Camouflaged Object Detection through Feature Selection and Enhancement

Yin Yuan[1], Kanjian Zhang[21], Jinxia Zhang[21],

1. The Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, P. R. China

2. Southeast University Shenzhen Research Institute, Shenzhen 518057, China

Abstract: Camouflaged object detection is a widely explored problem, the goal of which is to completely segment objects hidden in the background. Compared with general object segmentation and salient object detection tasks, this is very challenging. Existing methods are trying their best to mine the characteristics of the camouflaged objects or obtain some auxiliary information to find the discriminative features that distinguish the targets from the background. However, these methods of explicitly supplementing information have limited exploration and analysis of camouflaged features, and modules designed for specific information are needed. The effectiveness of these modules largely affects the final results. So in this paper, an FSENet (Feature Selection and Enhancement Network) is designed based on an implicit feature selection strategy and a feature enhancement strategy. This method comprehensively utilizes the advantages of CNN and Transformer to obtain the local features and global features of the target and capture as much detail as possible. Moreover, an Implicit Feature Selection (IFS) module based on feature channels is proposed to adaptively select more important features to retain and discard unimportant feature channels to reduce interference. In addition, stacked Multi-scale Region Attention (MRA) modules are designed to enrich the multi-scale information of features and maintain the consistency of multi-scale information. Our method is evaluated on 3 challenging benchmark datasets and outperforms 20 previous state-of-the-art methods under four widely used evaluation metrics.

Key Words: Camouflaged Object Detection, Image Segmentation, CNN, Transformer

1 Introduction

Animals continuously evolve to change their shape, color, and texture to better blend into the environment to avoid predators, and humans use artificial textures or patterns to blend into their surroundings. These are camouflaged targets that exist in both the natural and artificial world. Some camouflaged objects are shown in Fig.1. Camouflaged object detection is a task to distinguish objects that have a high degree of similarity with the background environment. Due to the widespread presence of camouflaged objects, this topic has a profound impact on many fields. For example, pest detection in agriculture[1], enemy reconnaissance in the military[2], and polyp segmentation in medicine[3] can all be achieved through camouflaged object detection methods. Because of its challenges and broad application prospects, more and more researchers are beginning to explore this promising field of camouflaged object detection.

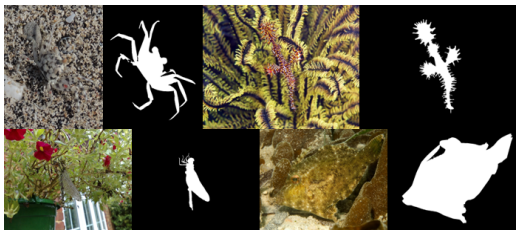

Fig. 1: Some of the camouflaged objects in different scenes.

Camouflaged object detection is a significant challenge for both human vision and computer vision perception systems. Early methods focused on using manually designed low-level features such as texture features and boundary features to separate the foreground and background, but these methods were very limited in effectiveness. Subsequently, the proposal of large-scale camouflaged object datasets led to the rapid development of deep learning methods for camouflaged object detection. Initially, methods like SINet[4] and PFNet[5], mimicked the two steps of location and recognition in the animal hunting process to design a two-stage network to perform a rough prediction and more detailed segmentation of the target. Some methods mimic the human observation process to fully mine the boundary information and gradient information of the target, add multi-view assistance and multi-scale input, and combine other tasks as assistance. These methods focus on explicitly mining and supplementing additional auxiliary information, intending to supplement the loss in the feature extraction process with this auxiliary information and enhance these features.

However, these methods require unique feature extraction stages, and additional supervision information is needed for these auxiliary features. Moreover, the extraction of auxiliary features also has a significant impact on the final prediction of the target. At the same time, these methods did not sufficiently explore, analyze, and utilize multiple auxiliary features. Therefore, we propose an implicit feature selection strategy based on the Implicit Feature Selection (IFS) modules to implicitly and adaptively select features that are more beneficial for target prediction and discrimination to fully utilize the helpful features and reduce judgment interference. Also, a feature enhancement strategy based on the Multi-scale Region Attention (MRA) modules is developed to pay attention to the multi-scale features of the target area.

