




# TPRNet: camouflaged object detection via transformer-induced progressive refinement network

Qiao Zhang<sup>1</sup> · Yanliang Ge<sup>1</sup> · Cong Zhang<sup>1</sup> · Hongbo Bi<sup>1</sup> 

Accepted: 24 June 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

Camouflaged object detection (COD) is a challenging task which aims to detect objects similar to the surrounding environment. In this paper, we propose a transformer-induced progressive refinement network (*TPRNet*) to solve challenging COD tasks. Specifically, our network includes a Transformer-induced Progressive Refinement Module (TPRM) and a Semantic-Spatial Interaction Enhancement Module (SIEM). In TPRM, high-level features with rich semantic information are integrated through transformers as prior guidance, and then, it is sent to the refinement concurrency unit (RCU), and the accurately positioned feature area is obtained through a progressive refinement strategy. In SIEM, we perform feature interaction to localized-accurate semantic features and low-level features to obtain rich fine-grained clues and increase the symbolic power of boundary features. Extensive experiments on four widely used benchmark datasets (i.e., CAMO, CHAMELEON, COD10K, and NC4K) demonstrate that our *TPRNet* is an effective COD model and outperforms state-of-the-art models. The code is available <https://github.com/zhangqiao970914/TPRNet>.

**Keywords** Deep learning · Camouflaged object detection · Transformer · Progressive refinement

## 1 Introduction

Camouflage is a long-term evolutionary survival skill for creatures in nature to hide from being discovered by changing colors, contrasts, etc. Camouflaged object detection (COD) is the identification of objects that perfectly blend into the surrounding environment and has a wide range of valuable applications, which involve medical imaging [41] (e.g., polyp detection [13] and lung infection segmentation [14,47,50]), surface defect detection [23], agriculture and security (e.g., disaster detection [1]) and surveillance (e.g., detection of pedestrians [56]).

Unlike salient object detection (SOD) [3,45,48,49], COD is more challenging, and a camouflaged object can fool the observer's visual system based on its texture similarity to the environment. Early methods are mostly based on hand-crafted low-level features such as texture [37], motion [18], 3D convexity [32] to distinguish camouflaged objects from the background. Recently, deep learning has achieved excellent performance in object detection, and the accuracy of COD methods based on deep learning has also been effectively improved. Although recently proposed methods [12,24,29,31,39,42,55] have achieved performance improvements to a certain extent, there is still a large room for exploring effective methods for COD. Existing methods perform well in detecting camouflaged objects from relatively simple scenes, but their performance is unsatisfactory in complex backgrounds and situations where the background is very similar to the camouflaged object. As shown in Fig. 1, state-of-the-art COD models cannot fully identify camouflaged objects in scenarios where the camouflaged object is highly similar to the background, resulting in poor visual effects. However, multi-stage feature refinement and fine-grained mining of spatial features are the keys to solving the above problems. Multi-stage feature refinement can explore the most representative features, and fine-grained mining of

---

✉ Cong Zhang  
cong Zhang98@126.com

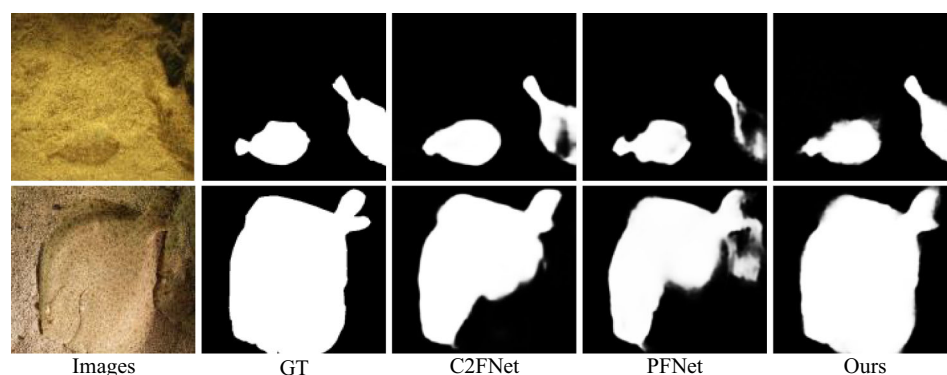
✉ Hongbo Bi  
bhbdq@126.com

Qiao Zhang  
qiaozhang\_dongyou@163.com

Yanliang Ge  
15804593399@139.com

<sup>1</sup> School of electrical information engineering, Northeast Petroleum University, Daqing 163000, China

**Fig. 1** Existing COD methods such as C<sup>2</sup>FNet [42] and PFNet [31] often fail to completely detect camouflaged objects in scenarios where the camouflaged object is highly similar to the background. In contrast, our method can accurately detect camouflaged objects



spatial features can obtain rich boundary information repeatedly.

To this end, we propose a new COD model called Transformer-induced Progressive Refinement Network (*TPRNet*). We first use the transformer in TPRM to integrate global contextual information from high-level features as a guide, which is then fed into the refinement concurrent unit (RCU) for more complete semantic features. The proposed RCU adopts a bilateral refinement strategy to exploit the intrinsic connection between reverse and forward features fully. Second, in the semantic-spatial interaction enhancement module (SIEM), we supplement high-level semantic features with shallow low-level features to obtain richer spatial information, thus making the detection results more complete. In summary, the main contributions of this paper are as follows:

- We propose a novel COD framework (TPRNet) consisting of two important modules, namely the transformer-induced progressive refinement module and the semantic-spatial feature interaction module, which considers high-level and low-level features with rich global contextual information and fine-grained spatial information.
- Our proposed refinement concurrent unit (RCU) employs a bilateral refinement strategy to explore the consistent learning between reverse and forward features.
- Extensive experiments on four challenging datasets, i.e., CAMO, CHAELEMON, COD10K, and NC4K, show that the proposed TPRNet achieves competitive performance.

## 2 Related work

### 2.1 Camouflaged object detection

Earlier methods relied on hand-crafted features and visual features to detect camouflaged objects. Pan et al. [32] proposed a 3D convexity-based camouflaged target detection method, which can continuously detect camouflaged targets in complex backgrounds. Liu et al. [27] proposed a novel top-

down information-based foreground object detection based on expectation maximization, and the foreground model can enhance the foreground detection of camouflaged objects. In [37], Sengottuvelan et al. proposed a recognition system for detecting camouflaged objects from a single input image with a co-occurrence matrix approach. Yin et al. [18] used optical flow to model the detection of moving objects with camouflage colors. Early methods are only applicable when the camouflaged object is relatively distinct from the background. However, if there is a high similarity between the foreground and background, it will significantly reduce the detection performance.

Recently, CNN has achieved excellent performance in different computer vision tasks, and some researchers have developed CNN-based models to solve the problem of camouflaged object detection. Fan et al. [12] proposed a network (SINet) consisting of a search module and an identification module. The search module is mainly used to store feature information at various levels, and the identification module is employed to locate and distinguish camouflaged objects. Wang et al. [42] proposed a new COD network consisting of a dual-branch feature extraction module and a stepwise refinement cross-fusion module, namely D<sup>2</sup>C-Net, which achieved good satisfactory detection results through dual-stage feature refinement. Sun et al. [39] proposed a novel context-aware cross-level fusion network (C<sup>2</sup>F-Net) to solve the challenging COD task, which utilizes rich global contextual information to achieve more accurate detection results. Li et al. [24] proposed an adversarial learning network using high-order similarity measures and network confidence estimation to enhance the detection of salient objects and camouflage. Dong et al. [7] proposed a new deep learning-based COD method that integrates a large receptive field and efficient feature fusion into a unified framework. Fan et al. [11] proposed an enhanced version of the network based on [12], which includes a neighbor connection decoder and group-reversal attention, achieving promising performance. Yang et al. [52] propose a new network that uses a probabilistic representation model combined with transformers to address uncertainty in camouflaged scenarios. Pang et al. [53]

proposed a mixed-scale triplet network, ZoomNet, which imitates human behavior when observing blurred images, i.e., zooming in and out, and achieves excellent performance. Unlike the above methods, we mainly consider multi-level feature refinement and spatial fine-grained feature mining. The proposed RCU starts from the direction of bidirectional refinement, which effectively promotes forward feature and reverse feature consistency learning.

