Integrating Part-Object Relationship and Contrast for Camouflaged Object Detection

Yi Liu, Dingwen Zhang, *Member, IEEE*, Qiang Zhang, and Jungong Han

***Abstract*— Object detectors that solely rely on image contrast are struggling to detect camouflaged objects in images because of the high similarity between camouflaged objects and their surroundings. To address this issue, in this paper, we investigate the role of the part-object relationship for camouflaged object detection. Specifically, we propose a Part-Object relationship and Contrast Integrated Network (POCINet) covering both search and identification stages, where each stage adopts an appropriate scheme to engage the contrast information and part-object relational knowledge for camouflaged pattern decoding. Besides, we bridge these two stages via a Search-to-Identification Guidance (SIG) module, in which the search result, as well as decoded semantic knowledge, jointly enhances the features encoding ability of the identification stage. Experimental results demonstrate the superiority of our algorithm on three datasets. Notably, our algorithm raises F_β of the best existing method by approximately 17 points on the CPD1K dataset. The source code will be released soon.**

***Index Terms*— Camouflaged object detection, contrast, part-object relationships, encoder-decoder, multi-stage.**

I. Introduction

IN VISUAL surveillance [1]–[6], camouflaged object detection is an interesting yet challenging task, where the goal is to search and segment out those objects concealed in their surroundings. High intrinsic similarity between the target object and the background makes camouflaged object detection much more challenging than the traditional visual detection tasks, such as salient object detection [7]–[15] and generic object detection [16]–[18]. Recently, camouflaged object detection has been receiving increasing attention due to its potential applications in real-life scenarios, including wild animals preservation, new species discovery, medical image segmentation [19], integrated circuits testing [20], [21], and art [22], [23], *etc*.

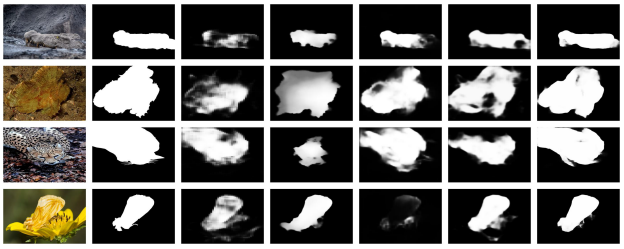


Fig. 1. Problem illustrations for the contrast and part-object relational cues used for camouflaged object detection. (a) Image; (b) GT; (c) PSPNet [28]; (d) BASNet [29]; (e) SINet [27]; (f) TSPOANet [30]; (g) OURS. Contrast-induced approaches (*i.e.*, (c)–(e) in the figure) miss some object parts, especially the object boundaries. Part-object relational method (*i.e.*, (f) in the figure) produces blurry object boundaries and some "holes".

Early camouflaged object detectors attempt to extract discriminative hand-crafted features [24], *e.g.*, color, edge, and texture. Despite their simpleness, such features have limited representation capacity in extracting useful visual patterns. In light of the extraordinary representation ability of deep features, research focus has shifted recently onto deep learning based frameworks for camouflaged object detection [25]–[27]. From a systemic perspective, these methods explore rich distinguishable features with primitive contrast information to identify the camouflaged object in a scene. However, a striking resemblance between foreground and background challenges the extraction of distinguishable features, giving rise to a failure to recognize the camouflaged object from the background. For instance, as shown in Fig. 1(c)-(e), some parts of the camouflaged object, especially the object boundaries, cannot be identified from its surroundings, resulting in the incomplete segmentation of the camouflaged object. The above observation reveals that the exploitation of the contrast information only could not be able to solve the problem.

In nature, an object is composed of several relevant parts, and on the other hand, associated parts can form a whole object. Such part-object relational property can aid in addressing the above problem of incomplete segmentation. Especially in [30], it successfully captures the complete salient object by finding relevant object parts, rather than distinctive regions, in a scene. Inspired by its promising results, this paper takes the initiative to incorporate such part-object relational property into camouflaged object detection. Fig. 1 shows

Manuscript received April 10, 2021; revised August 12, 2021 and October 15, 2021; accepted October 15, 2021. Date of publication November 4, 2021; date of current version November 11, 2021. This work was supported by the National Natural Science Foundation of China under Grant 62001341, Grant 61773301, and Grant 61876140. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alessandro Piva. *(Corresponding authors: Qiang Zhang; Jungong Han.)*

Yi Liu is with the School of Computer Science and Artificial Intelligence, the Aliyun School of Big Data, and the School of Software, Changzhou University, Changzhou, Jiangsu 213164, China (e-mail: liuyi0089@gmail.com).

Dingwen Zhang is with the School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi 710071, China (e-mail: zhangdingwen2006yyy@gmail.com).

Qiang Zhang is with the School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: qzhang@xidian.edu.cn).

Jungong Han is with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K. (e-mail: jungonghan77@gmail.com).

Digital Object Identifier 10.1109/TIFS.2021.3124734

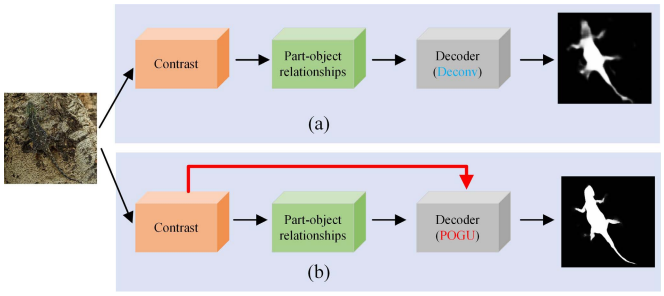


Fig. 2. Illustration for different interactions between contrast and part-object relational cues. (a) TSPOANet [30]; (b) OURS. The deconvolutional decoding mechanism pays ZERO attention to the engagement of the contrast and part-object relational cues. Alternatively, we integrate these two cues in the decoder via the proposed POGU module, which will be introduced in Sec. III.

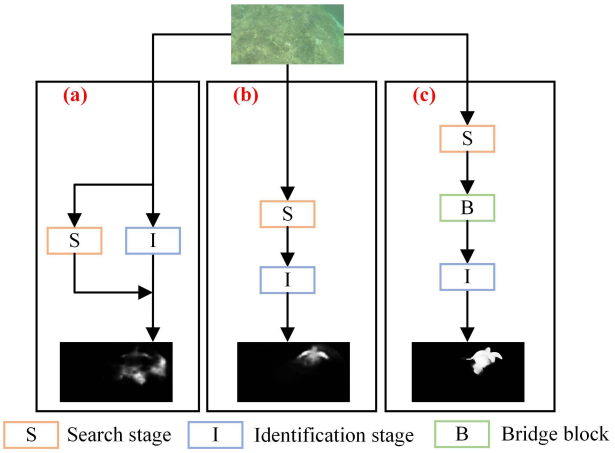


Fig. 3. Different camouflaged object detection pipelines. (a) An identification branch and a search branch are parallelized; (b) The features of the search branch are directly fed into the identification branch; (c) OURS: A bridge block is put in place to connect the search and identification stages.

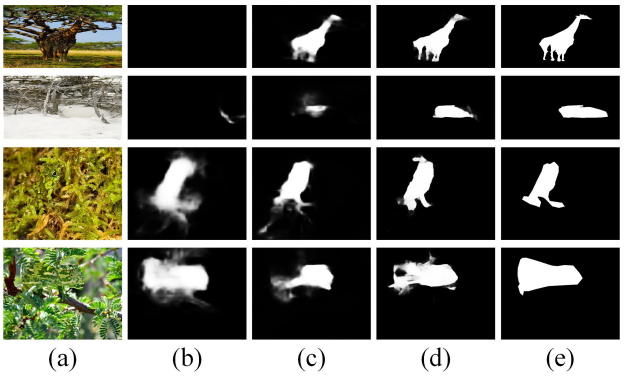


Fig. 4. Problem illustrations for different camouflaged object detection pipelines. (a) Image; (b) Parallel [26]; (c) Direct feeding [27]; (d) OURS; (e) GT. A parallel of two may cause a failure to search the camouflaged object (top two rows in the figure) or background noises (bottom two rows in the figure). Direct feeding may cause object details missed. Our framework can cater to these problems with accurate object locations and sufficient object details.

the comparison between the part-object relational approach (TSPOANet[1] [30], *i.e.*, Fig. 1(f)) and three contrast-induced approaches (Fig. 1(c)-(e)). It is clear that TSPOANet [30] is advantageous, especially when considering the completeness of the segmented objects. However, a closer look at the result generated by directly using TSPOANet [30] for camouflaged object detection reveals that it is far from satisfactory due to: 1) blurry object boundaries; 2) "holes" within the object. We further analyze the decoder of TSPOANet [30], and realize that the deconvolutional decoding mechanism pays ZERO attention to the engagement of part-object relational knowledge and contrast cues, which is shown in Fig. 2(a). No interaction of these two informative cues may adversely make both of their representation abilities weaker after prolonged deconvolutions, thus leading to the issues of the blurry boundary and the inner details deletion.