Specifically, we first obtain the local features and global features of the target at different scales through CNN blocks and Transformer blocks respectively. For these two features, a feature selection strategy is designed. The two features of the same scale pass through our proposed implicit feature selection (IFS) module and select the most important parts of local features and global features based on the channel characteristics. For the discriminative features, the weight

and proportion of this part of the features are increased, and this part of the features is selected and reserved so that the features that contribute more to prediction are strengthened and retained. While the features with a low contribution or with more background interference are eliminated to reduce the impact and interference on forecasts. In addition, we designed a feature enhancement strategy to enhance the target region of the features. The retained features pass through multi-scale region attention (MRA) modules. The multi-scale information of the retained features is obtained with different kernel sizes of convolution layers and fused with an attention map to maintain the consistency of the multi-scale features. Through this process, the features of target regions become more prominent and obvious.

Our proposed network has achieved excellent results on three benchmark data sets: CAMO, COD10K, and NC4K. To sum up, our contributions are as follows:

1. An FSENet is proposed to fully extract the local features and global features of the camouflaged object at different scales while implicitly selecting and enhancing the features.
2. An implicit feature selection module is designed to reserve the features that contribute the most to the result prediction among the local features and global features extracted based on the channel and eliminate the other part to reduce interference.
3. A multi-scale region attention module is designed to fully extract multi-scale features, make multi-scale features consistent, and apply spatial attention to enhance the characteristics of the target region.
4. Numerous experiments have demonstrated the effectiveness of our feature selection strategy and feature enhancement strategy. Our results outperform the previous 20 state-of-the-art methods on three benchmark datasets.

2 Related Works

Salient Object Detection (SOD). SOD is also a segmentation task fundamentally. Over the past few decades, researchers leveraged low-level features[6, 7]. With the development of deep learning, the field of SOD has witnessed the emergence of numerous high-performing approaches. Similar to COD, SOD also requires a focus on discriminative cues of the targets to achieve complete segmentation. Zhao et al.[8] propose an edge-guided salient object detection network based on the complementarity between salient edges and salient objects. Wei et al.[9] generate saliency maps with clear boundaries by explicitly supervising edge pixels. However, due to the high contrast between salient objects and backgrounds, and the relatively simple background in SOD, these methods perform poorly in camouflaged object detection. Specialized designs are needed to address the complex backgrounds and challenging edges associated with camouflage objects.

Camouflaged Object Detection (COD). Extensive research has been conducted on COD in both the academic and industrial sectors. In the early stages, traditional visual techniques were employed, including handcrafted patterns such as motion clues[10] as well as heuristic priors[11] like color, texture, and intensity. With the development of deep learning techniques and the introduction of benchmark datasets like COD10K[4] and NC4K[12], the focus of research has shifted from traditional computer vision and pattern recognition methods to deep learning-based approaches.

Inspired by animal hunting, some methods[4, 5, 13, 14] first locate the camouflaged objects with a coarse prediction map and then generate a detailed map to recognize them. SegMaR[15], also bio-inspired, iteratively refines rough predictions by sampling and zooming in on the predicted locations.

Many methods extract auxiliary information to assist the segmentation of camouflaged objects. Yan et al.[16] and Pang et al.[17] mimic human observation from different viewpoints, introducing auxiliary inputs of varying sizes and views. Sun et al.[18] explored extra object-related boundary semantics to guide representation learning of COD. Similarly, methods[19, 20] extracted boundaries and progressively refined boundary features to get comprehensive priors. Yang et al.[21] enhanced the model's reasoning capabilities, mined high-level semantics with Transformer and extracted clues from the frequency domain.