## 2.2 Self-attention mechanism

Self-attention mechanisms have been proposed to capture long-range information and applied to many tasks such as machine translation [40] and image feature extraction [44]. In the self-attention mechanism, the input tokens are linearly transformed into queries, keys, and values, in the embedding layer, and then, the long-range relationship between the tokens of the input sequence is computed through dot product attention. Limited by the locality of the convolution operator, the convolutional neural network (CNN) cannot model global contextual information. Wang et al. [44] proposed a non-local neural network, a classic implementation of the self-attention mechanism. Chen et al. [3] proposed a “double attention block” to reduce the complexity of traditional non-local modules. Fu et al. [15] proposed a dual attention model of spatial attention and channel attention in the field of scene segmentation to explore the potential relationship between different dimensions. In this paper, we exploit the similarity between self-attention affinity matrices to drive feature-consistency learning.

## 2.3 CNN-based model and transformer-based model

Convolutional Neural Network (CNN) can perform feature extraction on the target and then, classify, identify, predict or make decisions based on the features. Recently, Cui et al. [6] proposed a temporal feature blender network for video object detection based on a CNN model, which can plug into any detection network effortlessly to improve detection behavior. Liu et al. [25] proposed an end-to-end deep learning framework, called Motion-Aided Feature Calibration Network (MFCN), for video object detection. Liu et al. [26] exploited CNN to tackle the task of video instance segmentation from a new perspective and proposed a single-stage spatial granularity network (SGNet). Cui et al. [5] proposed an end-to-end annotator named DGLabeler based on CNN, whose architecture includes a new deep granularity module to model the spatial relationship of instances and help generate fine-grained instance masks. Recently, there has been an explosion of transformer-based models inspired by the successful ViT [8]. T2T [54] progressively structure the image to tokens by recursively aggregating neighboring tokens into one token. PVT [43]

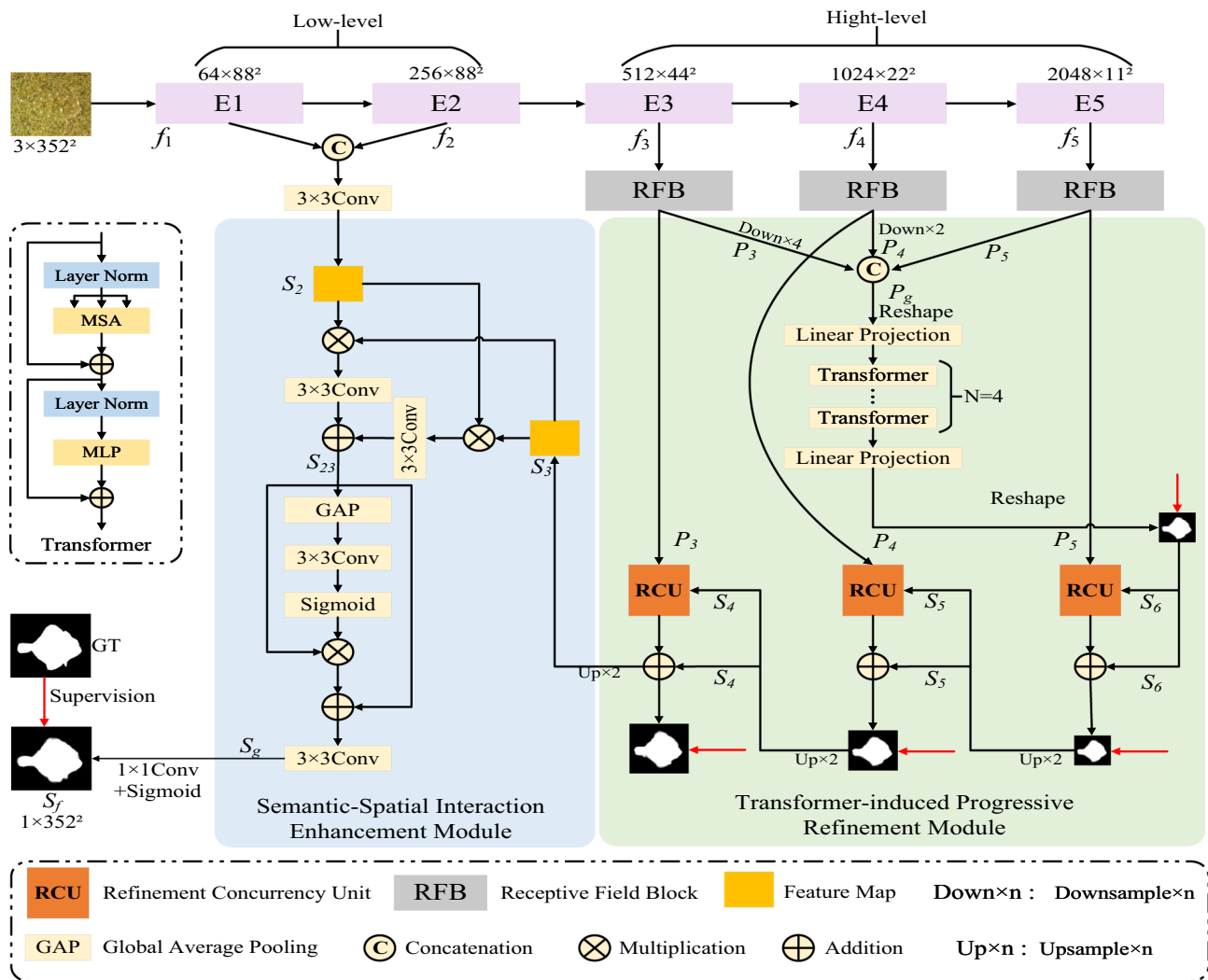
introduces a progressive shrinking pyramid to reduce the sequence length of the transformer. DPT [36] assembles tokens from multiple stages of the vision transformer and progressively combines them into full-resolution predictions using a convolutional decoder. Swin transformer [28] designs the shifted window-based multi-head attention to reduce the computation cost. The CNN-based network retains the original positional relationship after the convolution operation, which has great advantages for fine-grained spatial information extraction. Transformer-based networks can utilize self-attention to model long-term feature dependencies, facilitating global contextual information aggregation. In this paper, we integrate the strengths of CNN and Transformer, using Transformer to aggregate global information while exploiting CNN to mine local details.

## 3 Method

In this section, we will elaborate our proposed method for camouflaged object detection. Concretely, we first introduce the overall network architecture of the proposed method in Sect. 3.1. Then, we present the details of each specific module in Sects. 3.2 and 3.3, including transformer-induced progressive refinement module and semantic-spatial interaction enhancement module.

### 3.1 Architecture overview

Figure 2 shows the overall architecture of the proposed TPRNet. We adopt a pre-trained Res2Net network [16] as the baseline while applying RFB [12,48] at the high-level feature layer to expand the receptive field, which is combined with two proposed modules: transformer-induced progressive refinement module (TPRM) and semantic-spatial interaction enhancement module (SIEM) to predict result maps. Specifically, given a set of images as input, the Res2Net network is used to extract the backbone features. For simplicity, we define it as  $f_i$  ( $i = 1, 2, \dots, 5$ ). Then, we apply RFB [12,48] on the high-level feature  $f_i$  ( $i = 3, 4, 5$ ) to expand the receptive field to obtain rich semantic information and obtain the feature  $P_i$  ( $i = 3, 4, 5$ ). The setting of RFB is the same as [12,48]. In TPRM, we first adjust the size of  $P_3$  and  $P_4$  to be consistent with  $P_5$  and perform channel concatenation to obtain the enhanced feature  $P_g$  and then reshape  $P_g$  into a 1D sequence as the input of the transformer to extract global information. After that, we reshape the obtained global information into 2D feature  $S_6$  as a guide and input it together with feature  $P_5$  into the fifth-layer RCU for feature refinement and obtain feature  $S_5$  after residual connection with  $S_6$ . Similarly, we can also obtain feature  $S_4$  and  $S_3$ . In SIEM, we first perform channel concatenation of low-level features  $f_1$  and  $f_2$  to obtain feature  $S_2$  and then, interact with feature  $S_3$



**Fig. 2** The overall architecture of our proposed *TPRNet*, which consists of two main modules, namely, Transformer-induced Progressive Refinement Module (TPRM) and Semantic-Spatial Interaction Enhancement Module (SIEM)

to obtain rich fine-grained clues to obtain feature  $S_g$ , which is the final prediction result after convolution and sigmoid.

