sigma(\mathcal{C}_{1D}(Max(T_i))) \otimes T_i \tag{2}$$

where $\mathcal{C}_{1D}(\cdot)$ is a 1D convolution with kernel size 1×1.

Considering the inherent semantic discrepancy between CNN and Transformer features, the coherence branch aims to harmonize semantically diverse features by establishing global-local consistency. Specifically, the coherence branch synthesizes the strengths of these two complementary features. First, feature fusion is performed via element-wise multiplication, and fused feature is then passed through a multi-head self-attention unit to refine global and local attributes into a more coherent and contextually-aware representation. This process can be formulated as follows:

$$Q, K, V = Reshape(Split((R(C_i) \otimes T_i))) \tag{3}$$

$$F'_{ct} = MHSA(Q, K, V) + R(C_i) \otimes T_i \tag{4}$$

where $MHSA(\cdot)$ denotes multi-head self-attention.

Finally, the enhanced features C_i^e and T_i^e are guided by the feature F'_{ct} to focus on synergistic aspects, producing the output F_i and effectively mitigating knowledge discrepancies.

C. Cross-layer Feature Calibration (CFC) Module

After reasoning about the relationship between global representations and local details through the Coherence Harmonization module, we obtain coarse complementary features at each stage. However, this complementary information may be mixed with noise, affecting the accuracy of prediction results, especially in challenging scenarios where objects blend into the background. To this end, we propose an adaptive information integration scheme based on a learnable calibration strategy to ensure that features from various levels of the network contribute positively to the final prediction. As shown in Fig. 2, we first shuffle and split and regroup cross-layer features along the channel dimension, achieving $\{F_i^x, F_{i+1}^y\}(x, y = 1, 2, ..., t)$, respectively. Next, learnable parameters $\lambda_j (j = 1, 2, ..., t^2)$ are introduced to adaptively calibrate the grouped attention maps, balancing their contributions and producing a more harmonized integrated feature P_i. And the Attention module (AM) is composed of channel attention and spatial attention. It be formulated as follows:

$$M_j = \mathcal{C}_3([F_i^x, F_{i+1}^y]) \tag{5}$$

$$A_j = SA(CA(M_j) \otimes M_j) \otimes M_j \tag{6}$$

$$P_i = \sigma(\mathcal{C}_3(\sum_{j=1}^{t^2}(A_j \cdot \lambda_j))) \otimes \mathcal{C}_3([F_i, F_{i+1}]) \tag{7}$$

where $CA(\cdot)$ is channel attention and $SA(\cdot)$ is spatial attention. In our experiments, $t = 8$.

D. Loss Function

We adopt a hybrid loss function $\mathcal{L}_{hybrid}$, which consists of a weighted BCE loss ℓ_{bce}^{ω} [22] and a weighted IoU loss ℓ_{iou}^{ω} [22] to evaluate the differences between the prediction map generated by P_i and the ground truth. Therefore, the total loss of our network can be formulated as $\mathcal{L}_{hybrid} = \ell_{iou}^{\omega} + \ell_{bce}^{\omega}$ and $\mathcal{L}_{total} = \sum_{i=1}^{3} \mathcal{L}_{hybrid}(P_i, G)$, where G is the ground truth.

III. EXPERIMENTS AND RESULTS

A. Experiments

1) Datasets: To validate the effectiveness of our proposed model, we evaluate our FHCNet on three popular COD benchmark datasets, including CAMO [13], COD10K [9] and NC4K [15]. CAMO contains 1,250 camouflaged images, with 1,000 images used for training and 250 images for testing. COD10K includes a total of 5066 images, where 3,040 images are chosen for training and 2,026 images are used for evaluation. NC4K is a challenging COD dataset consisting of 4121 images for testing.

2) Evaluation Metrics: We employ four widely used metrics to evaluate the performance, including mean absolute error ($\mathcal{M}$) [19], E-measure (E_{ϕ}^{m}) [7], S-measure (S_m) [6] and weighted F-measure (F_{β}^{ω}) [16].

3) Implementation Details: Our proposed network is implemented by PyTorch with an NVIDIA GTX 3090 GPU (24GB memory). Adam is used as the optimizer, and the learning rate is initialized to 5e-5, adjusted by poly strategy with the power of 0.9. During training, all the input images are resized to 384×384 and augmented by random flipping, cropping, rotation, and color enhancement before being fed into our network. The batch size is set to 12 and the total number of training epochs is set to 50. The code will be available at https://github.com/LinAyq/FHCNet.

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

We compare our FHCNet with several representative methods, including PFNet [17], MGL_R [24], SINetV2 [8], ZoomNet [18], FPNet [4], FSPNet [12], FEDER [11], DINet [25], RISNet [21] and CamoFormer [23]. To ensure a fair comparison, all the predicted maps are provided by the authors or reproduced by the public released codes. Table. I presents the quantitative results of our proposed method against other competitors. Specifically, our FHCNet significantly outperforms competitors on the CAMO and NC4K datasets, and exhibits competitive performance on the COD10K dataset, further highlighting the superiority of the proposed network. In addition, Fig. 3 shows the visual comparisons of our FHCNet and some other state-of-the-art methods. It can be seen that our predictions are closer to the ground truth.