However, these methods still have problems such as incomplete segmentationand excessive interference. So we develop an FSENet to settle these problems.

3 Proposed Method

In this section, we first present the overall architecture of the proposed FSENet and then illustrate the details of each module and the total loss of the network.

3.1 Overall Architecture

The overall architecture of the proposed FSENet is illustrated in Fig. 2. Given an input image $I_{in} \in \mathbb{R}^{3\times H\times W}$, Res2Net50[22] and PVT v2 b1[23] are employed as the feature extraction networks to obtain multi-scale local features and global features of I_{in}. Then, the local feature and the global feature are concatenated and convoluted. The channel numbers of these features are reduced to 128. These 128-dimension features of the scale i are denoted as $f_i \in \mathbb{R}^{128\times \frac{H}{2^i}\times \frac{W}{2^i}} (i = 1, 2, 3, 4, 5)$. Subsequently, the three semantically rich features f_5, f_4, f_3 are integrated and selected based on feature channels through the IFS modules, while the detailed features f_2, f_1 preserve more texture features and detail features through convolution without special treatment. The features retained by the IFS module and the detail features processed by convolution are represented as $f_i^s (i = 1, 2, 3, 4, 5)$. Then, multi-scale information of these re-integrated features is obtained through the MRA modules to enrich the details and diversity of the features. The MRA modules also maintain the consistency of multi-scale features, apply extra attention to the key target regions in the features, and enhance the features of the target regions. The multi-scale enhanced features $f_i^e (i = 1, 2, 3, 4, 5)$ are connected from top to bottom through the Feature Pyramid Network (FPN) structure to fuse the features from different stages and obtain the output feature f_{pred}. Then f_{pred} is convoluted to a single channel as the final prediction P_{pred}, which is constrained by ground truth $\mathcal{GT}$ with loss $\mathcal{L}_{seg}$.

