Segment, Magnify and Reiterate: Detecting Camouflaged Objects the Hard Way

Qi Jia, Shuilian Yao, Yu Liu*, Xin Fan, Risheng Liu, Zhongxuan Luo
International School of Information Science & Engineering, Dalian University of Technology
{jiaqi,liuyu8824,xin.fan,rsliu,zxluo}@dlut.edu.cn, Shuilian_Yao@mail.dlut.edu.cn

Abstract

*It is challenging to accurately detect camouflaged objects from their highly similar surroundings. Existing methods mainly leverage a single-stage detection fashion, while neglecting small objects with low-resolution fine edges requires more operations than the larger ones. To tackle camouflaged object detection (COD), we are inspired by humans attention coupled with the coarse-to-fine detection strategy, and thereby propose an iterative refinement framework, coined **SegMaR**, which integrates **Seg**ment, **Ma**gnify and **R**eiterate in a multi-stage detection fashion. Specifically, we design a new discriminative mask which makes the model attend on the fixation and edge regions. In addition, we leverage an attention-based sampler to magnify the object region progressively with no need of enlarging the image size. Extensive experiments show our SegMaR achieves remarkable and consistent improvements over other state-of-the-art methods. Especially, we surpass two competitive methods 7.4% and 20.0% respectively in average over standard evaluation metrics on small camouflaged objects. Additional studies provide more promising insights into SegMaR, including its effectiveness on the discriminative mask and its generalization to other network architectures. Code is available at https://github.com/dlut-dimt/SegMaR.*

1. Introduction

Camouflaged object detection (COD) is a task which aims to identify any object hidden in the background [8, 22, 29]. It has been commonly useful for many applications in different fields [9, 38], including agriculture (*e.g.* locust detection to prevent invasion), art (*e.g.* photo-realistic blending and recreational art) and medical diagnosis (*e.g.* polyp segmentation). Biological and psychological studies have shown that various camouflage strategies can easily deceive the human's visual perceptual system [38], since the camouflaged objects always have similar visual features as the background surroundings. The major difficulty in COD is how to accurately distinguish the subtle differences between the target object and the background in the image.

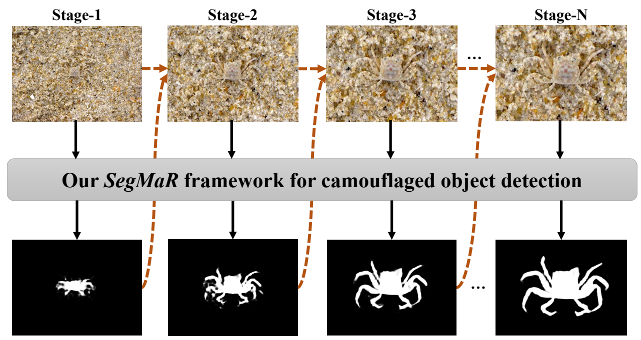

Figure 1. Illustration of our SegMaR (Segment, Magnify and Reiterate) framework for camouflaged object detection. Multiple stages are performed iteratively in the framework. Each stage involves two main steps: segment the camouflaged object (the solid line) and magnify the camouflaged object (the dotted line).

Different from traditional methods [5, 29, 51], a number of recent works [4, 7, 8, 20, 28, 45], by making use of sophisticated deep learning techniques [3, 43, 53], have achieved new state-of-the-art performance on all the COD benchmarks. Despite the quantitative performance by the latest methods looks promising (*e.g.* 0.80 of S_α on COD10K test set [8]), several difficulties in COD are still remaining unsolved. Particularly, when one certain camouflaged object accounts for a very small proportion of the whole image, it becomes more difficult to detect accurate edges around the object. For instance the crab in the first column in Fig. 1, its size is much smaller than the beach in the background. Unfortunately, existing COD methods fail to detect small camouflaged objects accurately.

Their detection and segmentation results lead to high deviation on low-resolution and small objects. One main reason is these methods employ a single-stage detection fashion, but many camouflaged objects are hardly detectable at the first time. In fact, when humans cannot watch any target object in the scene clearly, they will consciously move closer to the target till its resolution is large enough for visual recognition. We expect every person in front of the

*Corresponding author

screen is using such a manner to observe the small crab in Fig. 1. Motivated by this human behavior, our work aims to address the research question: *How to leverage more stages for gradually discovering more accurate camouflaged objects?*

To this end, we propose a new iterative refinement framework, coined *SegMaR*, which integrates *Segment*, *Magnify* and *Reiterate* via a multi-stage detection fashion, please see Fig. 2. First of all, our approach builds a new camouflaged segmentation network to generate an initial mask prediction. Next, an object magnification step takes as input both the original image and the mask prediction, and leverages an attention-based sampler to enlarge the camouflaged object adaptively. It can be observed that the image size is kept while the camouflaged object accounts for a larger proportion in the image. Moreover, we run iterative refinement by passing the image with magnified object back into the same network and fine-tuning the network parameters. After more refinement stages, SegMaR enables to refine and enrich the detected details, especially for small objects.

Importantly, SegMaR is an unified and general framework which shall be applicable to various camouflaged segmentation networks. Considering the significance of object localization and edge extraction, we advocate several special designs on the segmentation network for improving the COD performance further.

In particular, we introduce a distraction module to disentangle foreground and background features in order to capture more accurate edges.

Besides, we present a new and non-binary ground truth called *discriminative mask*, which combines the fixation and edge annotations together. Beyond the original ground truth based on binary mask, our discriminative mask makes the network attend more on the most significant textures and edges associated with the camouflaged objects.