C. Ablation Studies

1) Analysis of the Key Components: To validate the effectiveness of the proposed key components, a series of ablation experiments are performed, as shown in Table II. ”Base” represents the baseline model, where the proposed modules are replaced by convolution and concatenation operations.

TABLE I: Quantitative comparison with the SOTA methods on three benchmark datasets. Notes ↑ / ↓ denote the larger/smaller is better, respectively. "–" is not available. The top two models are **bolded** and underlined for highlighting, respectively.

Method	Publication	Backbone	CAMO				COD10K				NC4K			
			$S_m \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^m \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^m \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^m \uparrow$
PFNet	21CVPR	ResNet	0.782	0.695	0.085	0.841	0.800	0.660	0.040	0.877	0.829	0.745	0.053	0.887
MGL-R	21CVPR	ResNet	0.776	0.673	0.088	0.812	0.814	0.666	0.035	0.851	0.833	0.739	0.053	0.867
SINetV2	22TPAMI	Res2Net	0.820	0.743	0.071	0.882	0.815	0.680	0.037	0.887	0.847	0.770	0.048	0.903
ZoomNet	22CVPR	ResNet	0.820	0.752	0.066	0.877	0.838	0.729	0.029	0.888	0.853	0.784	0.043	0.896
FPNet	23MM	Pvt	0.851	0.802	0.056	0.905	0.850	0.755	0.028	0.912	-	-	-	-
FSPNet	23CVPR	Vit	0.856	0.799	0.050	0.899	0.851	0.735	0.026	0.895	0.879	0.816	0.035	0.915
FEDER	23CVPR	ResNet	0.802	0.738	0.071	0.867	0.822	0.716	0.032	0.900	0.847	0.789	0.044	0.907
DINet	24TMM	Res2Net	0.821	0.748	0.068	0.873	0.832	0.724	0.031	0.903	0.856	0.790	0.043	0.909
RISNet	24CVPR	Pvt	0.870	0.827	0.050	0.922	**0.873**	**0.799**	0.025	**0.931**	0.882	0.834	0.037	0.925
CamoFormer	24TPAMI	Swin	0.876	0.832	0.043	0.930	0.862	0.772	0.024	**0.931**	0.888	0.840	**0.031**	**0.937**
FHCNet	Ours	Swin-Res2Net	**0.885**	**0.842**	**0.039**	**0.934**	0.869	0.778	**0.023**	**0.931**	**0.893**	**0.843**	**0.031**	**0.937**
FHCNet	Ours	Swin-ResNet	0.884	**0.842**	0.041	0.932	0.864	0.770	0.025	0.928	0.888	0.838	0.033	0.934
FHCNet	Ours	Pvt-Res2Net	0.872	0.825	0.048	0.921	0.869	0.782	**0.023**	**0.931**	0.887	0.836	0.033	0.929
FHCNet	Ours	Pvt-ResNet	0.866	0.824	0.050	0.922	0.869	0.785	**0.023**	0.930	0.887	0.840	0.033	0.933

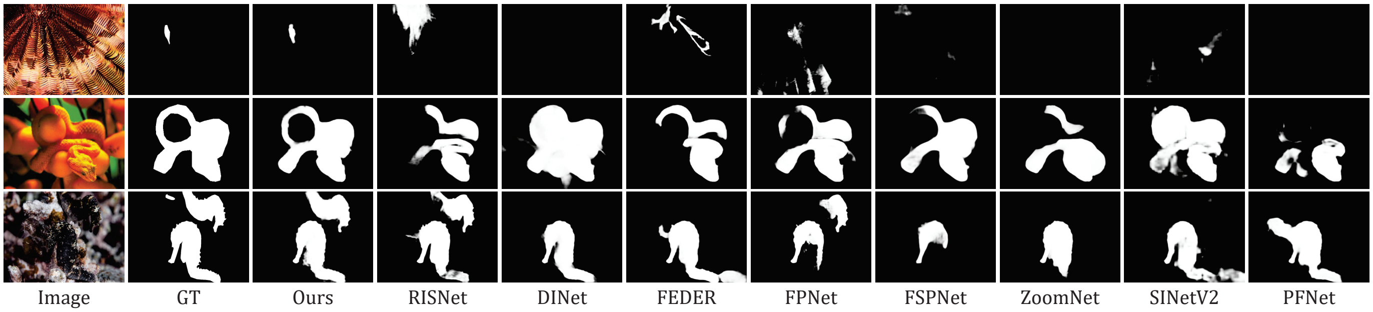

Fig. 3: Visual comparisons of several representative COD methods and our proposed network.