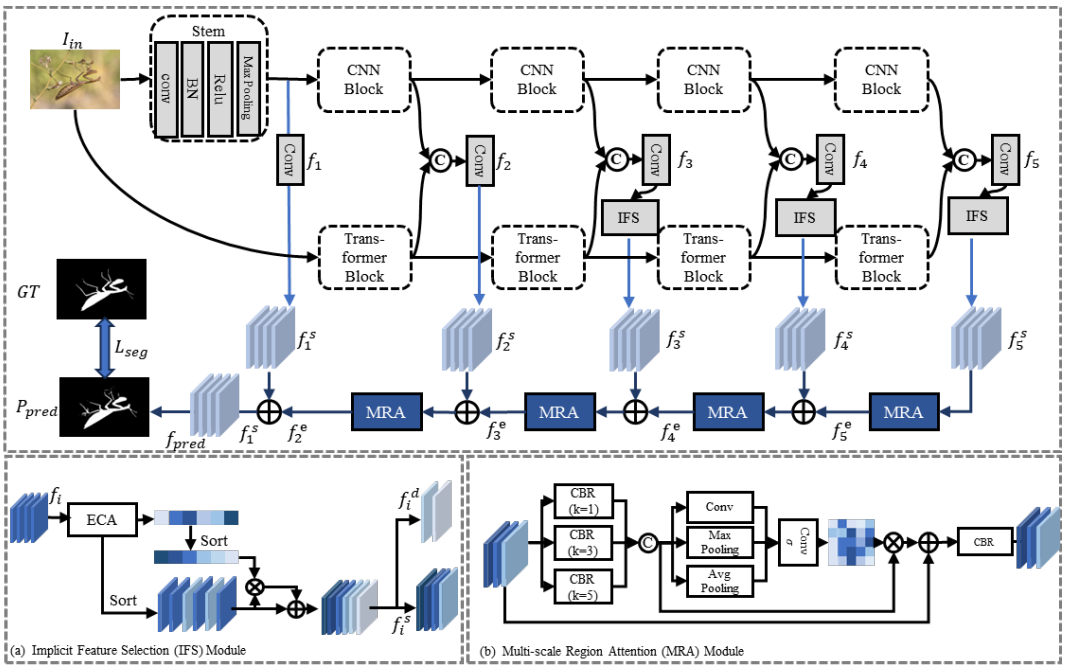

Fig. 2: The overall architecture of the proposed Network: the feature extraction part with a CNN backbone and a Transformer backbone, (a) the proposed Implicit Feature Selection (IFS) module, and (b) the proposed Multi-scale Region Attention (MRA) module.

3.2 Implicit Feature Selection Module

To achieve the adaptive interaction and complementarity of local and global features, a feature selection strategy based on the Implicit Feature Selection (IFS) module is designed to filter and integrate the local features and global features. The structure of the proposed Implicit Feature Selection (IFS) module is shown in Fig. 2.

Specifically, the local and global features are concatenated and fused by convolution layers to inject local and global information into $f_i(i = 1, 2, 3, 4, 5)$. Then, the weights w corresponding to the feature channels are calculated through ECA, and then the weights and the corresponding channels are sorted. The top half with the highest weights is selected for retention, which is denoted as $f_i^s(i = 1, 2, 3, 4, 5)$ in the figure. Since the surrounding environment of the camouflaged objects is complex, channels with smaller weights contain more background interference, so the half with smaller weights is eliminated and does not participate in subsequent processing. Here, the number of channels retained is set to 64 to minimize the computational load. Through this process, local and global features are fully integrated. Features that contribute more to the prediction results of camouflaged objects are screened out and the interference is eliminated. The features f_i^s possess both local and global information while retaining rich target semantic information. It is worth noting that the ECA used to calculate channel weights is an efficient channel attention module, as shown in Fig. 3. It efficiently realizes local cross-channel interaction through a global average pooling layer and a one-dimensional convolution, extracting the dependencies between channels. It can also be flexibly replaced by other modules to obtain channel weights.

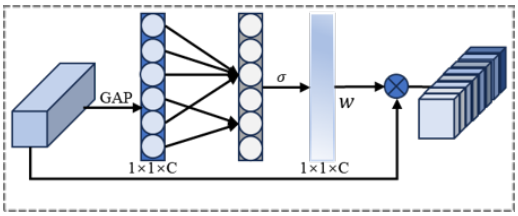

Fig. 3: Illustration of ECA and the calculating process of w.

3.3 Multi-scale Region Attention Module

Multi-scale information plays a significant role in enhancing target features and achieving complete segmentation, so a multi-scale region attention (MRA) module is designed to implement the feature enhancement strategy. As shown in Fig. 2, three CBRs (Convolution layer together with BN and Relu) with different kernel sizes are used to extract the multi-scale features of the target. These features from the three scales are concatenated and convoluted to 64 channels. Then, different attentions to the features are obtained through a convolution layer, a max pooling layer, and an average pooling. These three attentions are concatenated and convoluted to one channel to get the final attention map, which is used as weights for weighting multi-scale features. Finally, the original features are superimposed through residual connection to achieve multi-scale enhancement of the original features and features of the key attention regions. As shown in Fig. 2, by stacking MRA modules from top to bottom, the features of different stages are superimposed and fused to get the final prediction feature f_{pred}.