The contributions in this work are three-fold:

- **Framework contribution:** we propose SegMaR, which is the first to leverage an iterative refinement framework for camouflaged object detection. This work raises awareness of the importance of accomplishing COD in a multi-stage detection fashion.
- **Network contribution:** we implement an effective camouflaged segmentation network which introduces a distraction module to disentangle better object feature. In addition, we present a new discriminative mask to make the network attend on the most significant object regions.
- **Empirical contribution:** Our SegMaR achieves new state-of-the-art performance on three COD benchmarks, especially for small camouflaged objects. Besides, previous COD networks are easy to be applicable to SegMaR and witness remarkable accuracy boosts.

2. Related Work

Camouflaged Object Detection. Camouflaged or concealed objects [7, 18, 29, 51] are hardly detectable due to their subtle differences from the background surroundings. To overcome this difficulty, a increasing number of recent works [8, 20, 25, 28, 36] are devoted to adopting sophisticated SOD techniques [3, 42, 43, 53] for solving COD. For instance, SINet [8] was built on top of cascaded partial decoder [43] which has been widely used for SOD. The work in [25] introduced the reverse attention [3] in order to capture more spatial details. Besides, some other works [24, 49] focus on how to extract more accurate edges around the camouflaged objects. Zhai *et al*. [49] built an edge-Constricted Graph Reasoning module to guide feature representation learning of camouflaged objects. However, these existing methods are not robust to some more challenging yet practical cases, especially when the camouflaged objects are very small. *Different from the single-stage framework used in previous works, our SegMaR refines and enriches camouflaged detection results iteratively in a multi-stage framework.*

Iterative Refinement. This is a common and effective learning process for a variety of vision-oriented applications like object detection [1, 12], semantic segmentation [34, 50] and object localization [32, 35]. On the one hand, some studies [2,11,23,34] perform the refinement steps iteratively from shallow to deep convolutional layers within one single neural network. For example, the work in [34] addressed semantic segmentation with a refinement module and stacked such modules successively into a top-down refinement process. Likewise, Lin *et al*. [23] presented a multi-path refinement network which effectively combines high-level semantics and low-level features to generate high-resolution segmentation maps. On the other hand, several research works [47,50] re-train the same network iteratively by passing the result of the last training iteration into the next iteration. Representatively, CANet [50] proposed an iterative optimization module to refine predicted results for few-shot semantic segmentation.

Despite the significant improvements by iterative refinement, it has not been researched for solving COD yet. Besides, our SegMaR framework aims to magnify camouflaged objects gradually until capturing more accurate results.

Object Magnification. To increase the resolution of the target objects, some tasks [14, 17, 21, 30, 40, 41, 48] crop or sample the original image into sub-regions at finer scales and train neural networks recurrently. To reduce expensive and redundant calculation cost caused by the sub-regions, Marin *et al*. [27] proposed a content-adaptive down-sampling technique to sample locations near semantic edges of target objects. However, the increase of the background resolution is useless. To this end, the method of [54] provided an attention-based sampler to enlarge at-

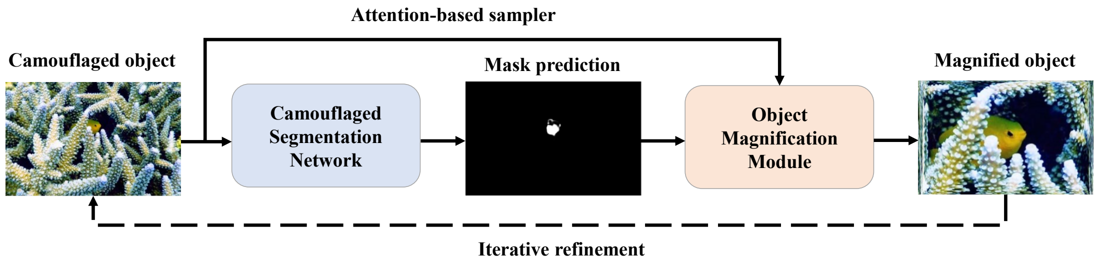

Figure 2. Pipeline of our SegMaR framework. The magnification module enlarges the proportion of the object while compressing that of the background, without increasing the image size. Due to limited space, we show the first stage only, while the following stages reiterate the same process. Please refer to the details in Section 3.

tended parts and decrease the background resolution, meanwhile the image size can be kept. One similar work to ours is [44] where they introduced the attention-based sampler [54] for SOD, while their method only magnified the object once. *Different from solving SOD, our work leverages more magnification steps for camouflaged objects and achieves further performance boosts.*

3. Segment, Magnify and Reiterate

Overview. This section introduces the SegMaR framework designed for COD. As depicted in Fig. 2, one can observe SegMaR is an iterative refinement framework trained in a multi-stage fashion. First of all, the input image is fed into a camouflaged segmentation network to generate a mask prediction with respect to the camouflaged object. Then it combines the input image and its mask into an attention based object magnification module, so as to enlarge the object while the image size can be kept. Next, we reiterate the segmentation process by taking as input the image with the magnified object. Consequently, the camouflaged object becomes more and more detectable from the background surroundings (Fig. 1).

Below, we detail the steps in the framework.

3.1. Camouflaged Segmentation Network

Like most of related works [8,43], our camouflaged segmentation network is built on top of a two-branch network architecture, see the left in Fig. 3. (1) For the first branch (shown in blue), it consists of four convolutional blocks and a discriminative decoder that generates a mask prediction P_{dis}. (2) The second branch (shown in green) adds three new convolutional blocks following the first block in the first branch. A binary decoder is responsible to inferring the final binary mask P_{bin} for COD. In addition, it is encouraged to use the first branch to help improve the learning process of the second branch. To make it, we merge the feature maps from the second convolutional block and the discriminative decoder in the first branch with the second branch, by using a holistic attention (HA) module [43].