### 3.2 TPRM: transformer-induced progressive refinement module

In previous work [11], multi-level aggregating features as an introductory guide and performing level-wise feature refinement can achieve satisfactory performance. To further explore the above issues, we propose a transformer-induced progressive refinement module (TPRM); the TPRM consists of transformers and three refinement concurrent units (RCU). Self-attention in transformers can learn pairwise similarities between input sequences, which can capture global contextual information with long-term dependencies. In the RCU, global context information is used as a priori guidance for

concurrent refinement with each layer side output to obtain accurate detection results.

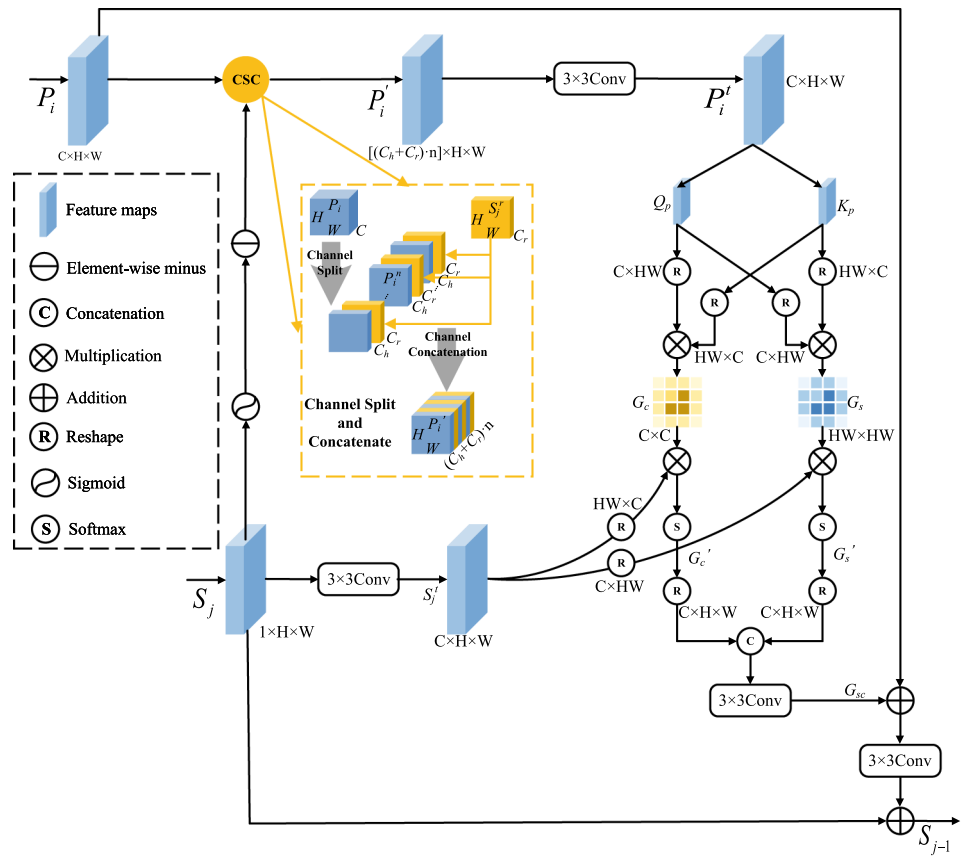
As shown in Fig. 2, we first apply transformers [40] to integrate the high-level features of the three layers to obtain vital global contextual information, and the whole process can be formulated as:

$$P_g = \text{Cat}(D_4(P_3), D_2(P_4), P_5) \quad (1)$$

$$S_6 = R(LP(\text{Transformer}(LP(R(P_g))))) \quad (2)$$

where  $\text{Cat}(\cdot)$  is the concatenation operation among channel dimensions.  $D_i, i = 2, 4$  stands for downsample operation.  $R(\cdot)$  defines a reshaping operator.  $LP(\cdot)$  is a linear projection. The transformer layer contains a multi-headed self-attention (MSA) and multilayer perceptron (MLP) sublayer. Layer normalization (LN) [2] is inserted before these two sublayers, and the residual connection is performed after these sub-

**Fig. 3** The details of our proposed refinement concurrent unit (RCU)



layers. In the transformer, the input patch size is  $14 \times 14$ , and the number of patches is 384. Second, we feed the obtained feature  $S_6$  as a guiding prior and the high-level feature side output  $P_i$ , ( $i = 3, 4, 5$ ) into RCU for refinement and concurrency and employ a residual learning strategy [17] to enhance the feature representation, and the whole process can be formulated as:

$$S_i = RCU(P_i, S_j) \oplus S_j, i = 3, 4, 5; j = 4, 5, 6 \quad (3)$$

where  $RCU(\cdot)$  represents refinement concurrency unit.  $\oplus$  denotes element-wise summation.

### 3.2.1 Refinement concurrency unit

Different from GRA [11], our RCU adopts a bidirectional refinement strategy, namely reverse feature refinement strategy and forward feature concurrency strategy. The reverse feature refinement strategy can explore useful information in reverse features, while the forward feature refinement strategy can promote consistent learning among features.

**Reverse feature refinement strategy** First, the feature  $P_i$  is divided into multiple groups (i.e.,  $n$  groups, we set  $n = 4$ ) along the channel dimension, each of which has  $C_h = C/n$

channels. After that, the reverse feature  $S_j^r$ , as a complementary feature, is periodically interpolated into the split feature groups to obtain a combined feature  $P_i'$  with  $(C_h + C_r) \cdot n$  channels. The above channel-wise split and concatenation operations can be formulated as:

$$\{P_i^1, \dots, P_i^n\} = F_{split}(P_i), i = 3, 4, 5 \quad (4)$$

$$P_i' = F_{cat}(\{S_j^r, P_i^1\}, \dots, \{S_j^r, P_i^n\}), \\ i = 3, 4, 5; j = 4, 5, 6 \quad (5)$$

where  $n \in \{1, 2, 3, 4\}$ , and  $F_{split}$  and  $F_{cat}$  represent the split and concatenation operations.

**Forward feature concurrency strategy** For features  $P_i'$  and  $S_j$ , we first restore the number of channels to 32 channels by a  $3 \times 3$  convolution, thereby obtaining features  $P_i^t \in R^{C \times H \times W}, i = 3, 4, 5$  and  $S_j^t \in R^{C \times H \times W}, j = 4, 5, 6$ . Inspired by [15], we consider both position-wise attention and channel-wise attention. To this end, we design a two-branch structure, i.e., the position branch and the channel branch, to compute them independently. As shown in Fig. 3, the two branches share the same embedding, namely  $Q_p$  and  $K_p$ , which are reshaped to the size of  $H \times W \times C$  and  $C \times H \times W$ , respectively. The position branch generates spatial



similarity matrix  $G_s \in R^{HW \times HW}$ . The channel branch generates channel similarity matrix  $G_c \in R^{C \times C}$  to explore the dependencies along channel dimension. The two branches are calculated as:

$$G_s = Q_p \odot K_p \quad (6)$$

$$G_c = K_p \odot Q_p \quad (7)$$

where  $G_s$  and  $G_c$  denote spatial similarity matrix and channel similarity matrix, respectively.  $\odot$  denotes matrix multiplication. Next, we multiply  $S_j$  by the spatial similarity matrix  $G_s$  and the channel similarity matrix  $G_c$ , respectively, for concurrency in the spatial dimension and the channel dimension. We reshape the feature maps ( $G'_s$  and  $G'_c$ ) into 2D features and then connect them in the channel dimension and obtain enhanced features  $G_{sc}$  via  $3 \times 3$  convolution operation. The  $G_{sc}$  can be expressed by the following formula:

$$G_{sc} = \text{Conv}_{3 \times 3} \left( F_{\text{cat}} \left( R \left( G'_c \right), R \left( G'_s \right) \right) \right) \quad (8)$$

where  $\text{Conv}_{3 \times 3}$  is a  $3 \times 3$  convolution operation. Finally, for better learning to refine feature representations, we introduce a residual strategy, which is denoted as:

$$S_{j-1} = \text{Conv}_{3 \times 3} \left( \text{Conv}_{3 \times 3} (G_{sc} \oplus P_i) \oplus S_j \right) \quad (9)$$

where  $i \in \{3, 4, 5\}$  and  $j \in \{4, 5, 6\}$ .  $S_5$ ,  $S_4$ , and  $S_3$  are the refinement features of the three layers, respectively.