3.4 Loss Function

The prediction feature f_{pred} is convoluted to one channel as the final prediction map P_{pred}. $\mathcal{L}_{seg}$ is utilized to constrain P_{pred} to get better prediction results. $\mathcal{L}_{seg}$ is divided into two parts, one of which is the commonly used loss in the

camouflaged object detection task–weighted binary cross-entropy (BCE) loss and weighted Intersection over Union (IoU) loss. They are mainly used to constrain the correctness of the prediction results in terms of structure and shape. During the prediction process of camouflaged objects, some pixels are difficult to accurately judge whether they belong to the background or the target, resulting in ambiguous prediction values. To improve the confidence of the prediction results, the UAL proposed by Pang et al.[17] is used in the second part. So overall, our loss $\mathcal{L}_{seg}$ can be expressed as:

$$\begin{aligned}\mathcal{L}_{seg-s} =& \mathcal{L}^{w}_{BCE}(\mathcal{P}_{pred}, \mathcal{GT}) + \mathcal{L}^{w}_{IoU}(\mathcal{P}_{pred}, \mathcal{GT}) \\ &+ \mathcal{UAL}(\mathcal{P}_{pred}, \mathcal{GT})\end{aligned} \tag{1}$$

In which, UAL can be expressed as:

$$\mathcal{UAL}(\mathcal{P}^{i,j}_{pred}, \mathcal{GT}) = 1 - |2P_{i,j} - 1|^2 \tag{2}$$

4 Experiments

4.1 Experiment Setup

DataSet. We evaluate our method on three benchmark datasets: CAMO[34], COD10K[4], and NC4K[12]. CAMO contains 1,250 camouflaged images covering different categories, which are divided into 1,000 training images and 250 testing images. COD10K includes 5,066 camouflaged images (3,040 for training and 2,026 for testing) downloaded from multiple photography websites, covering 5 super-classes and 69 sub-classes. NC4K includes 4,121 images downloaded from the Internet. We follow previous work to use the training set of CAMO and COD10K as the training set (4,040 images) and others as testing sets.

Evaluate Metrics. Following [4, 18, 19], four standard metrics are used to comprehensively evaluate the model performance: Structure measure S_α, weighted F-measure (F^β_ω)[35], Mean Absolute Error (MAE), and mean E-measure (E^m_ϕ)[36]. Structure-measure (S_α) is adopted to compute the structural similarity between object-aware and region-aware. E-measure focuses on evaluating the overall and the local accuracy of camouflaged object detection, which is related to human visual perception mechanisms. Weighted F-measure (F^β_ω) is a comprehensively reliable measure on both weighted precision and weighted recall. Mean absolute error (MAE) evaluates the element-wise difference between the normalized prediction and the ground truth.

Implementation details. We implement our proposed model on PyTorch[37]. Res2Net50[22] and PVT v2 b1[23], pre-trained on ImageNet, are utilized as the backbone. SGD optimizer is set with a weight decay of 2e-4. The learning rate is set to 2e-3 and decays follow a cosine curve. During training, the batch size is set to 8 and the maximum epoch is set to 40. The input image is resized to 384 × 384 and then fed into the network to obtain the predicted binary map. All the experiments are running with an Nvidia GeForce RTX 3090 GPU.

4.2 Comparisons with State-of-the-arts

Qualitative Evaluation. As shown in Fig. 4, our segmentation results for objects with different camouflage characteristics are compared with some previous SOTA methods. For targets with similar structure and color to the environment in row 1, our method can perfectly distinguish the target from the environment, and segment the target details very finely. For occluded targets in row 2, our method can segment the occluded area while also ensuring the integrity of the target. For large targets that are perfectly integrated with the background in row 3, our method can obtain more target semantic information than other methods. While ensuring the integrity of the target, the acquisition and segmentation results of the complex boundary of the target are also better. For scenarios with multiple small targets in row 4, our segmentation results are more accurate in positioning the target. In addition, the number of targets contained in the segmentation result is closer to GT than other methods. As shown in the last row, other methods can segment targets with a lower degree of camouflage. Targets with a higher degree of camouflage cannot be recognized completely. Only our method can segment two targets accurately with rich details.