To address the above problem, in this paper, we propose a Part-Object relationship and Contrast Integrated Network (POCINet) containing two stages of search and identification for camouflaged object detection, each of which adequately engages contrast information and part-object relational knowledge during decoding. Specifically, the encoder cascades a Contrast Information Exploration (CIE) subnetwork and a Part-Object Relationship Exploration (PORE) subnetwork. In such a way, contrast features learned by CIE could provide rich features for PORE at the deep layers to explore part-object relational knowledge. The decoder enables the combination of the two critical information via a Part-Object relationship Guidance Upsampling (POGU) module, which is shown in Fig. 2(b). Concretely, the part-object relational knowledge provides object completeness as prior information, which in turn guides the contrast features to achieve more primitive camouflaged cues. Doing so helps to locate the camouflaged object by extracting tight relevant object boundaries in the search stage, as well as grabbing inner object details and complete object shapes in the identification stage.

In addition, the existing approaches either parallel a classification network and a segmentation network [26] (as shown in Fig. 3(a)) or directly feed features from the search network into the identification network [27] (as shown in Fig. 3(b)) without equipping the well-coupled search and identification streams. These pipelines may cause some problems. As shown in Fig. 4, the parallel pipeline may cause a failure to serach the camouflaged object (top two rows of Fig. 4) or background noises (bottom two rows of Fig. 4), thereby having no abilities to segmenting the camouflaged object out from the scene. As shown in the second and fourth rows of Fig. 4, the direct feeding pipeline may cause some object details missed. To solve this problem, we develop a Search-to-Identification Guidance (SIG) module to bridge these two stages (as shown in Fig. 3(c)), resulting in a SIG-induced CIE subnetwork in the identification stage. Here, the search result and the hybrid features decoded from the search stage are employed to empower the feature extraction ability of the identification stage. As highlighted in Fig. 1 and Fig. 4, such a framework enables our model to detect more complete camouflaged objects than other models do.

To sum up, the contributions of this paper are described as follows.

[1]We re-train TSPOANet [30] using the camouflaged object detection benchmarks.

(1) We propose a POCINet to detect the camouflaged object, enabling to integrate the contrast information and part-object relational knowledge via a POGU module. To the best of our knowledge, it is the earliest attempt to involve the part-object relational property in camouflaged object detection, and also initiates the integration of the deep contrast and part-object relational cues.

(2) We present a SIG module to bridge the search and identification stages for camouflaged object detection so that the search result and decoded camouflaged cues of the search stage empower the feature extraction ability of the identification stage.

(3) Extensive experiments demonstrate the superiority of our algorithm on three datasets. Especially on the CPD1K [31] dataset, our algorithm surpasses the best existing algorithm up to 17 points on F_β.

The paper is organized as follows. Sec. II reviews the works related to our method. Sec. III details the proposed camouflaged object detection network. Sec. IV conducts experiment and analysis to evaluate the proposed method. Sec. V concludes this paper.

II. Related Work

In this section, we review the works that are highly relevant to our method, covering camouflaged object detection, part-object representation, and multi-stage strategy.

A. Camouflaged Object Detection

Research devoted to camouflaged region detection has a long and rich history [32], [33]. For example, Pan *et al.* [34] attempted to detect camouflaged objects via a 3D convexity model. Liu *et al.* [35] detected the foreground object by optimizing top-down information. Sengottuvelan *et al.* [35] recognized the camouflaged object via a co-occurrence matrix method. An overall review of this history can be found in [36].

The above methods encountered performance bottlenecks because of the limited representation abilities of handcrafted features. In view of the powerful representation ability of deep features, Zheng *et al.* [25] explored the possibility of using a deep CNN to detect camouflaged people. From the biological perspective, Le *et al.* [26] developed a segmentation stream to segment the camouflaged object out, and a classification stream to recognize the existence of the camouflaged object in parallel. Fan *et al.* [27] consolidated this research by proposing a deep search identification network. Despite growing interests, the research on deep camouflaged object detection requires more dedication, given unsatisfactory results until now.

Our work differs from the existing works in two aspects: 1) Instead of solely relying on deep contrast semantics for camouflaged object detection, we involve the part-object relational property in camouflaged object detection, and further integrate the contrast cues and part-object relational cues to predict the object details and object shape of the camouflaged object in the complex scene. 2) Rather than simply connecting the search and identification streams, we design a more sophisticated connection, *i.e.*, SIG, to bridge the gap between the search and identification stages.

B. Part-Object Representation

The study of part-object representation can date back several decades ago. For example, Biederman [37] proposed a recognition-by-component theory for image understanding. Krivic and Solina [38] recognized articulated objects based on part-level descriptions obtained by the Segmentor system [39]. Pentland [40] segmented an image into roughly convex component parts for further recognition and prediction via 3D deformable models. Felzenszwalb [41] used the deformable part models for cascade object detection. Girshick [42] designed a CNN to formulate the deformable part model using a distance transform pooling, object geometry filters, and maxout units. Hinton *et al.* [43]–[45] explored the part-whole spatial relationships by a Capsule Network (CapsNet), which routes low-level capsules (parts) to their familiar high-level ones (wholes). Liu *et al.* [30] involved the part-object relational property to solve the incomplete segmentation problem of salient object detection.

Inspired by [30] that explored the part-object relationships encoded in CapsNet [45] for salient object segmentation, in this paper, we adopt CapsNet as the PORE subnetwork. However, beyond simple exploitation of several deconvolutions for decoding [30], our decoder integrates the contrast information and the part-object relational knowledge, thus helping to predict the complete object shape with sufficient object details.

C. Multi-Stage Strategies

Multi-stage networks have been widely used and explored in many computer vision tasks. For example, Cheng *et al.* [46] proposed a multi-stage encoder-decoder structure for semantic segmentation, where a supervise-and-excite framework was designed to connect two stages. Newell *et al.* [47] stacked multiple hourglass networks for pose estimation. Yu *et al.* [48] repeatedly applied the segmentation probability map from the previous iterations as spatial priors to refine the current iteration. Shen *et al.* [49] utilized multiple side outputs with different-size receptive fields from the lower stage to provide multi-scale contextual boundary information for the consecutive learning. Wang *et al.* [50] refined the salient object detection performance via multiple stages, in each of which a refinement network merged high-level semantic knowledge encoded by the master network with rich low-level features encoded by the refinement network. Deng *et al.* [51] refined the initial saliency prediction map with a sequence of residual refinement blocks.

In contrast, we simultaneously adopt the coarse camouflaged map and features of the search stage to enhance the ability of feature extraction of the following identification stage.

III. Proposed Framework

Fig. 5 exhibits the overall architecture of the proposed deep camouflaged object detection network, which consists of the search and identification stages. Corresponding to each stage, a CIE subnetwork and a PORE subnetwork are cascaded in the encoder. On top of that, these two camouflaged semantics are integrated into the decoder via a POGU module. Especially,

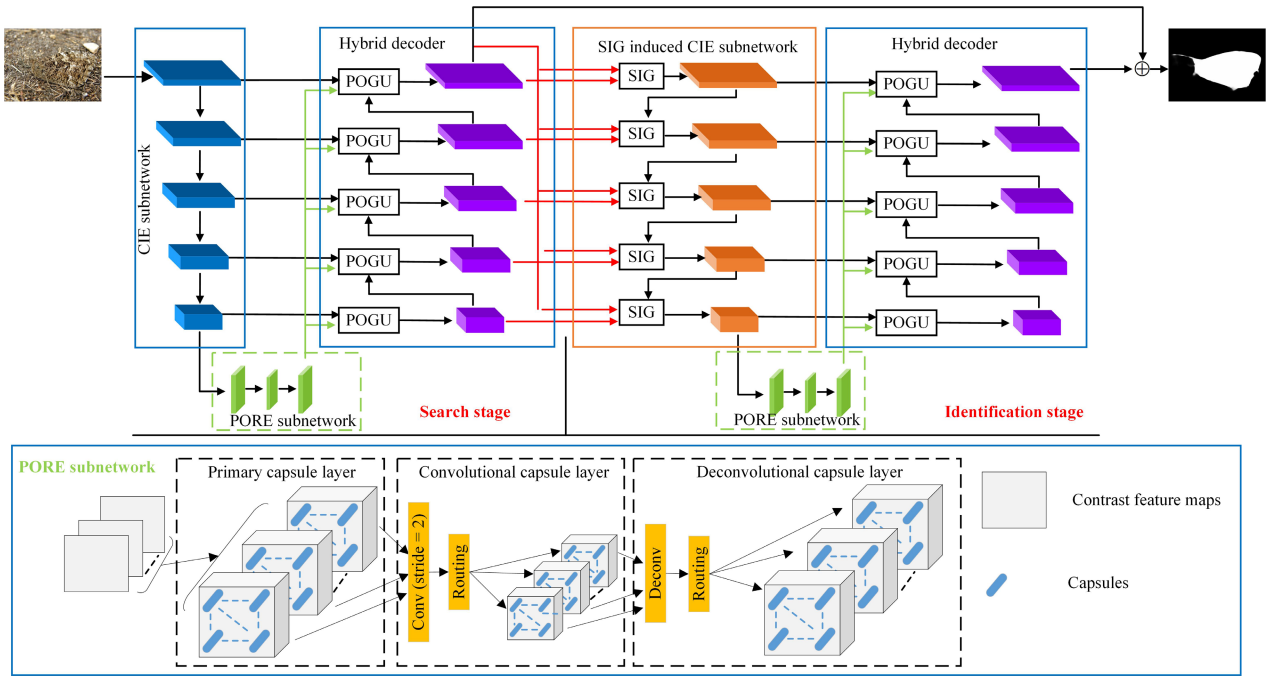


Fig. 5. The overall architecture of the proposed framework, which consists of two stages for search and identification. Within each stage, a CIE subnetwork and a PORE subnetwork are cascaded in the encoder, while these two camouflaged semantics are integrated into the decoder via a POGU module. Besides, a SIG module is designed to connect two stages, resulting in a SIG-induced CIE subnetwork in the identification stage. The final camouflaged map is achieved by integrating those maps output by two stages.

we bridge these two stages via a SIG module, resulting in a SIG-induced CIE subnetwork in the identification stage. The final camouflaged map is achieved by integrating those maps output by two stages. The details will be elaborated in the following.