The discriminative and binary decoders have the same network structure, see the right in Fig. 3. The input feature maps are firstly followed by atrous spatial pyramid pooling (ASPP) components [46] with dilation rate $D_r = 3, 6, 12, 18$, respectively. The aim is to achieve multi-scale receptive fields in the image. Then the pooling maps are concatenated together and passed into a distraction module (DM) [55]. DM is an effective technique to disentangle previous feature maps into foreground and background features, separately. We find this ability is significant particularly for *recognizing the subtle differences between camouflaged objects and background surroundings.* Different from [55], we tailor the DM module by adding two parallel residual channel attention blocks (RCAB) [52], which make the module concentrate more on informative channels and high-frequency information (*e.g.* edges, texture) in the feature maps. Afterwards, we use the element-wise subtraction to reverse the background feature and the element-wise addition to augment the foreground feature. The output feature f_d by the distraction operation is formulated by

$$f_d = BR\left(\beta f_a + BR(-\alpha f_b)\right), \tag{1}$$

where BR is the combination of batch normalization and ReLU, f_a and f_b represent the foreground and background features, respectively. α and β are two learnable parameters and initialized with 1. Lastly, another ASPP component following DM is added to make the output features.

Discriminative mask. In the wild, the *fixation regions* like the face or limbs, are the key clues for the predator being able to quickly locate camouflaged prey. Besides, the *edge regions* may also leak the location of the camouflaged objects, *e.g.* the hair of an animal. Thus, both fixation and edge regions are important to make the camouflaged object detectable. Typically, a binary mask (*i.e.* 255: object, 0: background) usually acts as the ground truth to train the COD model, which implies all the regions of the object weigh equally. However, this way neglects some important regions associated with the object. Although one recent work [25] adds new fixation annotations in addition to the binary mask, their fixation annotations have some wrong regions overflow the object region.

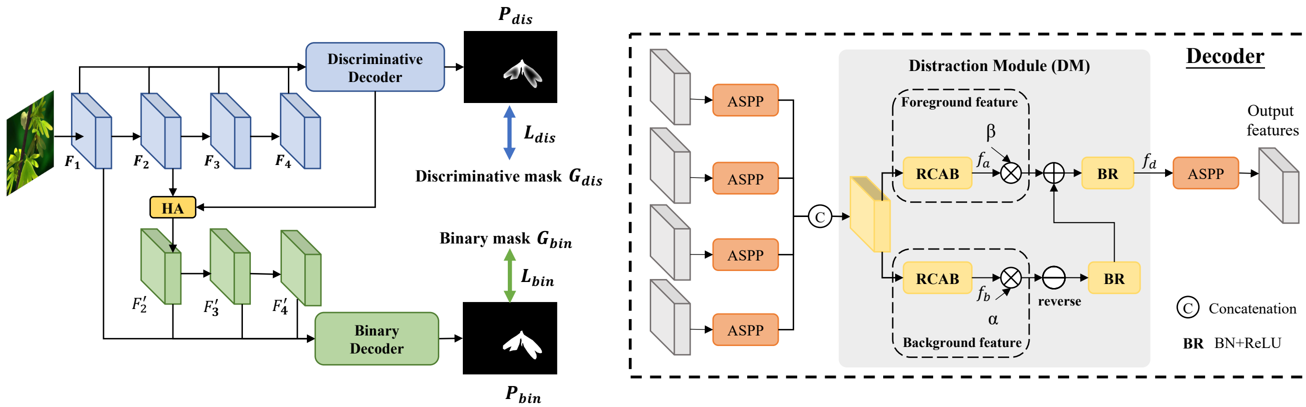

Figure 3. Our camouflaged segmentation network (Left) and its decoder (Right). The prediction P_{dis} by the first branch is supervised with the discriminative mask we present, while the prediction P_{bin} from the second branch is trained with the original binary mask. HA is holistic attention, ASPP is atrous spatial pyramid pooling, and RCAB is residual channel attention block.

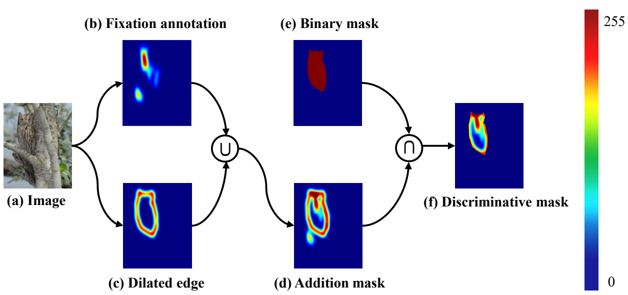

Figure 4. The calculation process of our discriminative mask. Here we use the colormaps for clearer visualization.

To solve this problem, we propose a richer and non-binary ground-truth annotation called *discriminative mask*. Beyond the original binary mask, our discriminative mask supervise the network to attend more on the fixation and edge regions. As for any image, we capture its edge annotation based on the binary mask, and then dilate the edges with Gaussian operation. The dilated edge captures more information around the object boundary. Then we merge the fixation annotation and dilated edge, resulting a addition mask. Lastly, we use the binary mask to subtract the overflow fixation region. Our discriminative mask G_{dis} is computed via

$$G_{dis} = G_{bin} \cap (G_{fix} \cup A(\sigma, \lambda, G_{edge})), \tag{2}$$

where $A(\cdot)$ is the Gaussian function with Gaussian blur σ = 15 and kernel size λ = 25. G_{bin} is binary mask based ground truth, G_{fix} and G_{edge} are fixation and edge annotations. Since G_{fix} is non-binary, G_{dis} is a non-binary mask ranging from 0 to 255. Figure 4 depicts the process of computing the discriminative mask. We illustrate some discriminative mask instances in Fig. 6, which renders stronger attention on significant regions.