### 3.3 SIEM: semantic-spatial interaction enhancement module

Generally, existing methods [12,31,39,42] focus more on the search and localization of semantic features of camouflaged objects while ignoring spatial information. Inspired by [7, 57], we design a semantic-spatial interaction enhancement module to implicitly interact semantic features with spatial features, resulting in valuable fine-grained clues.

As shown in Fig. 2, we first concatenate the features  $f_1$  and  $f_2$  in the channel dimension to ensure more spatial information and then, reduce the number of channels of the connected feature  $S_2$  to be consistent with  $S_3$  through a  $3 \times 3$  convolution operation, the whole process can be expressed as:

$$S_2 = \text{Conv}_{3 \times 3} (F_{\text{cat}} (f_1, f_2)) \quad (10)$$

Directly connecting features  $S_2$  with  $S_3$  may lead to inconsistencies, for which we employ an interactive connection strategy to ensure the accuracy and validity of spatial information. The interactive connection process can be expressed as:

$$S_{23} = \text{Conv}_{3 \times 3} (S_2 \otimes S_3) \oplus \text{Conv}_{3 \times 3} (S_3 \otimes S_2) \quad (11)$$

where  $\otimes$  denotes element-wise multiplication. Finally, we exploit the attention mechanism [19] to mine the most critical clues of interactive features and improve the representation ability of the features.

$$S_g = \text{Conv}_{3 \times 3} (\sigma (\text{Conv}_{3 \times 3} (\text{Avg} (S_{23}))) \otimes S_{23} \oplus S_{23}) \quad (12)$$

where  $\sigma$  is sigmoid function.  $\text{Avg}(\cdot)$  denotes global average pooling.

### 3.4 Loss function

Our loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}^w (\mathcal{F}^s, \mathcal{G}) + \mathcal{L}_{\text{IoU}}^w (\mathcal{F}^s, \mathcal{G}) \quad (13)$$

where  $\mathcal{L}_{\text{BCE}}^w$  and  $\mathcal{L}_{\text{IoU}}^w$  represent the binary cross entropy (BCE) loss and the weighted intersection-over-union (IoU) loss, and  $\mathcal{G}$  denotes the ground-truth. Unlike the standard IoU loss, which is widely used in image segmentation and object detection, the weighted IoU loss increases the weight of hard pixels to emphasize their importance. Besides, different from the standard BCE loss, the  $\mathcal{L}_{\text{BCE}}^w$  focuses more on hard pixels instead of assigning equal weights to all pixels. More details of these two losses can be found in [11,46].

## 4 Experiments

### 4.1 Experiment setup

In this section, we describe the benchmark datasets in the COD field, the evaluation metrics, the implementation details of experimental, comparisons with other deep learning models, and ablation study.

#### 4.1.1 Datasets

There are four COD datasets used in our experiments. CAMO [22] is a small-scale COD dataset containing 1000 training images and 250 testing images. CHAMELEON [38] dataset contains 76 images; it is currently the smallest COD dataset. The COD10K [12] is a challenging COD dataset with 3040 training images and 2026 testing images. It consists of 10 super-classes and 78 sub-classes gathered from several photography websites. NC4K [29] is the largest COD dataset and a newly released COD dataset, containing 4121 high-quality images with instance-level annotations. Following [12], our train dataset is a combination of COD10K and CAMO. For more details, please refer to [4].

### 4.1.2 Evaluation metric

We employ four widely used metrics for quantitative comparisons between different COD methods, including:

*S-Measure* [9] is to evaluate non-binary foreground maps. In this paper, we employ S-measure to assess the similarity between the feature map and the ground-truth map. The S-measure is defined as:

$$S = \alpha S_o + (1 - \alpha) S_r \quad (14)$$

According to the experience of previous work [9],  $\alpha$  is set to 0.5.

*F-Measure* [30] is the weighted harmonic average of Precision and Recall. F-measure is calculated by the following formula:

$$F_\beta = \frac{(\beta^2 + 1) P R}{\beta^2 P + R} \quad (15)$$

where  $\beta^2$  is generally set as 0.3 to emphasize more on precision.

*Mean absolute error (MAE)* [35] is designed to directly measures the absolute difference between the ground-truth value and the predicted value, which is formulated as MAE is computed as:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W |S(x, y) - G(x, y)| \quad (16)$$

where  $W$  and  $H$  denote the width and height of the feature map, respectively.  $S$  refers to the feature map, and  $G$  denotes the ground-truth.

*E-measure* [10] is mainly used to measure image-level statistics and local pixel matching information. The E-measure is defined as:

$$E = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H \phi_{FM}(x, y) \quad (17)$$

where  $\phi$  is the enhanced alignment matrix.

### 4.1.3 Implementation details

We use the Res2Net-50 [16] as our baseline. Our proposed model is implemented in Pytorch [34] and trained using an NVIDIA GTX-1080Ti GPU. All images are resized to  $352 \times 352$  for training and testing. The network is trained 100 epochs in total with the Adam optimizer [21]. During the training stage, the batch size is set to 14, and the learning rate starts at  $1e-4$ , dividing by 10 every 50 epochs.

## 4.2 Comparison with state of the art

We compare the proposed model with eighteen typical deep learning-based models: CPD [48], SCRNet [49], EGNNet [58], F3Net [46], MINet [33], PrNet [13], ANet-SRM [22], SINet [12], MirrorNet [51], ERRNet [20], CubeNet [59], D<sup>2</sup>CNet [42], C<sup>2</sup>FNet [39], LSR [29], MGL [55], PFNet [31], UAJNet [24] and UGTR [52]. Among them, ANet-SRM, SINet, MirrorNet, ERRNet, CubeNet, D<sup>2</sup>CNet, C<sup>2</sup>FNet, LSR, MGL, PFNet, UAJNet and UGTR are specially designed for COD detection. CPD, SCRNet, EGNNet, F3Net and MINet are designed for SOD tasks. PrNet is designed for polyp segmentation.

### 4.2.1 Quantitative comparison

Table 1 shows the quantitative results of different methods on four COD benchmark datasets. From the results, we can observe that UAJNet, as a state-of-the-art COD detection model, has superior metric performance than other competitors on the four COD benchmark datasets. Compared with UAJNet, our method has higher  $S_\alpha$ ,  $E_\phi^{\max}$ ,  $F_\phi^{\max}$ , smaller  $\mathcal{M}$  score, indicating that our method can identify camouflaged objects more accurately. Compared with UAJNet, our method has the highest scores except for the  $\mathcal{M}$  metric of CAMO dataset, the  $E_\phi^{\max}$  metric of CHAMELEON and COD10K.

### 4.2.2 Visual comparison

Figure 4 shows a visual comparison of the proposed method with other compared methods, and the red boxes represent the detection results of our model. Compared to other methods, our method is closer to the ground truth. Specifically, benefiting from the proposed TPRM, our method detects camouflaged objects more completely than the UGTR and PFNet visual results. At the same time, our SIEM can capture fine-grained cues, so our method can also detect accurate boundary features compared to other methods. All in all, compared with other methods, our proposed method better captures the semantic features and spatial features of camouflaged objects, thereby providing the best visual results, demonstrating its superiority in COD.