A. Search Stage

The search stage of POCINet consists of a CIE subnetwork and a PORE subnetwork in the encoder, and a POGU module in the decoder.

1) CIE Subnetwork: For the search stage, the CIE subnetwork stacks five layers with different sized receptive fields to encode different semantics. Following the previous works, we choose the VGG16 [52] model as the backbone network. Concretely, five layers are initialized by Conv1_2, Conv2_2, Conv3_3, Conv4_3, and Conv5_3 of the pretrained VGG16 [52] model, respectively.

2) PORE Subnetwork: Inspired by the success of the exploration of the part-object relationship by CapsNet in solving the problem of incomplete object segmentation in [30], we adopt CapsNet to implement PORE subnetwork. Concretely, we involve a mirror CapsNet to capture the part-object relational cues. As shown in the PORE subnetwork of Fig. 5, the features obtained by the CIE subnetwork are transformed into capsule feature maps by a Primary Capsule (*PrimaryCaps*) layer. On top of that, a Convolutional Capsule (*ConvCaps*) layer and a Deconvolutional Capsule (*DeconvCaps*) layer are designed for capsules routing via the EM routing algorithm [53] to form a mirror CapsNet, which is aimed to explore the part-object relationships of the input image. The details of *PrimaryCaps* and *ConvCaps* can be found in [30]. Especially in the *DeconvCaps* layer, the previous capsule feature maps are upsampled for routing, which can output high-resolution capsule feature maps while retaining part-object relationships.

3) POGU: The encoded contrast semantics and part-object relational semantics help to capture the object details and the object completeness, respectively. Therefore, two kinds of semantic knowledge can complement each other. Considering this point, the decoder is designed to integrate these two semantics with the purpose of generating more primitive camouflaged cues for further prediction, which is implemented by a POGU module. The architecture of POGU is shown in Fig. 6, which consists of three phases, *i.e.*, features combination, self-attention promotion, and Part-Object Relationship (POR) guidance. Suppose $\mathbf{X}_{\mathrm{CIE}}$ and $\mathbf{X}_{\mathrm{PORE}}$ are the encoded contrast and part-object relational semantics, respectively, and $\mathbf{X}_{\mathrm{Dec}}$ is the decoded deep-level semantics. W, H, and C represent the width, height, and channel number of the corresponding feature maps, respectively. Each phase will be elaborated in the following.

a) Features combination: Features combination intends to incorporate the encoded contrast semantics and the decoded semantics. As illustrated in Fig. 6, the combined features $\mathbf{F}_{\mathrm{Com}} \in R^{W\times H\times C}$ can be computed by

$$\mathbf{F}_{\mathrm{Com}} = f_{\mathrm{Cat}}\left(f_{\mathrm{Dia}}\left(\mathbf{X}_{\mathrm{CIE}}; \mathbf{W}_{\mathrm{Dia}}\right), f_{\mathrm{U}}\left(\mathbf{X}_{\mathrm{Dec}}; \mathbf{W}_{\mathrm{U}}\right); \mathbf{W}_{\mathrm{Cat}}\right), \tag{1}$$

where f_{Dia} is a stack of multiple dilation layers with dilation rates of 1, 3, 5, and 7, helping to capture rich context information under various receptive fields without increasing the network parameters. f_{U} is an upsampling layer. f_{Cat} is

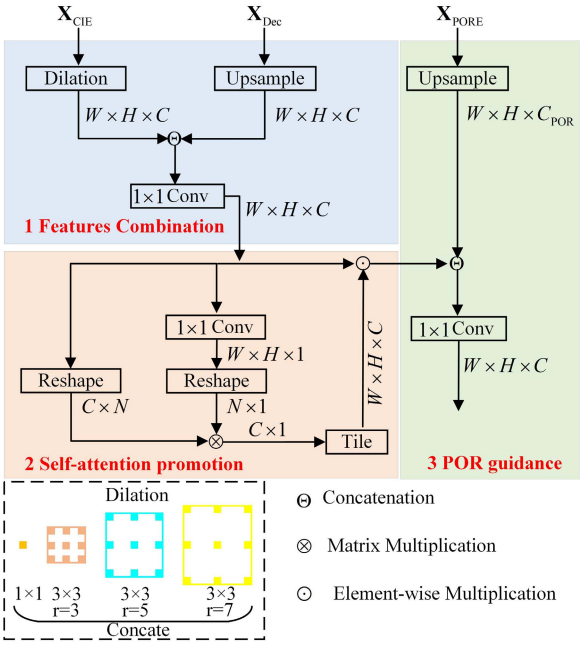


Fig. 6. Illustration for POGU, which consists of three phases, *i.e.*, features combination, self-attention promotion, and POR guidance. $\mathbf{X}_{CIE}$ and $\mathbf{X}_{PORE}$ are the encoded contrast and part-object relational semantics, respectively, and $\mathbf{X}_{Dec}$ is the decoded deep-level semantics. W, H, and C represent the width, height, and channel number of the corresponding feature maps, respectively. C_{POR} is the number of the corresponding capsule types. $N = W \times H$.

implemented by one concatenation and one 1×1 convolution. $\mathbf{W}_*$ represents the learned network parameters.

b) Self-attention promotion: Self-attention promotion is used to promote those informative features while suppressing less important ones by a channel-wise attention. As shown in Fig. 6, the process can be illustrated below.

Step 1: Channel attention. The combined feature maps $\mathbf{F}_{Com}$ is used to compute the channel attention map $\mathbf{f}^{Ch} \in R^{C\times 1}$, *i.e.,*

$$\mathbf{f}_{Ch} = f_{Res}\left(\mathbf{F}_{Com}\right) \otimes f_{Ch}\left(\mathbf{F}_{Com}; \mathbf{W}_{Ch}\right), \tag{2}$$

where f_{Res} represents the reshape operation. f_{Ch} is implemented by one 1×1 convolution and reshape. $\otimes$ means the matrix multiplication. $f_{Res}\left(\mathbf{F}_{Com}\right) \in R^{C\times N}$, $f_{Ch}\left(\mathbf{F}_{Com}; \mathbf{W}_{Ch}\right) \in R^{C\times 1}$, and $N = W \times H$.

Step 2: Features promotion. The promoted feature maps $\mathbf{F}_{Pro} \in R^{W\times H\times C}$ can be achieved by

$$\mathbf{F}_{Pro} = \mathbf{F}_{Com} \odot f_{tile}\left(\mathbf{f}_{Ch}\right), \tag{3}$$

where f_{tile} and $\odot$ are the operations of tensor expansion and element-wise multiplication, respectively.

c) Difference to self-attention in transformer networks: Transformer networks implement self-attention by learning query, key, and value components and then determining the self-attention by computing the similarity between query and key components. While our self-attention promotion simply computes the channel importance for further promoting those informative channels, which helps to promote the features themselves. Therefore, our self-attention promotion is simpler than that in transformer networks.

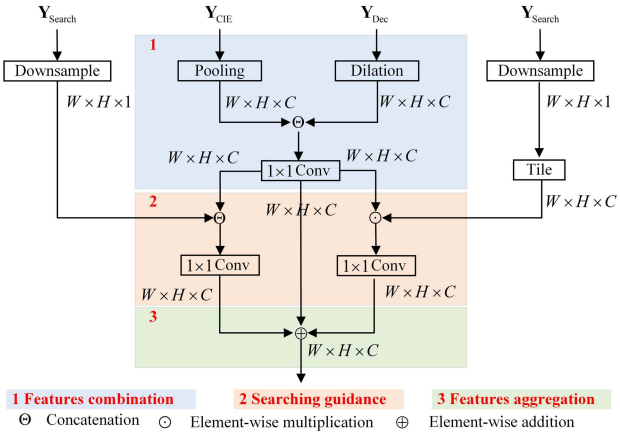


Fig. 7. Illustration for the SIG-induced CIE subnetwork, which consists of three phases, called features combination, searching guidance, and features aggregation. $\mathbf{Y}_{CIE}$ is the encoded features of the identification stage. $\mathbf{Y}_{Dec}$ and $\mathbf{Y}_{Search}$ are the decoded features and the coarse detection result of the search stage.

d) Difference to the attention mechanism in SINet [27]: The difference between the attention mechanisms in our model and SINet [27] can be illustrated as follows. SINet [27] adopted a Gaussian filter and a maximum function as the attention mechanism, which can filter out some noises while highlighting the detected regions, regardless of whether they are the camouflaged regions or not. Differently, we carry out a channel-wise attention via a series of operations to highlight those informative channels of feature maps, which helps to find the important channels that capture the camouflaged regions.

e) POR guidance: POR guidance is proposed to adopt the encoded part-object relational semantics, which captures the object wholeness prior, to guide the promoted features $\mathbf{F}_{Pro}$ for more accurate and complete camouflaged semantics.