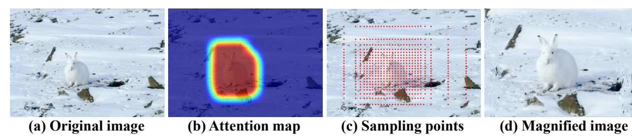

Figure 5. The attention based magnification process.

Loss function. The camouflaged segmentation network is trained end-to-end by two loss terms: the discriminative loss L_{dis} and the binary loss L_{bin}. L_{dis} indicates the loss cost between P_{dis} and G_{dis}, and L_{bin} is that between P_{bin} and G_{bin}. We adopt the structure loss in [42] to compute L_{dis} and L_{bin}. The structure loss $L_{str}(P, G)$ adds a weighted binary cross entropy (BCE) loss L^w_{bce} and a IoU loss L^w_{iou} by

$$L_{str}(P, G) = L^w_{bce}(P, G) + L^w_{iou}(P, G). \tag{3}$$

This structure loss is beneficial to maintain both pixel and global restrictions between the prediction and the ground truth. Finally, our total loss function is

$$L_{total} = L_{dis} + L_{bin} = L_{str}(P_{dis}, G_{dis}) + L_{str}(P_{bin}, G_{bin}). \tag{4}$$

3.2. Attention based Object Magnification

Camouflaged objects normally accounts for a very small proportion of the whole images, which makes it difficult to detect accurate object edges. Motivated by the fact that humans always move closer to the target in order to watch it more clearly, we propose to enlarge the camouflaged object while compressing the background information, as shown in Fig. 5.

Given the prediction mask P_{bin}, we further dilate it to be an attention map D via

$$D = Dilation(\sigma, \lambda, P_{bin}), \tag{5}$$

where Gaussian blur $\sigma = 15$ and kernel size $\lambda = 75$. The dilation operation aims to enlarge the original prediction area, and strengthen the integrity of the object region. In the second image of Fig. 5, the attention map completely covers the object. Then we employ an attention based sampler algorithm [54] to magnify the camouflaged object based on the attention map D. The attention map is used to calculate the mapping function between the coordinates of the original image and the sampled image, and the area with larger attention value is more likely to be sampled. We first decompose the attention map into two dimensions and obtain the marginal distribution by calculating the max values of the attention map D over x axis and y axis as

$$D_x = \max_{1 \le i \le w} D_i, D_y = \max_{1 \le j \le h} D_j, \tag{6}$$

where w and h is the width and height of D. Given the original image I, the sampling function $Sampler(I, D)$ is defined as

$$Sampler(I, D)_{i,j} = I_{D_x^{-1}(i), D_y^{-1}(j)}, \tag{7}$$

where $D^{-1}(\cdot)$ indicates the inverse function of $D(\cdot)$. Figure 5 demonstrates the area with high values in the attention map is dense sampled and magnified with its shape unchanged.

3.3. Iterative Refinement

The main advantage of SegMaR is its iterative refinement by replaying the Segment and Magnify steps in a multi-stage fashion. As shown in Fig. 1, the camouflaged crab becomes more detectable in an increasing resolution across the stages. During training period, all the stages share the same network parameters. In addition, we use the same hyper-parameters such as the Gaussian blur and kernel size for object magnification. The iterative refinement will terminate when the loss differences between two successive stages become subtle. Algorithm 1 summarizes the training steps in the SegMaR framework.

In terms of testing period, we need to restore the final mask prediction P_{bin} to the original object size, so that it can be aligned with the ground truth of the test images. We leverage a reversed sampling strategy of Eq. (7), which is denoted as $R_{sampler}(\cdot)$. The restored mask prediction is represented as $R_{sampler}(P_{bin})$.

Algorithm 1 Training SegMaR via Iterative Refinement

Input: Input images $I^{(i)}$ at the i-th stage, binary mask (G_{bin}), discriminative mask (G_{dis}), N stages
Output: COD network (Net)

```
for each stage i ∈ [1, N] do
// Segment step
    Net^(i) ⟵ train network with I^(i) as Eq. (4);
// Magnify step
    D^(i) ⟵ Dilation(σ, λ, G_bin^(i)) as Eq. (5)
    I^(i+1) ⟵ Sampler(I^(i), D^(i)) as Eq. (7)
    G_bin^(i+1) ⟵ Sampler(G_bin^(i), D^(i)) as Eq. (7)
    G_dis^(i+1) ⟵ Sampler(G_dis^(i), D^(i)) as Eq. (7)
// Reiterate step
    Initialize the next stage Net^(i+1) ⟵ Net^(i)
end for
```

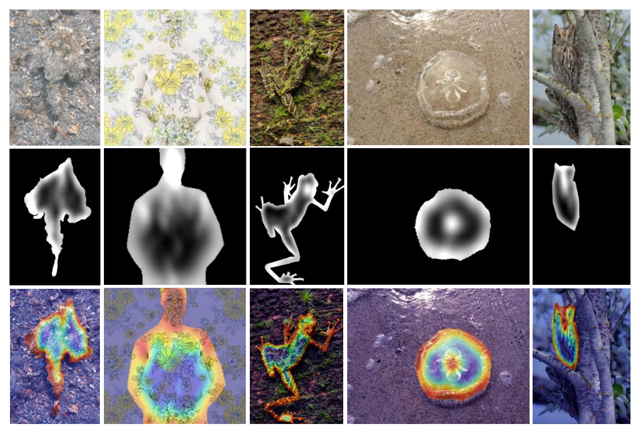

Figure 6. From top to bottom: original image, discriminative mask and colormap visualization. It can be seen that the fixation and edge regions have stronger attention.