### 4.3 Ablation study

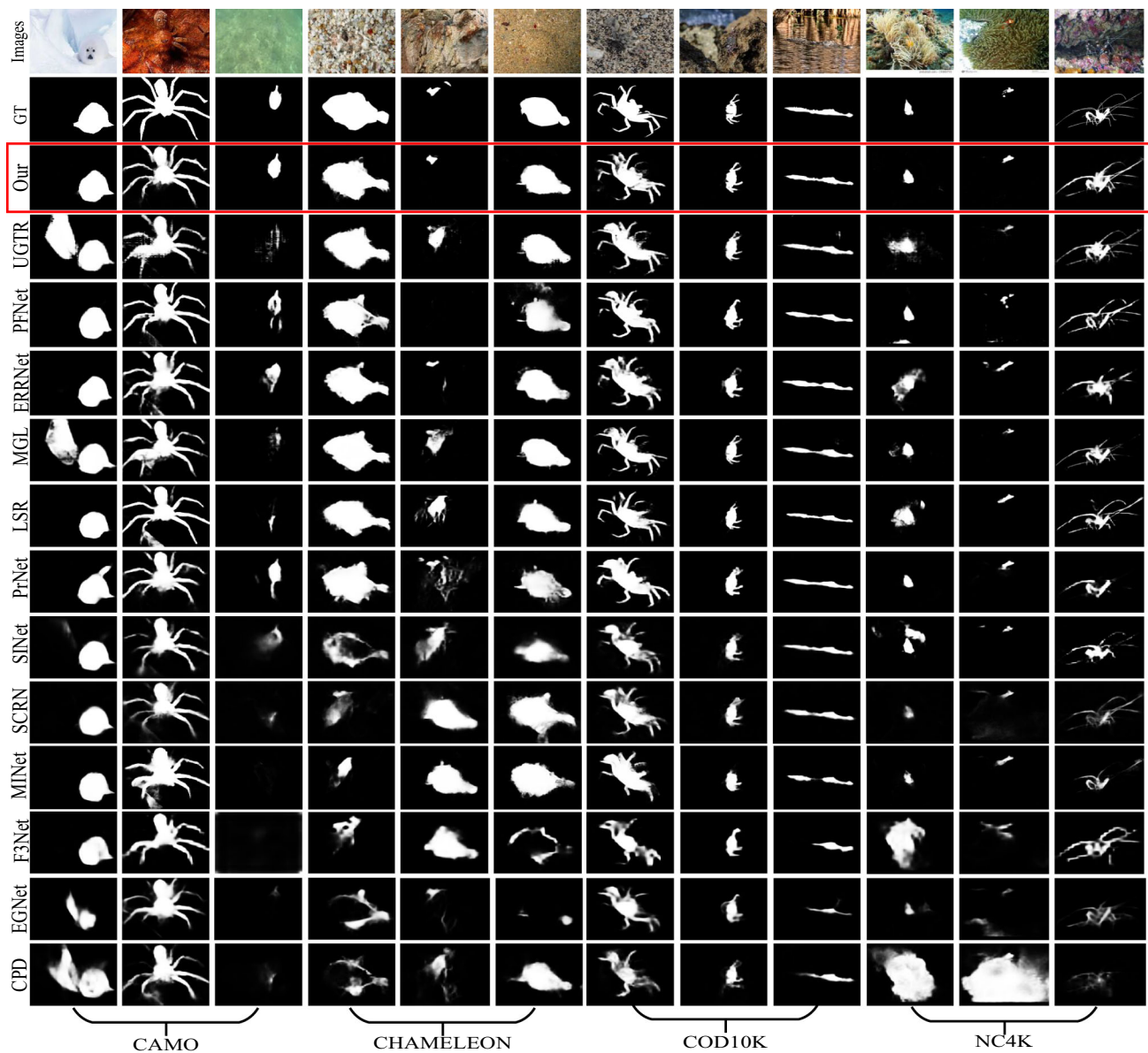
To verify the effectiveness of each key module, we designed four ablation experiments in Table 2. In No. 1, our Baseline model is a pre-trained Res2Net with three-layer RFB. In No. 2, we added TPRM based on Baseline. In No. 3, we replaced SIEM with multiplication. In No. 4, the experimental model is the complete TPRNet.

**Table 1** The performance comparison with eight state-of-the-art models on four datasets. maximum F-measure ( $F_{\phi}^{\max}$ , higher is better), maximum E-measure ( $E_{\phi}^{\max}$ , higher is better), S-measure ( $S_{\alpha}$ , higher is better), MAE( $\mathcal{M}$ , lower is better) are utilized to evaluate the performance of these models

Method	CAMO [22]				CHAMELEON [38]				COD10K [12]				NC4K [29]			
	$S_{\alpha}$	$\mathcal{M}$	$E_{\phi}^{\max}$	$F_{\phi}^{\max}$	$S_{\alpha}$	$\mathcal{M}$	$E_{\phi}^{\max}$	$F_{\phi}^{\max}$	$S_{\alpha}$	$\mathcal{M}$	$E_{\phi}^{\max}$	$F_{\phi}^{\max}$	$S_{\alpha}$	$\mathcal{M}$	$E_{\phi}^{\max}$	$F_{\phi}^{\max}$
CPD [48]	0.716	0.113	0.796	0.658	0.857	0.048	0.898	0.813	0.750	0.053	0.853	0.640	0.717	0.092	0.793	0.638
SCRN [49]	0.779	0.090	0.850	0.738	0.876	0.042	0.939	0.836	0.789	0.047	0.880	0.699	0.830	0.059	0.897	0.793
EGNet [58]	0.662	0.124	0.766	0.612	0.848	0.050	0.831	0.676	0.737	0.056	0.810	0.608	0.767	0.077	0.850	0.719
F3Net [46]	0.711	0.109	0.780	0.630	0.848	0.047	0.917	0.798	0.739	0.051	0.819	0.609	0.780	0.070	0.848	0.719
MINet [33]	0.748	0.090	0.838	0.706	0.855	0.036	0.937	0.821	0.770	0.042	0.859	0.671	0.812	0.056	0.887	0.775
PrNet [13]	0.769	0.094	0.837	0.728	0.860	0.044	0.935	0.830	0.789	0.045	0.879	0.704	0.828	0.058	0.883	0.775
ANet-SRM [22]	0.682	0.126	0.722	0.566	*	*	*	*	*	*	*	*	*	*	*	*
SINet [12]	0.751	0.100	0.831	0.706	0.869	0.044	0.935	0.830	0.771	0.051	0.868	0.676	0.808	0.058	0.883	0.775
MirrorNet [51]	0.785	0.077	0.850	0.756	*	*	*	*	*	*	*	*	*	*	*	*
ERRNet [20]	0.779	0.085	0.858	0.742	0.868	0.039	0.939	0.840	0.786	0.043	0.886	0.675	0.827	0.054	0.901	0.794
CubeNet [59]	0.788	0.085	0.860	0.750	0.873	0.037	0.945	0.849	0.795	0.041	0.883	0.715	*	*	*	*
D <sup>2</sup> CNet [42]	0.774	0.087	0.838	0.743	0.889	0.030	0.948	<b>0.868</b>	0.807	0.037	0.887	0.736	*	*	*	*
C <sup>2</sup> FNet [39]	0.796	0.080	0.864	0.771	0.888	0.032	0.946	0.863	0.813	0.036	<b>0.900</b>	0.743	0.838	0.049	0.904	0.810
LSR [29]	0.787	0.080	0.854	0.753	0.890	0.030	0.948	0.862	0.804	0.037	0.892	0.732	0.840	0.048	0.907	0.815
MGL [55]	0.775	0.088	0.842	0.740	0.893	0.030	0.941	0.860	0.814	0.035	0.890	0.738	0.833	0.052	0.893	0.800
PFNet [31]	0.782	0.085	0.855	0.758	0.882	0.033	0.945	0.853	0.800	0.040	0.890	0.725	0.829	0.053	0.898	0.799
UAJNet [24]	0.800	<b>0.073</b>	0.873	0.779	0.891	<b>0.030</b>	0.955	0.862	0.809	0.035	0.891	0.738	0.842	<b>0.047</b>	0.907	0.816
UGTR [52]	0.784	0.086	0.851	0.752	0.887	0.031	0.940	0.846	0.817	0.036	0.890	0.741	0.839	0.052	0.899	0.807
Ours	<b>0.814</b>	0.076	<b>0.875</b>	<b>0.783</b>	<b>0.901</b>	<b>0.029</b>	0.951	<b>0.868</b>	<b>0.829</b>	<b>0.034</b>	0.898	<b>0.752</b>	<b>0.854</b>	<b>0.047</b>	<b>0.908</b>	<b>0.821</b>

↑ Means the higher the score the better. ↓ Indicates the lower the score the better. \* Indicates the code or result is not available. Bold and italic denote the best and the second-best results, respectively