TABLE II: Ablation analyses of our proposed modules.

Method	COD10K			
	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^m \uparrow$
Base	0.839	0.702	0.032	0.900
Base+FCH	0.856	0.750	0.027	0.918
Base+FCH+CFC (Ours)	**0.869**	**0.778**	**0.023**	**0.931**
CNN-CNN	0.829	0.709	0.033	0.896
Transformer-Transformer	0.859	0.757	0.026	0.924
w/o CNN branch	0.852	0.738	0.028	0.913
w/o Transformer branch	0.853	0.743	0.027	0.914
w/o Coherence branch	0.845	0.727	0.029	0.910
t = 4	0.858	0.758	0.026	0.921
t = 8	**0.869**	**0.778**	**0.023**	**0.931**
t = 16	0.862	0.768	0.024	0.924

Different components are gradually added into the baseline. As seen in Table II, the performance improves progressively, indicating that each proposed component plays a positive role in the network, effectively leveraging the strengths of both CNN and Transformer. Additionally, we replace our CNN-Transformer backbone encoders with CNN-CNN and Transformer-Transformer configurations. The results validate the effectiveness of our CNN-Transformer strategy in feature extraction, as it successfully captures complementary information.

2) Analysis of the FCH: As shown in Table II, the individual removal of the CNN branch, Transformer branch, and Coherence branch results in varying degrees of performance degradation, further confirming the effectiveness of the module configuration.

3) Analysis of the CFC: In the CFC, we explore the impact of the number of splitted features t. Experimental results in Table II indicates that the module achieves an optimal performance when $t = 8$.

IV. Conclusion

In this paper, we propose a novel camouflaged object detection network, named FHCNet, which combines the strengths of different but complementary CNN and Transformer features to produce high-quality predictions. Specifically, the FCH module establishes global-local coherence to mitigate the knowledge gap, while the CFC module integrates features through a learnable calibration strategy. Comprehensive experimental results demonstrate that our proposed network surpasses state-of-the-art methods across several evaluation metrics on three widely-used benchmark datasets.

References

[1] Bao, L., Zhou, X., Zheng, B., Yin, H., Zhu, Z., Zhang, J., Yan, C.: Aggregating transformers and cnns for salient object detection in optical remote sensing images. Neurocomputing. **553**(C) (2023)

[2] Bhajantri, N.U., Nagabhushan, P.: Camouflage defect identification: a novel approach. In: International Conference on Information Technology. pp. 145–148. IEEE (2006)

[3] Cong, R., Liu, H., Zhang, C., Zhang, W., Zheng, F., Song, R., Kwong, S.: Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection. In: ACM International Conference on Multimedia. p. 406–416 (2023)

[4] Cong, R., Sun, M., Zhang, S., Zhou, X., Zhang, W., Zhao, Y.: Frequency perception network for camouflaged object detection. In: ACM International Conference on Multimedia. pp. 1179–1189 (2023)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

[6] Fan, D., Cheng, M., Liu, Y., Li, T., Botji, A.: Structure-measure:, a new way to evaluate foreground maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4548–4557 (2017)

[7] Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: International Joint Conference on Artificial Intelligence. pp. 698–704 (2018)

[8] Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6024–6042 (2022)

[9] Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2774–2784 (2020)

[10] Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: Parallel reverse attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 263–273 (2020)

[11] He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22046–22055 (2023)

[12] Huang, Z., Dai, H., Xiang, T.Z., Wang, S., Chen, H.X., Qin, J., Xiong, H.: Feature shrinkage pyramid for camouflaged object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5557–5566 (2023)

[13] Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Computer Vision and Image Understanding **184**, 45–56 (2019)

[14] Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In: European Conference on Computer Vision. pp. 404–419 (2018)

[15] Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11586–11596 (2021)

[16] Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2014)

[17] Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8768–8777 (2021)

[18] Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2160–2170 (2022)

[19] Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 733–740 (2012)

[20] Rustia, D.J.A., Lin, C.E., Chung, J.Y., Zhuang, Y.J., Hsu, J.C., Lin, T.T.: Application of an image and environmental sensor network for automated greenhouse insect pest monitoring. Journal of Asia-Pacific Entomology **23**(1), 17–28 (2020)

[21] Wang, L., Yang, J., Zhang, Y., Wang, F., Zheng, F.: Depth-aware concealed crop detection in dense agricultural scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17201–17211 (2024)

[22] Wei, J., Wang, S., Huang, Q.: F^{3}net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 12321–12328. No. 07 (2020)

[23] Yin, B., Zhang, X., Fan, D.P., Jiao, S., Cheng, M.M., Gool, L.V., Hou, Q.: Camoformer: Masked separable attention for camouflaged object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–14 (2024)

[24] Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., Fan, D.P.: Mutual graph learning for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12992–13002 (2021)

[25] Zhou, X., Wu, Z., Cong, R.: Decoupling and integration network for camouflaged object detection. IEEE Transactions on Multimedia (2024)