4. Experiment Results

4.1. Setup and Evaluation

Dataset. We evaluate our method on three widely used datasets: CHAMELEON [37], CAMO [19], and COD10K [8]. CHAMELEON [37] includes 76 high-resolution images, which are collected from the Internet by using 'camouflaged animal' as the keyword. CAMO [19] is a collection of $1,250$ images with 8 categories. COD10K [8] is currently the largest benchmark, containing $10,000$ images with 10 super-classes and 78 sub-classes collected from photography websites. Following previous works, our training includes $1,000$ images from CAMO dataset, and $3,040$ images from COD10K, and the test set merges $2,026$ images from COD10K, 76 images from CHAMELEON and 250 images from CAMO. In addition to the binary mask based ground truth provided in the benchmarks, we also employ the discriminative mask when training the network, please see the examples in Fig. 6.

Implementation details. A pretrained ResNet50 [13] on ImageNet dataset [16] is employed as the backbone of our camouflaged segmentation network. All input images are resized to 352×352, and the output predictions are resized back to the original object sizes to compare with their binary ground truths. Bilinear interpolation is employed for image resizing. We adopt Adam optimizer [15] with learning rate

Method	COD10K-Test (2,026 images)				CAMO-Test (250 images)				CHAMELEON-Test (79 images)			
	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$
CPD [43]	0.752	0.820	0.557	0.049	0.712	0.813	0.561	0.108	0.860	0.908	0.753	0.044
PraNet [9]	0.768	0.836	0.599	0.047	0.738	0.814	0.613	0.098	0.864	0.918	0.784	0.038
MINet-R [31]	0.759	0.832	0.580	0.045	0.749	0.835	0.635	0.090	0.844	0.919	0.746	0.040
SINet [8]	0.771	0.807	0.565	0.048	0.742	0.834	0.601	0.101	0.869	0.903	0.749	0.041
LSR [25]	0.767	0.861	0.611	0.045	0.712	0.791	0.583	0.104	0.846	0.913	0.767	0.046
PFNet [28]	0.800	0.868	0.660	0.040	0.782	0.852	0.695	0.085	0.882	0.942	0.810	0.033
C^2F-Net [39]	0.810	0.875	0.674	0.038	0.791	0.863	0.706	0.083	0.886	0.931	0.824	0.032
MGL [49]	0.811	0.865	0.666	0.037	0.775	0.847	0.673	0.088	0.893	0.923	0.813	0.030
SegMaR (Stage-1)	0.813	0.880	0.682	0.035	0.805	0.864	0.724	0.072	0.892	0.937	0.823	0.028
SegMaR (Stage-2)	0.830	0.890	0.718	0.034	0.808	0.863	0.739	0.074	0.902	0.944	0.851	0.027
SegMaR (Stage-3)	0.833	0.892	**0.725**	0.034	0.810	0.870	**0.745**	0.073	0.905	0.947	0.858	0.027
SegMaR (Stage-4)	**0.833**	**0.895**	0.724	**0.033**	**0.815**	**0.872**	0.742	**0.071**	**0.906**	**0.954**	**0.860**	**0.025**

Table 1. Comparison of our method with other state-of-the-art methods on three benchmarks in terms of S_α (larger is better), αE (larger is better), wF (larger is better), and M (smaller is better). Stage-i (i=1,2,3,4) denotes the iterative stages of our multi-stage framework. The best scores highlighted in bold indicate our SegMaR outperforms other methods by achieving new top-performing accuracy.

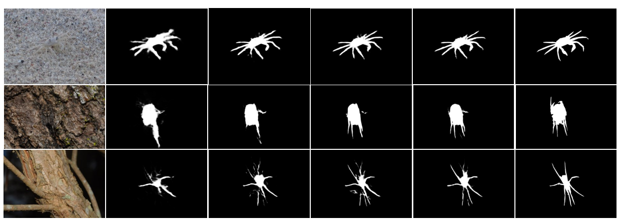

Figure 7. Visual comparison of our multi-stage detection framework. The fist stage has a rough contour of the object, while the following stages refine it. Please zoom-in to see the details.

of $2.5e-5$ and decay rate of 0.9.

We use PyTorch toolbox [33] to conduct the experiments on a GPU Tesla V100. Each training stage takes about 6 hours with batch size 24 and 50 epochs.

Evaluation Metrics. We employ four evaluation metrics, including mean absolute error (M), structure measure (S_α) [6], adaptive E-measure (αE) [10], and weight F-measure (wF) [26]. M is defined as element-wise difference between prediction map and binary ground truth. S_α is defined as $S_\alpha = \alpha S_o + (1-\alpha)S_r$, where S_o denotes object-aware structural similarity and S_r denotes region-aware structural similarity. αE evaluates the pixel-level similarity and the image level statistic simultaneously, which is related to human visual perception. wF is a comprehensive measure on both precision and recall, and recent works [6, 10] suggest that wF is more reliable than F-measure.

4.2. Comparison with the State-of-the-arts

We compare our SegMaR model with eight state-of-the-art COD methods, including CPD [43], PraNet [9], MINet [31], SINet [8], LSR [25], PFNet [28],C^2FNet [39] and MGL [49]. For a fair comparison, the results of these methods are directly provided by their authors or by their original trained model, and we test them with the same evaluation protocols. For our SegMaR model, we find the loss difference between two successive stages flattens within *four stages* for all three benchmarks. To validate the multi-stage learning framework, we list the performance of the proposed SegMaR in four stages, and compare them with other methods in Table 1. We can see the performance of SegMaR improves progressively with the increase of training stages, demonstrating the object magnification and iterative refinement help the model to achieve stronger detection ability. In addition to the quantitative results, Fig. 7 compares detection results across four stages qualitatively. Comparing with the ground truth in the last column, we can already obtain a rough area in the first stage. The detection results are improved with more details gradually in the following stages.