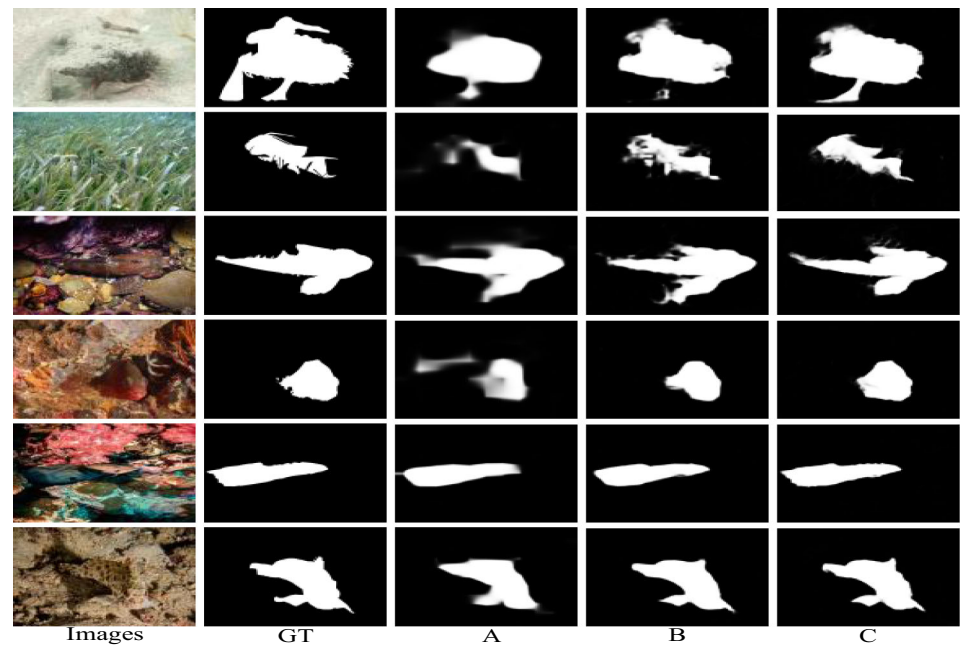
**Fig. 4** Visual comparison of COD detection results with eight models on the four public datasets. The red box is the visual effect result of our proposed TPRNet, and it can be clearly seen that our method has achieved excellent visual performance compared to other methods

**Table 2** Ablation study of our proposed model on COD10K and NC4K dataset

#	Baseline	TPRM	Mul	SIEM	COD10K				NC4K			
					$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$	$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$
No. 1	✓				0.792	0.044	0.883	0.688	0.831	0.056	0.899	0.785
No. 2	✓	✓			0.819	0.037	0.895	0.737	0.847	0.049	0.905	0.810
No. 3	✓	✓	✓		0.818	0.036	0.893	0.736	0.849	0.048	0.906	0.814
No. 4	✓	✓		✓	<b>0.829</b>	<b>0.034</b>	<b>0.898</b>	<b>0.752</b>	<b>0.854</b>	<b>0.047</b>	<b>0.908</b>	<b>0.821</b>

The *Mul* denotes semantic features, and spatial features are directly connected by multiplication. The best results are bolded

**Fig. 5** Visualization of the results of the ablation experiment. *A*, *B*, and *C* denote the Baseline, Baseline+TPRM and Baseline+TPRM+SIEM, respectively



**Table 3** Ablation experiments in TPRM on COD10K and NC4K dataset

#	PDC [48]	GRA [11]	RCU	Transformer [40]	COD10K				NC4K			
					$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$	$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$
No.1	✓		✓		0.826	0.035	0.895	0.747	0.852	<b>0.047</b>	0.906	0.816
No.2		✓		✓	0.825	0.036	0.899	0.745	0.852	0.048	0.905	0.815
No.3			✓	✓	<b>0.829</b>	<b>0.034</b>	<b>0.898</b>	<b>0.752</b>	<b>0.854</b>	<b>0.047</b>	<b>0.908</b>	<b>0.821</b>

The best results are bolded

#### 4.3.1 Effectiveness of TPRM

We investigate the importance of the TPRM. From Table 2, when adding TPRM to the baseline, the model performance improves significantly, clearly showing that aggregating multi-layer features and feature refinement is necessary for improving performance. Specifically, the metrics ( $S_\alpha$ ,  $\mathcal{M}$ ,  $E_\phi^{\max}$ ,  $E_\phi^{\max}$ ) is improved by 2.7%, 0.7%, 1.2%, 4.9% on the COD10K dataset, while the metrics ( $S_\alpha$ ,  $\mathcal{M}$ ,  $E_\phi^{\max}$ ,  $E_\phi^{\max}$ ) can be improved by 1.6%, 0.7%, 0.6%, 2.5% on the NC4K dataset (Fig. 5).

In addition, to verify the effectiveness of the components in TPRM, we replace Transformer and RCU with PDC [48] and GRA [11]. As shown in Table 3, comparing No. 1 and No. 3, Transformer is more suitable for integrating long-term dependencies of semantic information. Comparing No. 2 and No. 3, RCU is more suitable for feature refinement of camouflaged objects. At the same time, we show the RCU results of multiple stages in Fig. 6, and it can be seen that more detailed result graphs can be obtained by progressively passing through multiple RCU.

To verify the effect of multi-layer Transformer on TPRM, we set different layers to conduct ablation studies. As shown

in Table 4, we list the comparison results of  $T$  gradually increasing from 1 to 6. We can see that the performance is optimal when  $T = 4$ .

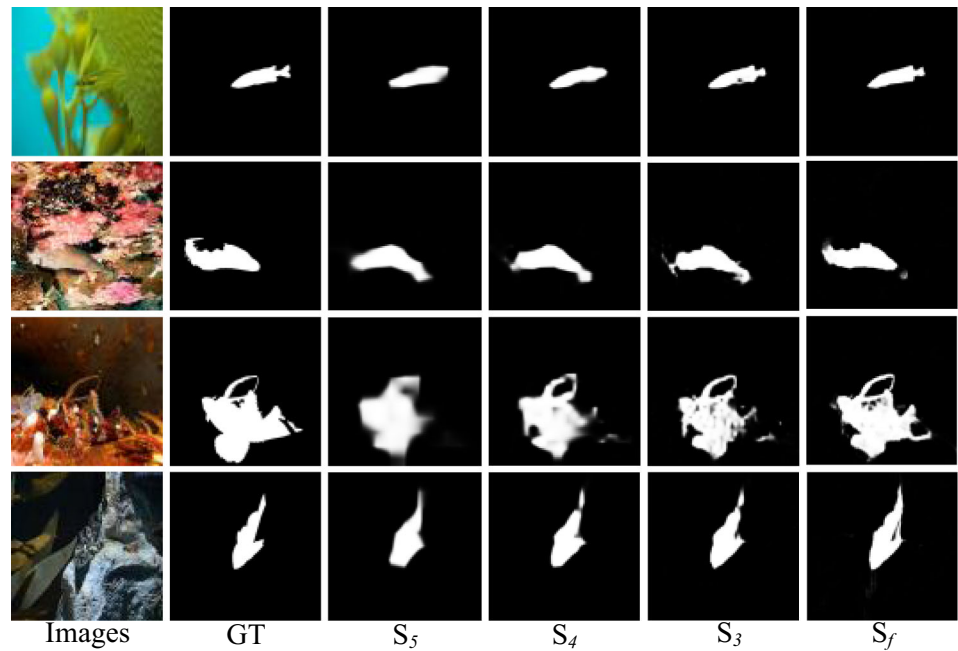
To demonstrate the effectiveness of channel split and concatenation, we set different grouping numbers, as shown in Table 5, and we see that the performance is best when  $n = 4$ .

In order to study the impact of different numbers of RCU on the model, as shown in Table 6, we set up multiple RCU in the height-level feature layer. It can be seen that when the number of RCU in each layer of advanced features is 1, the performance is the best.