$\mathbf{X}_{PORE}$ is first upsampled into $\mathbf{F}_{POR} \in R^{W\times H\times C_{POR}}$, where C_{POR} is the number of the corresponding capsule types. The current guided feature maps $\mathbf{F}_{Dec}$, which are also the feature maps of the current decoder layer, can be achieved by

$$\mathbf{F}_{Dec} = f_{Cat}\left(\mathbf{F}_{POR}, \mathbf{F}_{Pro}; \mathbf{W}_{Cat}\right). \tag{4}$$

$\mathbf{F}_{Dec}$ efficiently integrates the encoded contrast and part-object relational semantics. Such practice helps to decode more primitive camouflaged cues for prediction.

B. Identification Stage

The identification stage consists of a SIG-induced CIE subnetwork and a PORE subnetwork in the encoder, and a POGU module in the decoder. The PORE subnetwork and the POGU module are similar to those of the search stage.

1) SIG-Induced CIE Subnetwork: As shown in Fig. 7, the SIG-induced CIE subnetwork consists of three phases: features combination, searching guidance, and features aggregation. Suppose $\mathbf{Y}_{CIE}$ is the encoded features of the identification stage. $\mathbf{Y}_{Dec}$ and $\mathbf{Y}_{Search}$ are the decoded features and the coarse detection result of the search stage. The details of SIG will be presented in the following.

a) Features combination: Features combination is designed to combine the encoded contrast information $\mathbf{Y}_{CIE}$

of the identification stage and the decoded feature maps $\mathbf{Y}_{\text{Dec}}$ of the search stage. Doing so allows the decoded semantics $\mathbf{Y}_{\text{Dec}}$ of the search stage to improve the features encoding ability of the current encoder of the identification stage. To be specific, the combined features $\mathbf{G}_{\text{Com}} \in R^{W\times H\times C}$ can be computed by

$$\mathbf{G}_{\text{Com}} = f_{\text{Cat}}\left(f_{\text{pool}}\left(\mathbf{Y}_{\text{CIE}}\right), f_{\text{Dia}}\left(\mathbf{Y}_{\text{Dec}}; \mathbf{W}_{\text{Dia}}\right); \mathbf{W}_{\text{Cat}}\right), \quad (5)$$

where f_{pool} is the average pooling operation. f_{Dia} represents the dilation operation, as can be found in Fig. 6.

b) Searching guidance: Searching guidance is motivated to apply the search result $\mathbf{Y}_{\text{Search}}$ of the search stage to guide the combined feature maps $\mathbf{G}_{\text{Com}}$. On the one hand, $\mathbf{Y}_{\text{Search}}$ provides a coarse detection prior including the object location and rough details for $\mathbf{G}_{\text{Com}}$ to capture more accurate camouflaged cues. On the other hand, $\mathbf{G}_{\text{Com}}$ provides rich spatial details to refine $\mathbf{Y}_{\text{Search}}$. Concretely, $\mathbf{Y}_{\text{Search}}$ is downsampled and concatenated with $\mathbf{G}_{\text{Com}}$, *i.e.*,

$$\mathbf{G}_{\text{Cat}} = f_{\text{Cat}}\left(f_{\text{DW}}\left(\mathbf{Y}_{\text{Search}}; \mathbf{W}_{\text{DW}}\right), \mathbf{G}_{\text{Com}}; \mathbf{W}_{\text{Cat}}\right), \quad (6)$$

where f_{DW} represents the downsampling operation.

Simultaneously, the search prediction $\mathbf{Y}_{\text{Search}}$ of the search stage roughly predicts the camouflaged value of each position, which provides a pixel-level attention. Therefore, we utilize $\mathbf{Y}_{\text{Search}}$ to attend each channel of feature maps $\mathbf{G}_{\text{Com}}$. Specifically, $\mathbf{Y}_{\text{Search}}$ is first downsampled and then multiplied with $\mathbf{G}_{\text{Com}}$. The details can be formulated as

$$\mathbf{G}_{\text{Mul}} = f_{\text{tile}}\left(f_{\text{DW}}\left(\mathbf{Y}_{\text{Search}}; \mathbf{W}_{\text{DW}}\right)\right) \odot \mathbf{G}_{\text{Com}}. \quad (7)$$

$\mathbf{G}_{\text{Cat}}$ attends the features by the search prediction. $\mathbf{G}_{\text{Mul}}$ masks the features with the searched camouflaged value at each pixel position. Especially, $\mathbf{G}_{\text{Cat}}$ attends the features by jointly taking into account the search map and features from all the channels of $\mathbf{G}_{\text{Com}}$, while $\mathbf{G}_{\text{Mul}}$ attends the features by involving the search result and the current channel of $\mathbf{G}_{\text{Com}}$. Therefore, $\mathbf{G}_{\text{Cat}}$ tends to preserve accurate spatial details, whereas $\mathbf{G}_{\text{Mul}}$ prefers to produce an accurate camouflaged prediction for each pixel.

c) Features aggregation: To encode camouflaged cues with good accuracy and spatial details, we integrate these three types of information together, including $\mathbf{G}_{\text{Com}}$, $\mathbf{G}_{\text{Cat}}$, and $\mathbf{G}_{\text{Mul}}$, *i.e.*,

$$\mathbf{G}_{\text{CIE}} = \mathbf{G}_{\text{Com}} \oplus \mathbf{G}_{\text{Cat}} \oplus \mathbf{G}_{\text{Mul}}, \quad (8)$$

where $\oplus$ means the operation of element-wise addition.

$\mathbf{G}_{\text{CIE}}$ represents the feature maps of the current-layer SIG-induced CIE subnetwork. In Eq. (8), three enhanced features are integrated to enhance the features encoding ability of the identification stage, helping identify the camouflaged object in a complex scene.

C. Loss Function

We adopt the cross-entropy loss function (l_{ce}) used in [54] and the IoU boundary loss function (l_{iou}) to train the proposed camouflaged object detection network. Suppose P and Q are the predicted saliency map and corresponding ground truth. The cross-entropy loss function l_{ce} is written as follows

$$l_{ce}(P, Q) = -\sum_i \left[G_i \log(P_i) + (1 - Q_i)\log(1 - P_i)\right], \quad (9)$$

where i is the pixel index.

The IoU boundary loss function l_{iou} is defined as

$$l_{iou}(P, Q) = 1 - \frac{\sum_i P(i)\, Q(i)}{\sum_i \left[P(i) + Q(i) - P(i)\, Q(i)\right]}. \quad (10)$$

The joint loss function combines the cross-entropy loss function and the IoU Boundary loss function, *i.e.*,

$$l(P, Q) = l_{ce}(P, Q) + l_{iou}(P, Q). \quad (11)$$

IV. Experiment and Analysis

In this section, we will conduct various experiments to evaluate our proposed method.

A. Dataset

We evaluate the performance of our model on three benchmark datasets, details of which are described as follows.

CHAMELEON [55] is an unpublished dataset that has only 76 images collected from the Internet via the Google search engine using "camouflaged animal" as a keyword.

CPD1K [31] is the earliest dataset for camouflaged people detection, which contains 1000 images covering two scene types, namely woodland and snowfield. The test subset has 400 images.

COD10K [27], which is collected from multiple photography websites, contains 10000 images, including 5066 camouflaged images, 3000 background images, and 1934 non-camouflaged images. The test subset includes 2026 images.

CAMO [26] has 1250 images, which are divided into 1000 training images and 250 testing images.

B. Evaluation Metrics

We evaluate the performance of our model as well as other state-of-the-art methods using average weighted F-measure (F_β) [56], Mean Absolute Error (MAE) [56], S-measure (S_m) [57], and E-measure (E_m) [58].

A binary mask B is achieved by thresholding the saliency map P. Precision is defined as $Precision = |B \cap Q| / |B|$, and recall is defined as $Recall = |B \cap Q| / |Q|$, where Q is the corresponding ground truth. The PR curve is plotted under different thresholds. The F-measure is an overall performance indicator, which is computed by

$$F_\beta = \frac{\left(1 + \beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}. \quad (12)$$

As suggested in [56], $\beta^2 = 0.3$.

MAE is defined as

$$MAE = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} |P(i, j) - Q(i, j)|, \quad (13)$$

TABLE I

PERFORMANCE COMPARISONS FOR ABLATION ANALYSIS. (ED + POGU) = POCINET-SEARCH. (ED + POGU + SIG) = SIG = POCINET. THE BEST METHOD IS MARKED BY BOLD IN EACH SUBSECTION

Subsection	Model	COD10K [27]			
		$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$
(a) Sec. IV-D1	**ED**	0.4920	0.0746	0.6887	0.7319
	ED + POGU	0.5623	0.0598	0.7266	0.7902
	ED + POGU + SIG	**0.6143**	**0.0514**	**0.7515**	**0.8253**
(b) Sec. IV-D2	**ED + POGU**	**0.5623**	**0.0598**	**0.7266**	0.7902
	ED + POGU-C	0.5307	0.0652	0.6762	**0.7959**
(c) Sec. IV-D3	**POCINet-search**	**0.5623**	**0.0598**	**0.7266**	**0.7902**
	TSPOANet [30]	0.5119	0.0667	0.7162	0.7510
(d) Sec. IV-D4	**SIG**	**0.6143**	**0.0514**	**0.7515**	**0.8253**
	SIG-D	0.5884	0.0550	0.7448	0.8069
(e) Sec. IV-D5	**POCINet-PORE**	0.5380	0.0656	0.6987	0.7836
	POCINet-search	**0.5623**	**0.0598**	**0.7266**	**0.7902**

where w and h are the width and height of the image, respectively.