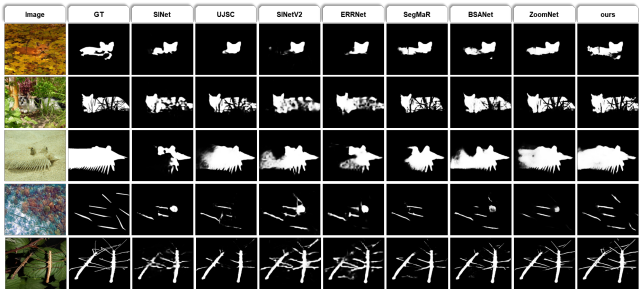

Fig. 4: Visual comparison of the proposed model with eight state-of-the-art COD methods. Our method is capable of accurately segmenting various camouflaged objects in difficult senses.

Quantitative Evaluation. Tab. 1 shows the quantitative results of our PONet compared with other 20 state-of-the-art methods on the three benchmark datasets. Our FSENet outperforms all the other methods on all four standard metrics. Specifically, our method improves S_α by 1.5%, F^β_w by 1.6%, MAE by 1.6% on CAMO, S_α by 0.7%, F^β_w by 0.5% on COD10K, On NC4K, compared with the second-ranked method S_α, it increased by 1.8%, F^β_w increased by 2.2%, MAE decreased by 9.3%, and E^m_ϕ increased by 0.8%. We also plot precision-recall and Fmeasure-Threshold curves of some SOTA methods and our method on 3 benchmarks. As can be seen in Fig. 5, our method achieves better results.

4.3 Ablation Study

We conduct ablation experiments to demonstrate the effectiveness of our components, including IFSs and MRA. The baseline A only uses the backbone network and an FPN structure composed of several convolutions for prediction. We install the components one by one on the baseline model to evaluate their performance. The results are shown in Tab. 2.

To explore a more appropriate training strategy, we adjusted the learning rate and experimented. The experimental results in Tab. 3 show that the best effect is achieved when the initial learning rate of our network is set to 2e-3.

Table 1: Quantitative results of our method and other 20 state-of-the-art methods on three benchmark datasets. "↑(↓)" after the four metrics means that higher (lower) values indicate better performance. The best results are in bold, and the second-ranked results are underlined.

Model	CAMO(250)				COD10K(2026)				NC4K(4121)			
	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$E_\beta^m\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$E_\beta^m\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$MAE\downarrow$	$E_\beta^m\uparrow$
Salient Object Detection Methods												
EGNet [8]	0.732	0.604	0.109	0.800	0.736	0.517	0.061	0.810	0.777	0.639	0.075	0.841
CPD [24]	0.726	0.553	0.114	0.723	0.748	0.509	0.058	0.766	0.788	0.632	0.074	0.804
MINet [25]	0.737	0.613	0.096	0.767	0.769	0.601	0.043	0.820	0.805	0.705	0.060	0.846
ITSD [26]	0.750	0.610	0.102	0.780	0.767	0.557	0.051	0.808	0.811	0.680	0.064	0.845
UCNet [27]	0.739	0.640	0.094	0.787	0.776	0.633	0.042	0.865	0.811	0.729	0.055	0.871
Camouflaged Object Detection Methods												
SINet [4]	0.751	0.606	0.100	0.771	0.771	0.551	0.051	0.806	0.808	0.723	0.058	0.872
PFNet [5]	0.782	0.695	0.085	0.852	0.800	0.660	0.036	0.890	0.829	0.745	0.053	0.898
MGL [28]	0.775	0.673	0.088	0.847	0.814	0.666	0.035	0.865	0.833	0.739	0.053	0.893
LSR [12]	0.787	0.696	0.080	0.838	0.804	0.673	0.037	0.880	0.840	0.766	0.048	0.895
UJSC [29]	0.800	0.728	0.073	0.859	0.809	0.684	0.035	0.884	0.842	0.771	0.047	0.898
UGCT [21]	0.785	0.686	0.086	0.823	0.818	0.667	0.035	0.853	0.839	0.747	0.052	0.874
C2FNet [30]	0.796	0.719	0.080	0.854	0.813	0.686	0.036	0.890	0.838	0.762	0.049	0.897
SINet-v2 [31]	0.820	0.743	0.070	0.882	0.815	0.680	0.037	0.887	0.847	0.770	0.048	0.903
BSANet [20]	0.794	0.717	0.079	0.851	0.818	0.699	0.034	0.891	0.841	0.771	0.048	0.897
OCENet [32]	0.802	0.723	0.080	0.852	0.827	0.707	0.033	0.894	0.853	0.785	0.045	0.902
ERRNet [19]	0.779	0.679	0.085	0.842	0.786	0.630	0.043	0.867	0.827	0.737	0.054	0.887
BGNet [18]	0.812	0.749	0.073	0.870	0.831	0.722	0.033	0.901	0.851	0.788	0.044	0.907
SegMaR [15]	0.815	0.753	0.071	0.874	0.833	0.724	0.034	0.899	0.841	0.781	0.046	0.896
ZoomNet [17]	0.820	0.752	0.063	**0.895**	0.838	0.729	**0.029**	0.888	0.853	0.784	0.043	0.896
CINet [33]	0.827	0.763	0.066	0.888	0.830	0.710	0.033	0.904	0.855	0.789	0.043	0.910
FSENet(ours)	**0.839**	**0.775**	**0.062**	**0.895**	**0.844**	**0.733**	**0.029**	**0.908**	**0.870**	**0.806**	**0.039**	**0.917**