From the results reported in Table 1, the performance of our 1-*st* stage already outperforms other methods, which verifies the advantage of our camouflaged segmentation network. Furthermore, our 4-*th* stage achieves new state-of-the-art performance on three benchmarks. Specifically, SegMaR outperforms previous methods by a large margin on the most challenging dataset COD10K, like surpassing MGL [49] 3.5% on αE, and 8.7% on wF. We outperform MGL 2.0% in average over all metrics on CAMO dataset, and 2.4% in average on CHAMELEON dataset. Additionally, Fig. 8 shows the qualitative comparison of our method with other methods. We can see that our detection results are the closest to the ground-truth annotations, in terms of not only large camouflaged objects (*e.g.* the first row), but also small ones (*e.g.* the last four rows). This is mainly because the discriminative mask can provide the initial location of the camouflaged objects and enforce the attention on the contours. Moreover, benefited by the magnification process in the multi-stage training, our method can capture

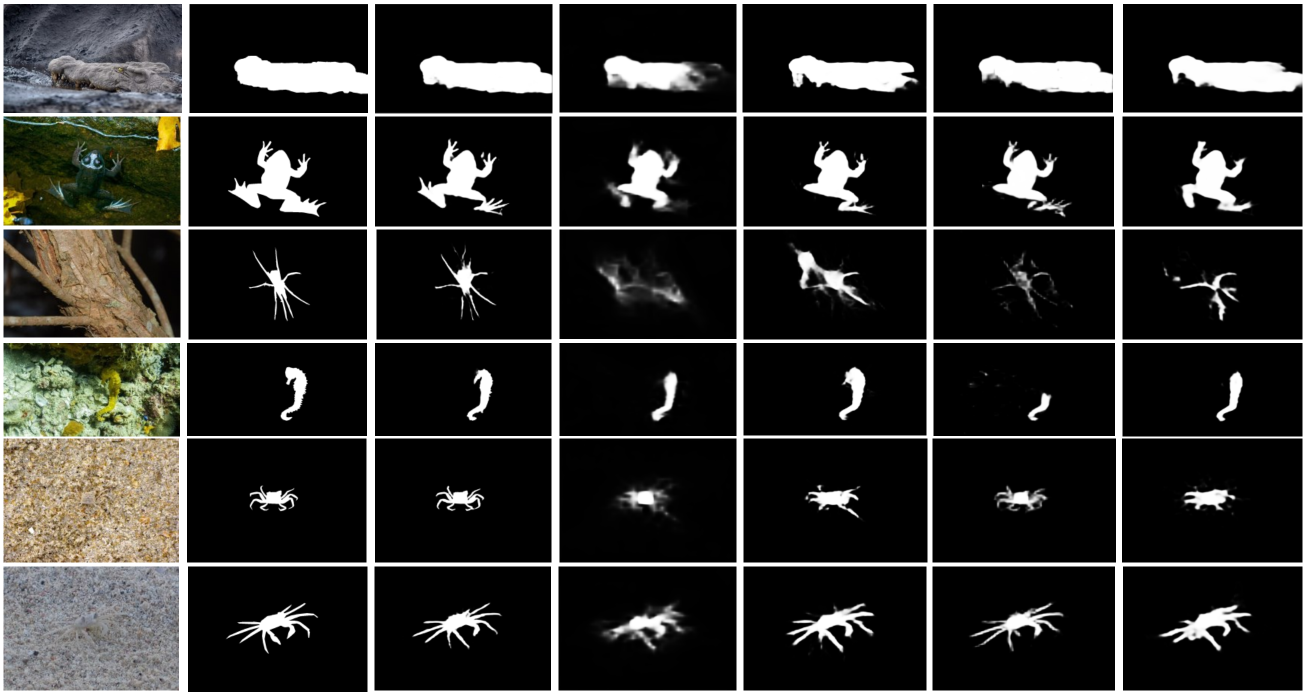

Figure 8. Visual comparison of the proposed SegMaR with recent state-of-the-art methods. Our method can distinguish the edges of camouflaged objects more clearly than other methods.

detailed distraction information and thus has the ability to finely segment the camouflaged objects with complex structures. The proposed SegMaR is far beyond other methods on the fine contour with only bits of pixels, such as the fingers of frog in the first row, the legs of spider in the second row, and crab's claws in the last two rows.

4.3. Ablation Study

We conduct ablation studies to validate our key components tailored specifically for accurate COD, including discriminative mask, small object detection, distraction module (DM) and generalization analysis.

Effectiveness of discriminative mask. Table 2 compares the performance of training SegMaR in the first stage based on four different types of ground truth, including fixation annotation, edge annotation, binary mask and our discriminative mask. Notice that these ground-truths are used to train the discriminative decoder in our segmentation network, while the binary decoder is always trained with the binary mask for a consistent comparison with previous works. Our discriminative mask surpasses fixation annotation and binary mask across all metrics. Edge annotation achieves better performance sometimes, but still under-performs our discriminative mask.

Small object detection. One can expect that it is challenging to segment small camouflaged objects with fine edges composed of limited pixels, such as fur or legs of living creatures.

In order to validate the effectiveness of SegMaR on small objects, we divide the testing set on COD10K into 'small' and 'non-small' subsets. The small subset contains $1,084$ images where the objects occupy less than $1/4$ of the image size, and the left 924 images belong to the 'non-small' subset. As shown in Table 4, we compare our performance at Stage-4 with two competitive methods, *i.e*. SINet and MGL. Our method has remarkable improvements over the two methods on small testing set, by surpassing SINet 8.0% on S_α, 16.8% on αE, 35.4% on wF, and outperforming MGL 7.4% in average over three metrics.