S-measure [57] computes the object-aware and region-aware structure similarities, denoted as S_o and S_r, between the prediction and the ground truth. The S-measure value S_m can be computed as

$$S_m = \alpha S_o + (1 - \alpha) S^r, \tag{14}$$

where α is set to 0.5 [57].

E-measure [58] (E_m) combines local pixel values with the image-level mean value to jointly evaluate the similarity between the prediction and the ground truth.

C. Implementation Details

The proposed model is implemented in Tensorflow [59]. To avoid over-fitting caused by training from scratch, the five stacked convolutional layers in the search stage are initialized by the Conv1_2, Conv2_2, Conv3_3, Conv4_3, and Conv5_3 of the pretrained VGG16 [52], respectively. The other weights are initialized randomly with a truncated normal ($\sigma = 0.01$), and the biases are initialized to 0. The Adam optimizer [60] is used to train our model with an initial learning rate of 10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The training datasets of CAMO [26] and COD10K [27] are chosen as the training dataset with horizontal flipping as the data augmentation technique. We adopt the joint loss function [30], including the cross-entropy loss function and the IoU loss function, to train our deep framework. The inference time for each image cropped into 352×352 is 0.1s, which is twice faster than SINet [27].

D. Ablation Analysis

In this subsection, we will carry out a series of experiments to investigate the role of each component in our framework.

1) Different Components: To better understand our framework, we train different components for comparisons, including the Encoder-Decoder (ED) model, ED + POGU, and ED + POGU + SIG. As shown in Table I(a), the proposed POGU module significantly improves the performance of ED, which benefits from the integration of the contrast information and the part-object relational knowledge. The proposed SIG module achieves a further performance improvement, which demonstrates the importance of the guidance from the search stage for the feature extraction of the identification stage. Besides, Fig. 8 displays the detection results of different components. Specifically, ED that relies on contrast information only can hardly identify the camouflaged object with low contrast between foreground and background (top two rows of Fig. 8) or just can capture discriminatively local regions (bottom two rows of Fig. 8). With the aid of the proposed POGU module that integrates the contrast information and the part-object relational knowledge, the camouflaged object can be localized distinguishably. Furthermore, the SIG-induced identification stage helps to segment out the camouflaged object wholly.

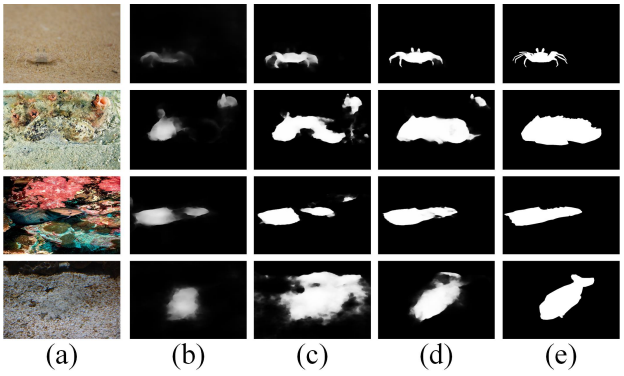


Fig. 8. Visual comparisons for different components. (a) Image; (b) ED; (c) ED + POGU; (d) ED + POGU + SIG; (e) GT.

2) POGU: To investigate the effectiveness of the proposed POGU module, we compare (ED + POGU) and a modified version, *i.e.*, (ED + POGU-C), which is implemented by directly concatenating the contrast information and the part-object relational cues. Quantitatively in Table I(b), our POGU achieves better performance with respect to F_β, MAE, and S_m, compared to POGU-C. Visually in the top two rows of Fig. 9, our POGU achieves better object wholeness than POGU-C does. As illustrated in the bottom two rows of Fig. 9, our POGU achieves sufficient object details while POGU-C produces some holes. This efficiently verifies the superiority of the intelligent integration between the contrast information and the part-object relational cues in our POGU.

3) The Part-Object Relationship Decoded Strategy: To investigate the superiority of the part-object relationship decoded strategy adopted in our model, we compare the search stage of POCINet (POCINet-search) with TSPOANet [30]. For a fair comparison, we re-train TSPOANet for camouflaged object detection. As shown by Table I(c), POCINet-search achieves a significant performance gain over TSPOANet. Besides, as shown in the top two rows of Fig. 10, POCINet-search gets more clear and complete object shapes than TSPOANet does. As illustrated in the bottom two rows of Fig. 10, TSPOANet misses some object details and thereby produces some holes, which can be addressed by POCINet-search. This benefits from the proposed POGU enabling the integration between the contrast information and the part-object relational cues, which helps to grab more object details and more accurate part-object relationships in a complex scene.

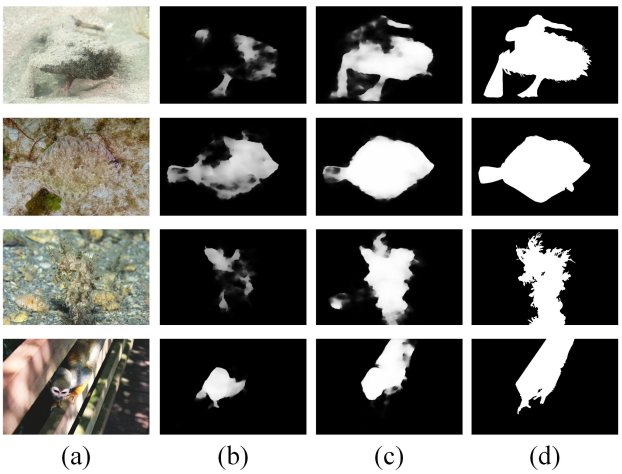


Fig. 9. Visual comparisons for POGU. (a) Image; (b) ED + POGU-C; (c) ED + POGU; (d) GT.

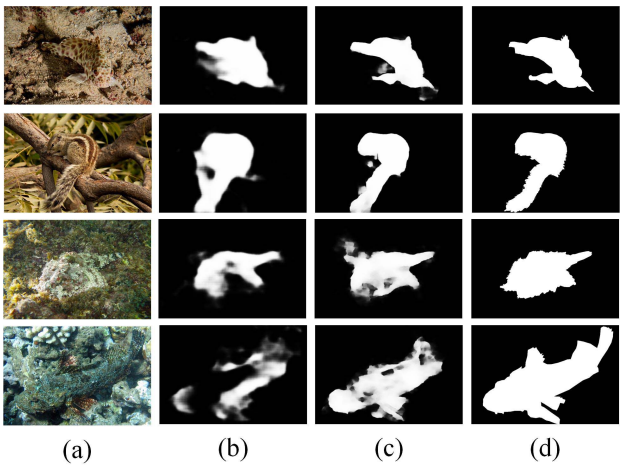


Fig. 10. Visual comparisons for the part-object relationship decoded strategies. (a) Image; (b) TSPOANet [30]; (c) POCINet-search; (d) GT.

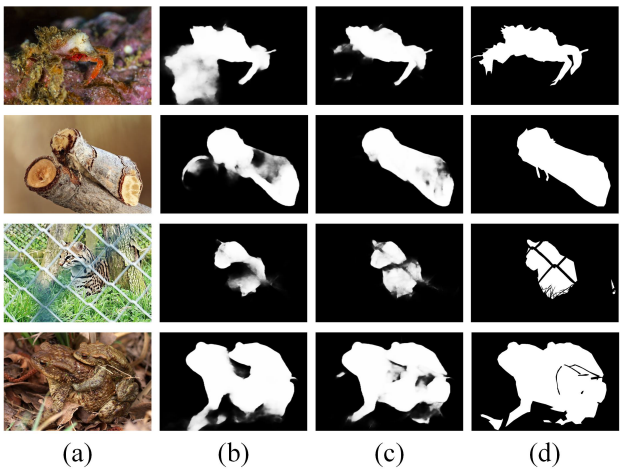


Fig. 11. Visual comparisons for SIG. (a) Image; (b) SIG-D; (c) SIG; (d) GT.

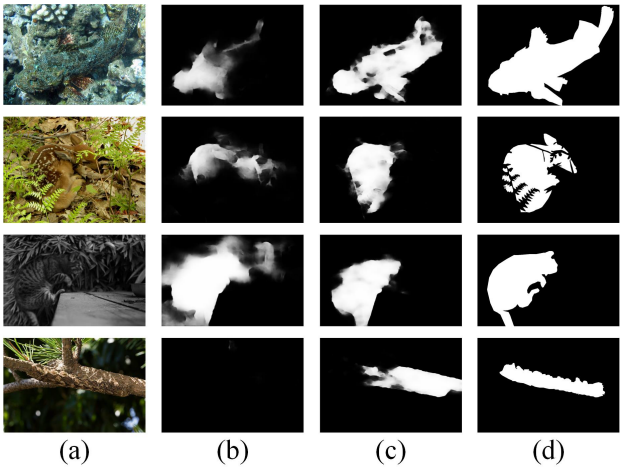


Fig. 12. Visual comparisons for the role of PORE in the identification stage. (a) Image; (b) POCINet-PORE; (c) POCINet-search; (d) GT.

4) SIG: To investigate the effectiveness of our connections between search and identification stages, we compare our entire model with a modified version, called SIG-D, which is implemented by directly concatenating the search result and decoded features of the search stage, and the encoded features of the identification stage. As shown in Table I(d), our SIG outperforms SIG-D by a clear margin. As illustrated in Fig. 11, SIG suppresses confusing background noises (the top two rows of Fig. 11), and captures more complete object shapes (the bottom two rows of Fig. 11), compared to SIG-D. This improvement arises from the primitive integration between contrast and part-object relational cues in the SIG module.