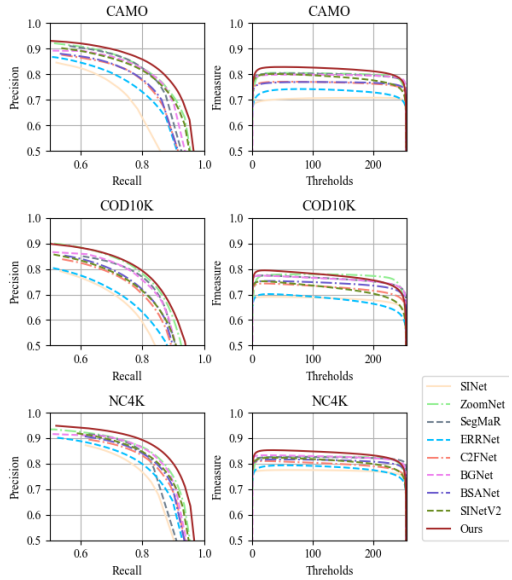

Fig. 5: Precision-Recall and Fmeasure-Threshold curves of some of the SOTA methods and our method on 3 benchmarks. The closer the PR curve is to the upper-right corner, the better the performance is. The higher the F-measure curve is, the better the performance the better the model works.

5 Conclusion

In this paper, we propose a Feature Selection and Enhancement Network (FSENet) to adaptively screen and integrate the local and global features of camouflaged targets, and to enhance the target region at multiple scales. We designed an implicit feature selection (IFS) module to implicitly select the features that contribute more to the target prediction, and remove a part of the interference. We also design a multi-scale region attention (MRA) module to perform multi-scale enhancement of features. It utilizes multi-scale information to obtain precise attention to the target region and weights this region. Experiments show that our method outperforms the previous 20 SOTA methods.

Table 2: Ablation studies of the components.

model	CAMO		COD10K		NC4K	
	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$
A. baseline	0.821	0.752	0.815	0.712	0.854	0.779
B. A+AFSs	0.833	0.768	8.830	0.725	0.862	0.794
C. A+AFSs+MRAs	**0.839**	**0.775**	**0.844**	**0.733**	**0.870**	**0.806**

Table 3: Studies of different settings of learning rate.

learning rate	CAMO		COD10K		NC4K	
	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$
1e-3	0.836	0.771	0.841	0.720	0.868	0.798
2e-3	**0.839**	**0.775**	**0.844**	0.733	**0.870**	**0.806**
3e-3	0.830	0.762	0.842	**0.734**	**0.870**	0.805
4e-3	0.825	0.760	0.842	0.732	0.867	0.803