#### 4.3.2 Effectiveness of SIEM

We further investigate the contribution of SIEM; we observe that compared to Baseline + TPRM, SIEM can improve metrics ( $S_\alpha$ ,  $\mathcal{M}$ ,  $E_\phi^{\max}$ ,  $E_\phi^{\max}$ ) by 1.0%, 0.3%, 0.3%, and 1.5% on COD10K dataset, while the metrics ( $S_\alpha$ ,  $\mathcal{M}$ ,  $E_\phi^{\max}$ ,  $E_\phi^{\max}$ ) can be improved by 0.7%, 0.2%, 0.3%, 1.1% on the NC4K dataset. In No. 3, we replace SIEM with multiplication and let the semantic features interact directly with the spatial features. The experimental results show that the direct interaction may introduce noise and does not improve the

**Fig. 6** Visual results at different stages of RCU.  $S_3$ ,  $S_4$ , and  $S_5$  represent the refinement results of the RCU of the fifth layer, the RCU of the fourth layer, and the RCU of the third layer, respectively.  $S_f$  denotes the final prediction



**Table 4** Ablation study of transformers in TPRM on COD10K and NC4K dataset

#	Transformer	COD10K				NC4K			
		$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$	$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$
No.1	$T = 1$	0.824	0.036	0.896	0.743	0.852	<b>0.047</b>	<b>0.908</b>	0.817
No.2	$T = 2$	0.826	0.035	<b>0.898</b>	0.747	0.850	0.048	0.904	0.817
No.3	$T = 3$	0.823	0.035	0.896	0.742	0.852	<b>0.047</b>	0.906	0.816
No.4	$T = 4$	<b>0.829</b>	<b>0.034</b>	<b>0.898</b>	<b>0.752</b>	<b>0.854</b>	<b>0.047</b>	<b>0.908</b>	<b>0.821</b>
No.5	$T = 5$	0.826	0.035	0.897	0.747	0.852	<b>0.047</b>	0.907	0.818
No.6	$T = 6$	0.826	<b>0.034</b>	0.900	0.749	0.850	<b>0.047</b>	0.906	0.816

The best results are bolded

**Table 5** Ablation study of the number of channel groupings of feature  $P_i$  in the reverse feature refinement strategy in TPRM

#	Group size	COD10K				NC4K			
		$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$	$S_\alpha$	$\mathcal{M}$	$E_\phi^{\max}$	$F_\phi^{\max}$
No.1	$n = 1$	0.824	0.036	0.896	0.743	0.853	0.048	0.908	0.820
No.2	$n = 2$	0.824	0.036	0.896	0.745	0.852	0.048	0.904	0.815
No.3	$n = 4$	<b>0.829</b>	<b>0.034</b>	<b>0.898</b>	<b>0.752</b>	<b>0.854</b>	<b>0.047</b>	<b>0.908</b>	<b>0.821</b>
No.4	$n = 8$	0.824	0.037	0.892	0.742	0.854	0.048	0.906	0.818
No.5	$n = 16$	0.826	0.036	0.898	0.745	0.854	0.047	0.907	0.818
No.6	$n = 32$	0.825	0.035	0.893	0.742	0.851	0.047	0.906	0.815

The best results are bolded

**Table 6** Ablation experiments on the number of RCUs in each level (high-level features)

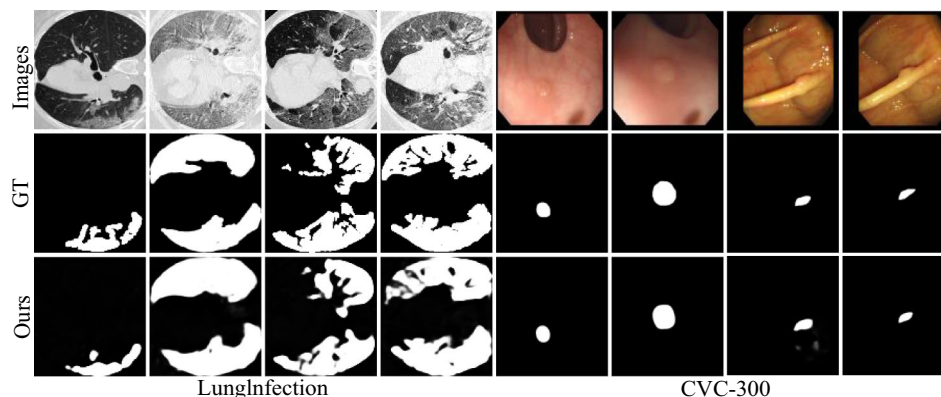
#	Different components	COD10K				NC4K			
		$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^{\max} \uparrow$	$F_\phi^{\max} \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^{\max} \uparrow$	$F_\phi^{\max} \uparrow$
No. 1	<i>A</i>	<b>0.829</b>	<b>0.034</b>	<b>0.898</b>	<b>0.752</b>	<b>0.854</b>	0.047	<b>0.908</b>	<b>0.821</b>
No. 2	<i>B</i>	0.826	0.035	0.897	0.746	0.853	<b>0.046</b>	0.905	0.815
No. 3	<i>C</i>	0.823	<b>0.034</b>	0.896	0.742	0.852	0.047	0.907	0.820

*A* represents our proposed method, and only one RCU is included in each layer of high-level features. *B* represents that each layer of high-level features contains two RCUs. *C* represents three RCUs in each layer of high-level features. The best results are bolded

**Table 7** Comparison with state-of-the-art methods in FLOPs, Parameters and FPS

Model Type	SINet <i>C</i>	C2FNet <i>C</i>	LSR <i>C</i>	MGL <i>C</i>	PFNet <i>C</i>	UGTR <i>T</i>	Ours <i>C&amp;T</i>
FLOPs (G)	38.8	26.2	66.6	553.9	53.2	1007.0	25.8
Params (M)	48.9	28.4	50.9	63.6	46.5	48.9	33.1
FPS	29.7	32.1	31.6	6.3	34.5	8.5	27.6
COD10K( $F_{\phi}^{\max} \uparrow$ )	0.676	0.743	0.732	0.738	0.725	0.741	0.752

'*C*' represents the COD model based on CNN, '*T*' represents the COD model on Transformer, and '*C&T*' represents the COD model combined with CNN and Transformer

**Fig. 7** Visual Results in Medical Segmentation. The right side is the polyp segmentation, the left side is the lung infection segmentation, and the third row is the detection result of the proposed TPRNet

performance. Overall, our proposed SIEM is more capable of capturing fine-grained cues.

We provide result maps at different stages, demonstrating the effectiveness of the proposed module. As shown in Fig. 5, the baseline model can detect camouflaged objects, but this is not satisfactory. When adding TPRM to the baseline, we can observe that more comprehensive camouflaged regions can be located, but some boundary features are missing. The network can capture some boundary features when adding TPRM and SIEM to the baseline. It can be seen that the proposed three modules effectively facilitate camouflaged object detection.

#### 4.3.3 Comparison of efficiency

As listed in Table 7, we also compare our proposed method with recently proposed CNN-based COD models and Transformer-based models in FLOPs, Params, and FPS. Our method is very competitive in FLOPs, Params, and FPS compared to CNN-based COD models. Similarly, compared with the Transformer-based COD model, our method significantly reduces both FLOPs and Params while greatly improving inference speed. In conclusion, our method outperforms CNN-based COD and Transformer-based COD models regarding FLOPs, Params, FPS, and performance.

#### 4.4 Potential applications

Generally, camouflage object detection has potential application value in clinical medical diagnosis. As shown in Fig. 7,

we will test the proposed TPRNet to the polyp segmentation dataset and the lung infection segmentation dataset. It can be seen that TPRNet can accurately segment out the diseased area.

## 5 Conclusions

In this paper, we propose a transformer-induced progressive refinement network (*TCNet*) for camouflaged object detection. Our network includes a transformer-induced progressive refinement module (TPRM) and a semantic-spatial interaction module (SIEM), where TPRM is mainly responsible for capturing semantic features, and SIEM is responsible for providing spatial information. Extensive experiments on four challenging datasets of COD datasets (i.e., CAMO, CHAMELEON, COD10K, and NC4K) under four widely-used evaluation metrics demonstrate that the proposed TPRNet outperforms the eighteen state-of-the-art methods for camouflaged object detection. Furthermore, we hope that the proposed method can be applied in more engineering fields.

**Acknowledgements** The paper is supported by Anhui Province Key Laboratory of Infrared and Low-Temperature Plasma under No. IRKL2022KF07.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.