5) Role of PORE in the Identification Stage: To investigate the role of the PORE subnetwork in the identification stage, we train a modified version, in which the identification stage directly shares the PORE subnetwork of the search stage, called POCINet-PORE. As shown in Table I(e), POCINet-PORE is inferior to POCINet-search. This might be because the searched part-object relationships just coarsely locate the target object with some noisy reasonings, which will degrade the performance when applied to the identification stage. As well, as illustrated in Fig. 12, POCINet-PORE misses some object parts (top two rows of Fig. 12), introduces some background noises (third row of Fig. 12), or even cannot identify the camouflaged object (bottom row of Fig. 12), compared to POCINet-search, which is due to the over-search. Therefore, an individual PORE subnetwork is essential in the identification stage for segmenting out the camouflaged object, as can be verified by the performance improvement of the entire model.

6) Robustness Compared With SINet [27]: To explore the robustness of our model with different initial parameters and randomnesses of the training, we re-train our model and SINet [27] for another four times. Together with the final results of our model and SINet [27] listed in this paper, we compute the standard deviations of different metrics for different datasets. Table II lists the standard deviations. It can be seen that our standard deviations of different metrics are mostly smaller than that of SINet [27] (except E_m for CAMO [26]). This demonstrates that our model achieves more consistent performance with different initial randomized parameters when training than SINet [27] does, showing better robustness than SINet [27].

TABLE II

STANDARD DEVIATIONS OF DIFFERENT METRICS FOR OUR METHOD AND SINET [27]. SMALLER STANDARD DEVIATION IS MARKED BY BOLD

	CHAMELEON [61]				CPD1K [31]				CAMO [26]				COD10K [27]			
	F_β	MAE	S_m	E_m	F_β	MAE	S_m	E_m	F_β	MAE	S_m	E_m	F_β	MAE	S_m	E_m
OURS	**0.0042**	**0.0012**	**0.0034**	**0.0038**	**0.0090**	**1.7889e-04**	**5.8052e-04**	**0.0098**	**0.0084**	**0.0013**	**0.0067**	0.0108	**0.0029**	**4.0373e-04**	**8.1854e-04**	**0.0042**
SINet [27]	0.0133	0.0041	0.0099	0.0072	0.0228	0.0010	0.0083	0.0222	0.0087	0.0045	0.0109	**0.0060**	0.0113	0.0022	0.0069	0.0100

TABLE III

F_β, MAE, S_m, AND E_m VALUES OF DIFFERENT METHODS. TOP TWO METHODS ARE MARKED BY RED AND BLUE, RESPECTIVELY

		CHAMELEON [61] (76 images)				CPD1K [31] (400 images)				CAMO [26] (250 images)				COD10K [27] (2026 images)			
	Backbone	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$
OURS	VGG	0.7976	0.0417	0.8663	0.9080	0.6717	0.0067	0.8640	0.8682	0.6623	0.1101	0.7017	0.7768	0.6143	0.0510	0.7515	0.8253
TSPOANet [30]	VGG	0.7038	0.0586	0.8276	0.8574	0.4830	0.0097	0.8195	0.7270	0.6410	0.1179	0.7088	0.7985	0.5119	0.0667	0.7162	0.7510
PFANet [62]	VGG	0.5168	0.1443	0.6792	0.7322	0.2185	0.0425	0.6580	0.4545	0.5511	0.1721	0.6589	0.7350	0.3848	0.1282	0.6357	0.6186
SINet [27]	ResNet	0.7755	0.0438	0.8685	0.8988	0.5066	0.0103	0.8489	0.7247	0.7086	0.0997	0.7515	0.8345	0.5931	0.0511	0.7710	0.7971
PoolNet [63]	ResNet	0.6486	0.0811	0.7764	0.8243	0.2633	0.0256	0.6316	0.5360	0.6275	0.1286	0.7025	0.7895	0.4788	0.0744	0.7051	0.7081
HTC [64]	ResNet	0.3271	0.1292	0.5166	0.4898	0.4725	0.0192	0.7026	0.6657	0.3378	0.1722	0.4764	0.4423	0.2984	0.0878	0.5478	0.5209
EGNet [65]	ResNet	0.7373	0.0502	0.8483	0.8785	0.4051	0.0188	0.5881	0.6924	0.6802	0.1036	0.7319	0.8267	0.5480	0.0561	0.7369	0.7772
BASNet [29]	ResNet	0.5458	0.1179	0.6874	0.7419	0.4841	0.0178	0.7563	0.7507	0.5252	0.1590	0.6182	0.7191	0.4213	0.1054	0.6343	0.6756
PiCANet [11]	ResNet	0.6676	0.0847	0.7686	0.8363	0.2967	0.0240	0.7540	0.5430	0.5382	0.1560	0.6087	0.7534	0.4225	0.0899	0.6491	0.6782
FPN [66]	ResNet	0.6758	0.0750	0.7935	0.8351	0.3917	0.0126	0.7861	0.6217	0.6416	0.1310	0.6838	0.7907	0.4837	0.0747	0.6972	0.7109
MaskRCNN [67]	ResNet	0.6119	0.0992	0.6430	0.7802	0.3214	0.0371	0.6145	0.6723	0.5207	0.1511	0.5738	0.7164	0.4702	0.0805	0.6132	0.7504
MSRCNN [68]	ResNet	0.5290	0.0914	0.6372	0.6881	0.5784	0.0099	0.7425	0.7916	0.5442	0.1327	0.6171	0.6704	0.4860	0.0734	0.6413	0.7077
PSPNet [28]	ResNet	0.6501	0.0850	0.7734	0.8139	0.3173	0.0167	0.7650	0.5524	0.6053	0.1390	0.6630	0.7779	0.4507	0.0801	0.6778	0.6876

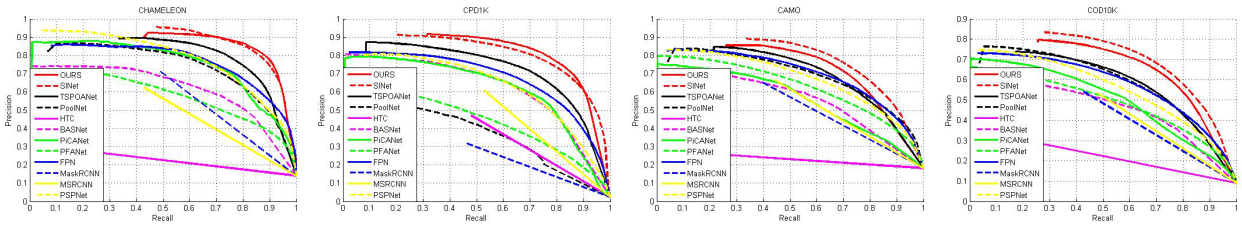


Fig. 13. PR and F-measure curves of different methods on camouflaged object detection datasets.

E. Comparison With the State-of-the-Art Methods

In this subsection, to verify the superiority of our model, we compare our method with one deep camouflaged object detector, *i.e.*, SINet [27]. Besides, due to the lack of deep camouflaged object detection networks, ten deep object detectors are taken into account for comparisons, including TSPOANet [30], PoolNet [63], HTC [64], BASNet [29], PiCANet [11], PFANet [62], FPN [66], MaskRCNN [67], MSRCNN [68], and PSPNet [28], which are re-trained for camouflaged object detection.

1) Quantitative Comparison: Table III illustrates the quantitative comparisons. For CHAMELEON [61], our method performs best in terms of F_β, MAE, and E_m, and is slightly inferior to SINet [27] with respect to S_m. For CPD1K [31] that only contains camouflaged persons, we achieve the best performance. Especially, we achieve 16.51 and 14.35 points higher than the best compared method, *i.e.*, SINet [27], with respect to F_β and E_m, respectively, which indicates our model can especially address the camouflaged people detection. For COD10K [27], we beat the other approaches with respect to F_β, MAE, and E_m, but are slightly inferior to SINet [27] in terms of S_m. Obviously, we achieve consistently superior performance on these three datasets. For CAMO [26], we are inferior to SINet [27] but superior to the other methods. Fig. 13 plots the PR curves of different methods. Specifically, on CHAMELEON [61] and CPD1K [31], we achieve the best PR performance. On COD10K [27] and CAMO [26], our method is inferior to SINet [27] but significantly better than the other methods. However, it is worth noting that our proposed method uses a primitive backbone VGG [52] while other competitors, including SINet [27], take advantage of a ResNet [69] backbone, which is well-known for its better performance. The reason for using VGG [52] is to make our network as thin and lightweight as possible.

Fig. 14 illustrates the detection results of different methods on images with various distortions, including scaling, slender objects, and various shapes. To be specific, for those objects with different sizes, the compared methods usually miss some object parts of the large object, and hardly identify the small object. While our method can segment out the large objects with good wholeness and the small object clearly under the complex scenes. For the slender objects, our approach can identify them and segment them out with clear object boundaries, while the others fail to recognize these objects owing to the high similarity between foreground and background. For those objects with various shapes, the compared methods mostly cannot label the whole object boundaries and thereby fail at segmenting them out, while our method can well detect the whole object shape with clear boundaries.