Effectiveness of distraction module (DM). To further investigate the effectiveness of the tailored components of our camouflaged segmentation network, Table 3 compares the performance with and without our distraction module. Specifically, adding the DM obtains 4.2% performance gain in terms of wF on COD10K test set. This validates the rationality of our design to learn distractions from the attentive input features.

Generalization analysis of SegMaR framework. We state that, SegMaR is a unified and general framework which shall be applicable to other camouflaged segmentation networks. To validate its generalization ability, we reiterate the segmentation, magnification steps on SINet [8]

SegMaR (Stage-1)	COD10K-Test (2,026 images)				CAMO-Test (250 images)				CHAMELEON-Test (76 images)			
	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$
fixation	0.806	0.871	0.669	0.037	0.784	0.855	0.693	0.083	0.884	0.931	0.808	0.032
edge	0.809	**0.882**	0.679	0.036	0.796	0.863	0.712	0.075	0.890	**0.944**	0.822	0.030
binary	0.810	0.877	0.680	0.036	0.799	0.857	0.719	0.075	0.887	0.920	0.818	0.031
discriminative	**0.813**	0.880	**0.682**	**0.035**	**0.805**	**0.864**	**0.724**	**0.072**	**0.892**	0.937	**0.823**	**0.028**

Table 2. Ablation analysis of using different ground-truth annotations to train the network. 'fixation' and 'edge' indicate fixation and edge annotations. 'Binary' indicates the binary mask, while 'discriminative' is our discriminative mask. Overall, our discriminative mask achieves the best performance on almost all evaluation metrics.

SegMaR (Stage-1)	COD10K-Test (2,026 images)				CAMO-Test (250 images)				CHAMELEON-Test (76 images)			
	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$
without DM	0.799	0.866	0.654	0.039	0.795	**0.865**	0.706	0.077	0.881	0.926	0.799	0.033
with DM	**0.813**	**0.880**	**0.682**	**0.035**	**0.805**	0.864	**0.724**	**0.072**	**0.892**	**0.937**	**0.823**	**0.028**

Table 3. Ablation analysis of our distraction module (DM) and its effect on our SegMaR framework. Overall, introducing DM brings considerable performance gains across datasets and metrics. We show the results at Stage-1 only due to limited space, while other stages witness consistent improvements.

Method		COD10K(2,026 images)					
		Small (1,084 images)			**Non-Small (924 images)**		
		$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$
SINet [8]		0.764	0.743	0.500	0.779	0.881	0.639
MGL [49]		0.796	0.823	0.598	0.832	0.911	0.743
SegMaR	stage-1	0.797	0.852	0.620	0.832	0.913	0.753
	stage-2	0.821	0.866	0.667	0.841	0.917	0.775
	stage-3	0.825	0.871	0.677	0.843	0.917	0.782
	stage-4	0.825	0.868	0.677	0.842	0.921	0.779

Table 4. Performance results on small and non-small test sets. 'Small' indicates the region of the camouflaged object is less than $1/4$ of the image. We show the results of SegMaR at Stage-4. The best performance in bold demonstrates our method surpass SINet and MGL by a large margin, particularly for small objects.

Method	Stage	COD10K-Test (2026 images)			
		$S_\alpha\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	$M\downarrow$
SINet [8] with SegMaR framwork	Stage-1	0.771	0.807	0.565	0.048
	Stage-2	0.795	0.847	0.639	0.043
	Stage-3	0.801	0.862	0.658	0.041
	Stage-4	**0.805**	**0.869**	**0.667**	**0.041**

Table 5. Generalization analysis of the proposed SegMaR framework, by applying multi-stage iterative refinement to SINet [8].

which is a recent and competitive baseline in the field. Comparing the results from Stage-1 to Stage-4 in Table 5, SINet gains 4.4%, 7.7%, and 18.0% improvements in terms of S_α, αE, and wF, validating our potential and strong generalization to other alternatives.

4.4. Limitations and Discussions

We discuss two potential limitations in this work:

Q1: Why the whole SegMaR framework is not end-to-end trainable. The main reason is the object magnification module we introduce is a non-parametric approach. Alternatively, we did consider leveraging a neural network to implement the magnification module. However, this solution requires new ground-truth annotations with respect to the magnified objects. Otherwise, it is hard to supervise the magnification network and achieve desirable results. We will devote to learning an unsupervised magnification network, with no need of extra annotations.

Q2: When to terminate the iterative refinement stages. Here, we terminate the iterative refinement when the loss difference between two successive stages is subtle. As a result, our SegMaR reaches saturation after four stages only. This training process is simple yet lacks theoretical evidence. Instead, it is promising to design new algorithms to optimize the multi-stage training process, so as to reiterate more stages and achieve better performance.

5. Conclusion

To simulate humans attention which segments camouflaged objects in a coarse-to-fine manner, we have proposed an iterative refinement framework SegMaR to integrate Segment, Magnify and Reiterate in a multi-stage detection fashion. We also designed a new discriminative mask and distraction module to make the network segment more object regions. Extensive experiments have demonstrated our top-performing performance on three benchmarks especially for small camouflaged objects. In the future, it is promising to study more sophisticated magnification algorithms.

Acknowledgements. This work is partially supported by the NSF of China under grant Nos. 61876030, 61733002, 61572105, 61922019, 62027826 and 62102061.