References

[1] A. Albanese, M. Nardello, and D. Brunelli, "Automated pest detection with dnn on the edge for precision agriculture,"

IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 11, no. 3, pp. 458–467, 2021. 1

[2] I. Forsyth, "Designs on the desert: camouflage, deception and the militarization of space," *Cultural Geographies*, vol. 21, no. 2, pp. 247–265, 2014. 1

[3] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2020, pp. 263–273. 1

[4] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, 2020, pp. 2777–2787. 1, 2, 4, 5

[5] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *CVPR*, 2021, pp. 8772–8781. 1, 2, 5

[6] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, June 2013. 2

[7] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, June 2013. 2

[8] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *CVPR*, 2019, pp. 8779–8788. 2, 5

[9] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *CVPR*, 2020, pp. 13 025–13 034. 2

[10] S. P. McKee, S. N. Watamaniuk, J. M. Harris, H. S. Smallman, and D. G. Taylor, "Is stereopsis effective in breaking camouflage for moving targets?" *Vision Research*, vol. 37, no. 15, pp. 2047–2055, 1997. 2

[11] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann, "Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1382–1402, 2001. 2

[12] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *CVPR*, 2021, pp. 11 591–11 601. 2, 4, 5

[13] H. Mei, K. Xu, Y. Zhou, Y. Wang, H. Piao, X. Wei, and X. Yang, "Camouflaged object segmentation with omni perception," *IJCV*, pp. 1–16, 2023. 2

[14] X. Yan, M. Sun, Y. Han, and Z. Wang, "Camouflaged object segmentation based on matching–recognition–refinement network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[15] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *CVPR*, 2022, pp. 4713–4722. 2, 5

[16] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen, "Mirrornet: Bio-inspired camouflaged object segmentation," *IEEE Access*, vol. 9, pp. 43 290–43 300, 2021. 2

[17] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *CVPR*, 2022, pp. 2160–2170. 2, 4, 5

[18] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," in *IJCAI*, 2022, pp. 1335–1341. 2, 4, 5

[19] G.-P. Ji, L. Zhu, M. Zhuge, and K. Fu, "Fast camouflaged object detection via edge-based reversible re-calibration network," *PR*, vol. 123, p. 108414, 2022. 2, 4, 5

[20] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin, "I can find you! boundary-guided separated attention network for camouflaged object detection," in *AAAI*, vol. 36, no. 3, 2022, pp. 3608–3616. 2, 5

[21] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *ICCV*, 2021, pp. 4146–4155. 2, 5

[22] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019. 2, 4

[23] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, October 2021, pp. 568–578. 2, 4

[24] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *CVPR*, 2019, pp. 3907–3916. 5

[25] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *CVPR*, 2020, pp. 9413–9422. 5

[26] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *CVPR*, 2020, pp. 9141–9150. 5

[27] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *CVPR*, 2020, pp. 8582–8591. 5

[28] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *CVPR*, 2021, pp. 12 997–13 007. 5

[29] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *CVPR*, 2021, pp. 10 071–10 081. 5

[30] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, and T. Zhou, "Camouflaged object detection via context-aware cross-level fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6981–6993, 2022. 5

[31] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6024–6042, 2021. 5

[32] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1445–1454. 5

[33] X. Li, H. Li, H. Zhou, M. Yu, D. Chen, S. Li, and J. Zhang, "Camouflaged object detection with counterfactual intervention," *Neurocomputing*, vol. 553, p. 126530, 2023. 5

[34] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Computer Vision and Image Understanding*, vol. 184, pp. 45–56, 2019. 4

[35] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *CVPR*, 2014, pp. 248–255. 4

[36] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function," *Scientia Sinica Informationis*, vol. 6, no. 6, 2021. 4

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 4