2) Deeper Insight Into CAMO [26]: As the dataset CAMO [26] contains real-life (CAMO-RE) images (*e.g.*, Fig. 14) and synthetic (CAMO-SY) images (e.g., Fig. 15), we divide the whole dataset into two subsets: CAMO-RE and CAMO-SY, respectively. We believe, the comparisons on two subsets, as well as the whole dataset, are fair and will be more insightful. As shown in Table IV, we achieve the best performance on CAMO-RE while obtaining the second-best performance On CAMO-SY, which is unfortunately inferior to SINet [27].

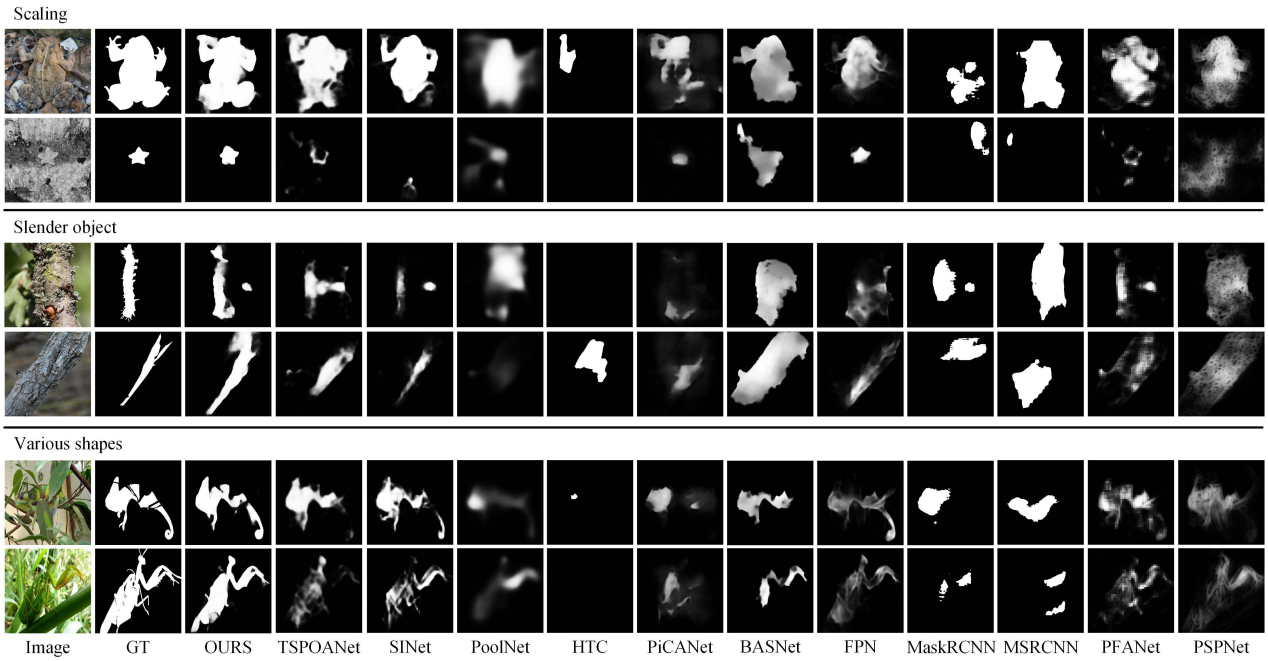


Fig. 14. Detection results of different methods.

TABLE IV

F_β, MAE, S_m, AND E_m VALUES OF DIFFERENT METHODS ON CAMO [26]. TOP TWO METHODS ARE MARKED BY RED AND BLUE, RESPECTIVELY

	CAMO-RE		CAMO-SY	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
OURS	0.7476	0.0681	0.5654	0.1577
SINet [27]	0.7413	0.0682	0.6715	0.1354
TSPOANet [30]	0.6792	0.0866	0.5975	0.1534
PoolNet [63]	0.6627	0.0950	0.5876	0.1669
HTC [64]	0.3637	0.1475	0.3084	0.2002
PiCANet [11]	0.6040	0.1220	0.4633	0.1946
PFANet [62]	0.5743	0.1512	0.5247	0.1959
FPN [66]	0.6654	0.0984	0.6145	0.1681
MaskRCNN [67]	0.5871	0.1203	0.4453	0.1682
MSRCNN [68]	0.6118	0.0991	0.4673	0.1708
PSPNet [28]	0.6361	0.1099	0.5702	0.1721

a) Why SINet [27] is superior to our method on CAMO-SY?: The reason why SINet is superior to our model can be interpreted as follows. SINet relies on the deep network to extract the robust features of the input image to carry out the task of camouflaged object detection. While we attempt to detect the camouflaged object from the part-object relational perspective, which identifies the whole camouflaged object by finding the relevant object parts. This manner prefers to work for real-life scenes because part-object relationships appear truly on the real-life scenes. In contrast, the part-object relationships (at the feature layers) in those man-made synthetic images on CAMO-SY have been strained due to unnatural pixels. Consequently, our model based on the part-object relational view is inferior on CAMO-SY, compared to SINet. However, SINet is inferior to our model on real-life scenes because SINet has a limited ability for feature extraction on low-contrast real-life scenes, which can be addressed easily by our model because that our part-object relationships extraction is not weakened on various camouflaged scenes.

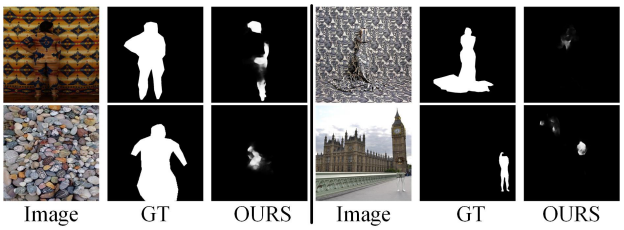


Fig. 15. Confusing synthetic images.

TABLE V

DISTRIBUTIONS OF REAL-LIFE AND SYNTHETIC IMAGES ON CAMO [26]

	Training dataset		Testing dataset	
	CAMO-RE + COD10K	CAMO-SY	CAMO-RE	CAMO-SY
Number of images	867 + 3640 = 4507	133	133	117
Ratio	33.9	1	1.1	1

As illustrated in Fig. 15, some synthetic scenes confuse the proposed camouflaged object detector, resulting in a failure at recognizing the camouflaged object from the surroundings. This observation reveals that synthetic images might be less useful for training the network with a CapsNet structure. Again, we want to emphasize that our entire framework is built upon a primitive VGG [52] backbone.

To have a better understanding of the reason of why the proposed model performs unsatisfactorily on CAMO-SY, we take a study on the data distribution of the real-life images against synthetic images during training and testing. As illustrated in Table V, when training the proposed camouflaged object framework, the number of real-life images and synthetic images are 4507 and 133, respectively, resulting in a ratio of 33.9 : 1. However, when testing for the dataset CAMO [26], the real-life images and synthetic images become 133 and 117, respectively, resulting in a ratio of 1.1 : 1. Obviously, the marginal number of synthetic images, whose statistical distribution is not in accordance with that of real-life images,

TABLE VI

AVERAGE OF F_β, MAE, S_m, AND E_m VALUES ON FOUR DATASETS, INCLUDING CHAMELEON [61], CPD1K [31], COD10K [27], AND CAMO [26], OF DIFFERENT METHODS. THE TOP TWO METHODS ARE MARKED BY RED AND BLUE, RESPECTIVELY

Method	Metric			
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$
OURS	0.6321	0.0496	0.7665	0.8294
SINet [27]	0.5961	0.0497	0.7832	0.7928
TSPOANet [30]	0.5247	0.0628	0.7336	0.7548
PoolNet [63]	0.4657	0.0724	0.6961	0.6937
HTC [64]	0.3281	0.0866	0.5630	0.5339
PiCANet [11]	0.4215	0.0862	0.6640	0.6697
PFANet [62]	0.3794	0.1202	0.6423	0.6085
FPN [66]	0.4900	0.0708	0.7116	0.7086
MaskRCNN [67]	0.4571	0.0811	0.6106	0.7368
MSRCNN [68]	0.5059	0.0701	0.6537	0.7160
PSPNet [28]	0.4509	0.0764	0.6918	0.6796

in the training subset cannot train our PORE subnetwork that is a CapsNet structure to explore primitive part-object relational cues, which further wearkens the proposed camouflaged object detector to perform that promising on synthetic images. This will be taken into account in our future work.

3) Overall Comparison on CHAMELEON [61], CPD1K [31], COD10K [27], and CAMO [26]: To better exhibit the performance of the proposed method, we compute an overall performance on four datasets, including CHAMELEON [61], CPD1K [31], COD10K [27], and CAMO [26]. Specifically, we propose a simple yet effective indicator to calculate overall values of F_β, MAE, S_m, and E_m. The indicator can be represented as

$$\phi_j = \sum_{i=1}^{4} (\alpha_i \varphi_{ij}) \Big/ \sum_{i=1}^{4} \alpha_i, \tag{15}$$

where $i\,(i = 1, 2, 3, 4)$ and $j\,(j = 1, 2, 3, 4)$ represent different datasets (CHAMELEON [61], CPD1K [31], COD10K [27], and CAMO [26]) and different metric values (F_β, MAE, S_m, and E_m). φ_{ij} is the jth metric value for the ith dataset. α_i is the number of images on the ith dataset. ϕ_j is the jth overall metric value. Using Eq. (15), Table VI lists the overall performance of different metrics. It can be found in Table VI,the proposed method achieves the best performance in terms of F_β, MAE, and E_m, and competitive performance in terms of S_m that is slightly inferior to SINet [27]. Taking into account Table VI, III, and IV, our method prefers to solve the problem of camouflaged object detection on real-life scenes. Again, it is noticed that such a promising performance is achieved based on a primitive VGG [52] backbone, showing that the integration of the contrast information and part-object relational knowledge indeed helps detect camouflaged objects.