References

[1] Sayanti Bardhan. Salient object detection by contextual refinement. In *CVPR*, pages 1464–1472, 2020. 2

[2] Arantxa Casanova, Guillem Cucurull, Michal Drozdzal, Adriana Romero, and Yoshua Bengio. On the iterative refinement of densely connected representation levels for semantic segmentation. In *CVPR Workshop*, pages 978–987, 2018. 2

[3] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, volume 11213, pages 236–252, 2018. 1, 2

[4] Bo Dong, Mingchen Zhuge, Yongxiong Wang, Hongbo Bi, and Geng Chen. Towards accurate camouflaged object detection with mixture convolution and interactive fusion. *CoRR*, abs/2101.05687, 2021. 1

[5] Hui Du, Xiaogang Jin, and Xiaoyang Mao. Digital camouflage images using two-scale decomposition. *Comput. Graph. Forum*, 31(7):2203–2212, 2012. 1

[6] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017. 6

[7] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *CoRR*, abs/2102.10274, 2021. 1, 2

[8] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2774–2784, 2020. 1, 2, 3, 5, 6, 7, 8

[9] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention, MICCAI*, volume 12266, pages 263–273, 2020. 1, 6

[10] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 51(9), 2021. 6

[11] Golnaz Ghiasi and Charless C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 519–534, 2016. 2

[12] Jicheng Gong, Zhao Zhao, and Nic Li. Improving multistage object detection via iterative proposal refinement. In *BMVC*, page 223, 2019. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[14] Chen Jin, Ryutaro Tanno, Moucheng Xu, Thomy Mertzanidou, and Daniel C. Alexander. Foveation for segmentation of ultra-high resolution images. *CoRR*, abs/2007.15124, 2020. 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012. 5

[17] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. In *CVPR*, pages 3668–3677, 2016. 2

[18] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Asian Conference on Computer Vision,ACCV*, pages 488–503, 2020. 2

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

[20] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *CVPR*, pages 10071–10081, 2021. 1, 2

[21] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015. 2

[22] Liyuan Li, Weimin Huang, Irene Y. H. Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.*, 13(11):1459–1472, 2004. 1

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 5168–5177, 2017. 2

[24] Jiawei Liu, Jing Zhang, and Nick Barnes. Confidence-aware learning for camouflaged object detection. *CoRR*, abs/2106.11641, 2021. 2

[25] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021. 2, 3, 6

[26] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *CVPR*, pages 248–255, 2014. 6

[27] Dmitrii Marin, Zijian He, Peter Vajda, Priyam Chatterjee, Sam S. Tsai, Fei Yang, and Yuri Boykov. Efficient segmentation: Learning downsampling near semantic boundaries. In *ICCV*, pages 2131–2141, 2019. 2

[28] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, pages 8772–8781, 2021. 1, 2, 6

[29] Andrew Owens, Connelly Barnes, Alex Flint, Hanumant Singh, and William T. Freeman. Camouflaging an object from many viewpoints. In *CVPR*, pages 2782–2789. IEEE Computer Society, 2014. 1, 2

[30] Anuj Pahuja, Avishek Majumder, Anirban Chakraborty, and R. Venkatesh Babu. Enhancing salient object segmentation through attention. In *CVPR*, pages 27–36, 2019. 2

[31] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9410–9419, 2020. 6

[32] Alejandro Pardo, Humam Alwassel, Fabian Caba Heilbron, Ali K. Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. In

IEEE Winter Conference on Applications of Computer Vision, pages 3318–3327, 2021. 2

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 6

[34] Pedro Oliveira Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 75–91, 2016. 2

[35] Rakesh Nattoji Rajaram, Eshed Ohn-Bar, and Mohan M. Trivedi. Refinenet: Iterative refinement for accurate object localization. In *IEEE International Conference on Intelligent Transportation Systems*, pages 1528–1533, 2016. 2

[36] Jingjing Ren, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Yangyang Xu, Weiming Wang, Zijun Deng, and Pheng-Ann Heng. Deep texture-aware features for camouflaged object detection. *CoRR*, abs/2102.02996, 2021. 2

[37] P Skurowski, H Abdulameer, J Błaszczyk, T Depta, A Kornacki, and P Kozieł. Animal camouflage analysis:chameleon database, 2018. 5

[38] Martin Stevens and Sami Merilaita. Animal camouflage: Current issues and new perspectives. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364:423–7, 12 2008. 1

[39] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In Zhi-Hua Zhou, editor, *IJCAI*, pages 1025–1031, 2021. 6

[40] Jing Tan, Pengfei Xiong, Zhengyi Lv, Kuntao Xiao, and Yuwen He. Local context attention for salient object segmentation. In *Asian Conference on Computer Vision,ACCV*, pages 706–722, 2020. 2

[41] Lv Tang, Bo Li, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. *CoRR*, abs/2108.03551, 2021. 2

[42] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. *CoRR*, abs/1911.11445, 2019. 2, 4

[43] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019. 1, 2, 3, 6

[44] Binwei Xu, Haoran Liang, Ronghua Liang, and Peng Chen. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *AAAI*, pages 3004–3012, 2021. 3

[45] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021. 1

[46] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018. 3

[47] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K. Fishman, and Alan L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *CVPR*, pages 8280–8289, 2018. 2

[48] Yi Zeng, Pingping Zhang, Zhe L. Lin, Jianming Zhang, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7233–7242, 2019. 2

[49] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *CVPR*, pages 12997–13007, 2021. 2, 6, 8

[50] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *CVPR*, pages 5217–5226, 2019. 2

[51] Xiang Zhang, Ce Zhu, Shuai Wang, Yipeng Liu, and Mao Ye. A bayesian approach to camouflaged moving object detection. *IEEE Trans. Circuits Syst. Video Technol.*, 27(9):2001–2013, 2017. 1, 2

[52] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, volume 11211, pages 294–310, 2018. 3

[53] Jiaxing Zhao, Jiangjiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8778–8787, 2019. 1, 2

[54] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, pages 5012–5021, 2019. 2, 3, 5

[55] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. Distraction-aware shadow detection. In *CVPR*, pages 5167–5176, 2019. 3