## References

- Amit, S.N.K.B., Shiraishi, S., Inoshita, T., Aoki, Y.: Analysis of satellite images for disaster detection. In: 2016 IEEE International geoscience and remote sensing symposium (IGARSS), pp. 5189–5192. IEEE (2016)
- Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint [arXiv:1607.06450](https://arxiv.org/abs/1607.06450) (2016)
- Bi, H., Wang, K., Lu, D., Wu, C., Wang, W., Yang, L.: C 2 net: a complementary co-saliency detection network. *Vis. Comput.* **37**(5), 911–923 (2021)
- Bi, H., Zhang, C., Wang, K., Tong, J., Zheng, F.: Rethinking camouflaged object detection: models and datasets. *IEEE Trans. Circuits Syst. Video Technol.* (2021). <https://doi.org/10.1109/TCSVT.2021.3124952>
- Cui, Y., Cao, Z., Xie, Y., Jiang, X., Tao, F., Chen, Y.V., Li, L., Liu, D.: Dg-labeler and dgl-mots dataset: Boost the autonomous driving perception. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 58–67 (2022)
- Cui, Y., Yan, L., Cao, Z., Liu, D.: Tf-blender: Temporal feature blender for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8138–8147 (2021)
- Dong, B., Zhuge, M., Wang, Y., Bi, H., Chen, G.: Towards accurate camouflaged object detection with mixture convolution and interactive fusion. arXiv preprint [arXiv:2101.05687](https://arxiv.org/abs/2101.05687) 1(2) (2021)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
- Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp. 4548–4557 (2017)
- Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint [arXiv:1805.10421](https://arxiv.org/abs/1805.10421) (2018)
- Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021). <https://doi.org/10.1109/TPAMI.2021.3085766>
- 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. 2777–2787 (2020)
- 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: International conference on medical image computing and computer-assisted intervention, pp. 263–273. Springer (2020)
- Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: automatic covid-19 lung infection segmentation from ct images. *IEEE Trans. Med. Imaging* **39**(8), 2626–2637 (2020)
- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146–3154 (2019)
- Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: a new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(2), 652–662 (2019)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
- Hou, J.Y.Y.H.W., Li, J.: Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Eng.* **15**, 2201–2205 (2011)
- Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141 (2018)
- Ji, G.P., Zhu, L., Zhuge, M., Fu, K.: Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recogn.* **123**, 108414 (2022)
- Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
- Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.* **184**, 45–56 (2019)
- Le, X., Mei, J., Zhang, H., Zhou, B., Xi, J.: A learning-based approach for surface defect detection using small image datasets. *Neurocomputing* **408**, 112–120 (2020)
- Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., Dai, Y.: Uncertainty-aware joint salient object and camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10071–10081 (2021)
- Liu, D., Cui, Y., Chen, Y., Zhang, J., Fan, B.: Video object detection for autonomous driving: Motion-aid feature calibration. *Neurocomputing* **409**, 1–11 (2020)
- Liu, D., Cui, Y., Tan, W., Chen, Y.: Sg-net: Spatial granularity network for one-stage video instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9816–9825 (2021)
- Liu, Z., Huang, K., Tan, T.: Foreground object detection using top-down information based on em framework. *IEEE Trans. Image Process.* **21**(9), 4204–4217 (2012)
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
- 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. 11591–11601 (2021)
- Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 248–255 (2014)
- 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. 8772–8781 (2021)
- Pan, Y., Chen, Y., Fu, Q., Zhang, P., Xu, X.: Study on the camouflaged target detection method based on 3d convexity. *Mod. Appl. Sci.* **5**(4), 152 (2011)
- Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9413–9422 (2020)
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
- Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition, pp. 733–740. IEEE (2012)
- Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12179–12188 (2021)
- Sengottuvelan, P., Wahi, A., Shanmugam, A.: Performance of decamouflaging through exploratory image analysis. In: 2008 First International Conference on Emerging Trends in Engineering and Technology, pp. 6–10. IEEE (2008)



38. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Koziel, P.: Animal camouflage analysis: Chameleon database. Unpublished manuscript 2(6), 7 (2018)
39. Sun, Y., Chen, G., Zhou, T., Zhang, Y., Liu, N.: Context-aware cross-level fusion network for camouflaged object detection. arXiv preprint [arXiv:2105.12555](https://arxiv.org/abs/2105.12555) (2021)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
41. Wang, D., Hu, G., Lyu, C.: Frnet: an end-to-end feature refinement neural network for medical image segmentation. *Vis. Comput.* **37**(5), 1101–1112 (2021)
42. Wang, K., Bi, H., Zhang, Y., Zhang, C., Liu, Z., Zheng, S.: D 2 c-net: a dual-branch, dual-guidance and cross-refine network for camouflaged object detection. *IEEE Trans. Ind. Electron.* **69**, 5364 (2021)
43. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578 (2021)
44. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803 (2018)
45. Wang, X., Wang, W., Bi, H., Wang, K.: Reverse collaborative fusion model for co-saliency detection. *The Visual Computer* pp. 1–11 (2021)
46. Wei, J., Wang, S., Huang, Q.: F<sup>3</sup>net: fusion, feedback and focus for salient object detection. *Proc. AAAI Conf. Artif. Intell.* **34**, . 12321–12328 (2020)
47. Wu, Y.H., Gao, S.H., Mei, J., Xu, J., Fan, D.P., Zhang, R.G., Cheng, M.M.: Jcs: an explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Trans. Image Process.* **30**, 3113–3126 (2021)
48. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3907–3916 (2019)
49. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7264–7273 (2019)
50. Xiao, H., Ran, Z., Mabu, S., Li, Y., Li, L.: Saunet++: an automatic segmentation model of covid-19 lesion from ct slices. *Vis. Comput.* (2022). <https://doi.org/10.1007/s00371-022-02414-4>
51. Yan, J., Le, T.N., Nguyen, K.D., Tran, M.T., Do, T.T., Nguyen, T.V.: Mirrornet: bio-inspired camouflaged object segmentation. *IEEE Access* **9**, 43290–43300 (2021)
52. Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., Fan, D.P.: Uncertainty-guided transformer reasoning for camouflaged object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4146–4155 (2021)
53. Youwei, P., Xiaoqi, Z., Tian-Zhu, X., Lihe, Z., Huchuan, L.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. arXiv preprint [arXiv:2203.02688](https://arxiv.org/abs/2203.02688) (2022)
54. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567 (2021)
55. 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. 12997–13007 (2021)
56. Zhang, X., Wang, X., Gu, C.: Online multi-object tracking with pedestrian re-identification and occlusion processing. *Vis. Comput.* **37**(5), 1089–1099 (2021)
57. Zhang, Y., Han, S., Zhang, Z., Wang, J., Bi, H.: Cf-gan: cross-domain feature fusion generative adversarial network for text-to-image synthesis. *Vis. Comput.* (2022). <https://doi.org/10.1007/s00371-022-02404-6>
58. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnnet: Edge guidance network for salient object detection. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8779–8788 (2019)
59. Zhuge, M., Lu, X., Guo, Y., Cai, Z., Chen, S.: Cubenet: X-shape connection for camouflaged object detection. *Pattern Recogn.* **127**, 108644 (2022)

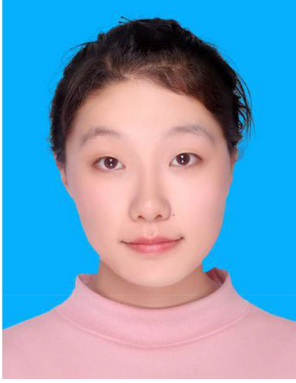
**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Qiao Zhang** is pursuing his master degree at the Northeast Petroleum University, Daqing, China. His current research interests include co-saliency detection, Camouflaged Object Detection and deep learning.



**Yanliang Ge** received his bachelor degree in 2002 from NorthEast Petroleum University in communications engineering. He received his master degree in 2008 from NorthEast Petroleum University in Oil and gas information and control engineering. Currently he is a Associate professor in School of Electrical Information Engineering in NorthEast Petroleum University. His main research interests contains digital watermarking, signal processing, digital video processing, etc.



**Cong Zhang** is pursuing her master degree at the Northeast Petroleum University, Daqing, China. Her current research interests include camouflaged object detection and deep learning.



**Hongbo Bi** received Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2013. He worked as a Postdoc Fellow (PDF) with Harbin Engineering University, Harbin, China, from 2014 to 2017. He worked as a Visiting Scholar with the University of Waterloo, Waterloo, ON, Canada, from 2014 to 2015. He is currently an Associate Professor with the School of Electrical Information Engineering, NorthEast Petroleum University. His research interests include

camouflaged object detection, saliency detection, computer vision, deep learning, etc.