F. Failure Cases

Although the proposed model has achieved good performance on various cases, there still exist some challenging cases to be solved. Fig. 16 illustrates some confusing detection results of our model on complex scenes. More precisely, those objects in the top two rows of Fig. 16 cannot be segmented out wholly, which is because that these large objects have unclear part-object hierarchies within their identical inner regions.

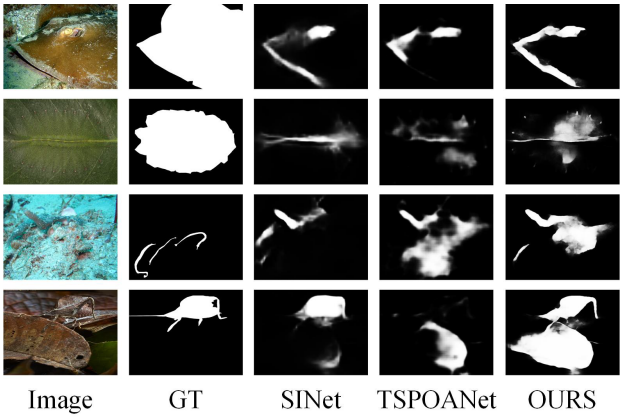


Fig. 16. Some failure cases.

Besides, those images in the bottom two rows of Fig. 16 introduce noises in their camouflaged maps, which owes to the confusing backgrounds that conceal the camouflaged patterns. Also, these scenes are challenging for the contrast based method (*i.e.*, SINet [27]) and the part-object relational method (*i.e.*, TSPOANet [30]). In the future, we will dig into more robust part-object relationships exploration mechanisms to improve our method for various camouflaged patterns via taking into account robust deep learning [70].

V. CONCLUSION

In this paper, we have proposed a POCINet covering the search stage and the identification stage for camouflaged object detection by engaging the contrast information and the part-object relational knowledge for decoding. Besides, a SIG module is designed to biologically connect two stages for location and segmentation of the camouflaged object in complex scenes. Extensive experiments have verified the superiority of the proposed camouflaged object detection network. One possible future work might be the enhancement of our part-object relationships exploration for various camouflaged scenes by incorporating robust deep learning mechanisms. Another possible future work might be the performance improvement of our model by using more powerful backbone networks, *e.g.*, ResNet [69] and DenseNet [71].

REFERENCES

[1] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1359–1370, Jul. 2015.

[2] G. Chen *et al.*, "Neuroaed: Towards efficient abnormal event detection in visual surveillance with neuromorphic vision sensor," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 923–936, 2020.

[3] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proenca, "The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1696–1708, 2021.

[4] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 988–998, Jun. 2014.

[5] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 717–732, Mar. 2018.

[6] N. Y. Almudhahka, M. S. Nixon, and J. S. Hare, "Semantic face signatures: Recognizing and retrieving faces by verbal descriptions," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 706–716, Mar. 2018.

[7] L. Zheng, Y. Lei, G. Qiu, and J. Huang, "Near-duplicate image detection in a visually salient Riemannian space," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1578–1593, Oct. 2012.

[8] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[9] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6609–6617.

[10] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[11] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[12] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 478–487.

[13] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE Trans. Image Process.*, vol. 29, pp. 360–374, 2020.

[14] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[15] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient object detection via two-stage graphs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1023–1037, Apr. 2018.

[16] M.-M. Cheng *et al.*, "BING: Binarized normed gradients for objectness estimation at 300 fps," *Comput. Vis. Media*, vol. 5, no. 1, pp. 3–20, 2019.

[17] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[18] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[19] D.-P. Fan *et al.*, "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 263–273.

[20] M. Yasin, O. Sinanoglu, and J. Rajendran, "Testing the trustworthiness of IC testing: An oracle-less attack on IC camouflaging," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2668–2682, Nov. 2017.

[21] K. Shamsi, T. Meade, M. Li, D. Z. Pan, and Y. Jin, "On the approximation resiliency of logic locking and IC camouflaging schemes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 2, pp. 347–359, Feb. 2019.

[22] S. Ge, X. Jin, Q. Ye, Z. Luo, and Q. Li, "Image editing by object-aware optimal boundary searching and mixed-domain composition," *Comput. Vis. Media*, vol. 4, no. 1, pp. 71–82, Mar. 2018.

[23] H.-K. Chu, W.-H. Hsu, N. J. Mitra, D. Cohen-Or, T.-T. Wong, and T.-Y. Lee, "Camouflage images," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 1–51, 2010.

[24] S. K. Singh, C. A. Dhawale, and S. Misra, "Survey of object detection methods in camouflaged image," *Proc. IERI*, vol. 4, pp. 351–357, Jan. 2013.

[25] Y. Zheng, X. Zhang, F. Wang, T. Cao, M. Sun, and X. Wang, "Detection of people with camouflage pattern via dense deconvolution network," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 29–33, Jan. 2019.

[26] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Understand.*, vol. 184, pp. 45–56, Jul. 2019.

[27] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2777–2787.

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[29] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[30] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1232–1241.

[31] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[32] G. H. Thayer and A. H. Thayer, *Concealing-Coloration in the Animal Kingdom: An Exposition of the Laws of Disguise Through Color and Pattern Being a Summary of Abbott H. Thayer's Discoveries*. New York, NY, USA: Macmillan Co, 1909.

[33] E. B. Poulton, "Adaptive coloration in animals," *Nature*, vol. 146, no. 3692, pp. 144–145, Aug. 1940.

[34] Y. Pan, Y. Chen, Q. Fu, P. Zhang, and X. Xu, "Study on the camouflaged target detection method based on 3D convexity," *Mod. Appl. Sci.*, vol. 5, no. 4, p. 152, Aug. 2011.

[35] Z. Liu, K. Huang, and T. Tan, "Foreground object detection using top-down information based on EM framework," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4204–4217, Sep. 2012.

[36] M. Stevens and S. Merilaita, "Animal camouflage: Current issues and new perspectives," *Philos. Trans. Royal Soc. B, Biol. Sci.*, vol. 364, no. 1516, pp. 423–427, 2009.

[37] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychol. Rev.*, vol. 94, no. 2, p. 115, 1987.

[38] J. Krivic and F. Solina, "Part-level object recognition using superquadrics," *Comput. Vis. Image Understand.*, vol. 95, no. 1, pp. 105–126, Jul. 2004.

[39] A. Jaklic, "Construction of cad models from range images," Ph.D. dissertation, Dept. Comput. Inf. Sci., Univ. Ljubljana, Kongresni Trg, Ljubljana, Slovenia, 1997.

[40] A. P. Pentland, "Automatic extraction of deformable part models," *Int. J. Comput. Vis.*, vol. 4, no. 2, pp. 107–126, Mar. 1990.

[41] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2241–2248.

[42] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 437–446.

[43] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. Int. Conf. Artif. Neural Netw.*, 2011, pp. 44–51.

[44] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.

[45] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 3856–3866.

[46] B. Cheng *et al.*, "SPGNet: Semantic prediction guidance for scene parsing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5218–5228.

[47] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 483–499.

[48] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8280–8289.

[49] W. Shen, B. Wang, Y. Jiang, Y. Wang, and A. Yuille, "Multi-stage Multi-recursive-input fully convolutional networks for neuronal boundary detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2391–2400.

[50] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

[51] Z. Deng *et al.*, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

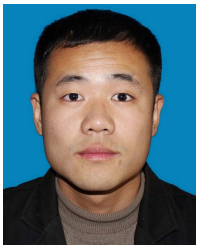
[53] S. Sabour, N. Frosst, and G. Hinton, "Matrix capsules with em routing," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[54] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

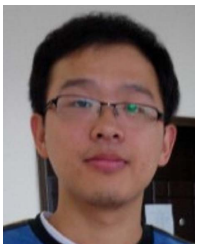
[55] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, "Animal camouflage analysis: Chameleon database," *Unpublished Manuscript*, vol. 2, no. 6, p. 7, 2018.

[56] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[57] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.
[58] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*.
[59] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[61] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.
[62] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.
[63] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.
[64] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4974–4983.
[65] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.
[66] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[67] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[68] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.
[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[70] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.
[71] G. Huang, Z. Liu, L. Van D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.



Yi Liu received the B.S. degree from the Nanjing Institute of Technology, China, in 2012, the M.S. degree from Dalian University, China, in 2015, and the Ph.D. degree from Xidian University, China, in 2019. From 2018 to 2019, he was a Visiting Scholar at Lancaster University. He is currently a Lecturer at Changzhou University. His research interests include computer vision and deep learning.



Dingwen Zhang (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2018. From 2015 to 2017, he was a Visiting Scholar at The Robotics Institute, Carnegie Mellon University. He is currently a Professor with the School of Automation, Northwestern Polytechnical University. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, temporal action localization, and weakly supervised learning.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a Professor with the Department of Automatic Control, Xidian University. His current research interests include image processing and pattern recognition.

Jungong Han is currently a Chair Professor and the Director of the Research of Computer Science, Aberystwyth University, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning.