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 Segment, Magnify and Reiterate 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.

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 $S_a$ 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...
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-Concept 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-
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 $P_{\text{dis}}$. (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 $P_{\text{bin}}$ 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

$$f_d = BR (\beta f_a + BR(-\alpha f_b)),$$

where $BR$ is the combination of batch normalization and ReLU, $f_a$ and $f_b$ represent the foreground and background features, respectively. $\alpha$ and $\beta$ 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 $P_{\text{dis}}$ by the first branch is supervised with the discriminative mask we present, while the prediction $P_{\text{bin}}$, 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 $G_{\text{dis}}$ is computed via

$$G_{\text{dis}} = G_{\text{bin}} \cap (G_{\text{fix}} \cup A(\psi, \lambda, G_{\text{edge}})),$$

where $A(\cdot)$ is the Gaussian function with Gaussian blur $\psi = 15$ and kernel size $\lambda = 25$. $G_{\text{bin}}$ is binary mask based ground truth, $G_{\text{fix}}$ and $G_{\text{edge}}$ are fixation and edge annotations. Since $G_{\text{fix}}$ is non-binary, $G_{\text{dis}}$ 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.

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

$$L_{\text{str}}(P, G) = L_{\text{bce}}^w(P, G) + L_{\text{iou}}^w(P, G).$$

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

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

3.2. Attention based Object Magnification

Camouflaged objects normally account 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 $P_{\text{bin}}$, we further dilate it to be an attention map $D$ via

$$D = \text{Dilation}(\sigma, \lambda, P_{\text{bin}}).$$
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

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

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

$$\text{Sampler}(I, D)_{i,j} = I_{D^{-1}_x(i), D^{-1}_y(j)},$$

where $D^{-1}()$ indicates the inverse function of $D()$. 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 $P_{\text{bin}}$ 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 $R_{\text{Sampler}}(\cdot)$. The restored mask prediction is represented as $R_{\text{Sampler}}(P_{\text{bin}})$.

### 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.

Algorithm 1 Training SegMaR via Iterative Refinement

**Input:** Input images $I^{(i)}$ at the $i$-th stage, binary mask ($G_{\text{bin}}^{(i)}$), discriminative mask ($G_{\text{dis}}^{(i)}$, $N$ stages

**Output:** COD network ($\text{Net}$)

1: for each stage $i \in [1, N]$ do
2: // Segment step
3: $\text{Net}^{(i)} \leftarrow$ train network with $I^{(i)}$ as Eq. (4);
4: // Magnify step
5: $D^{(i)} \leftarrow \text{Dilation}(\sigma, \lambda, G_{\text{bin}}^{(i)})$ as Eq. (5)
6: $I^{(i+1)} \leftarrow \text{Sampler}(I^{(i)}, D^{(i)})$ as Eq. (7)
7: $G_{\text{bin}}^{(i+1)} \leftarrow \text{Sampler}(G_{\text{bin}}^{(i)}, D^{(i)})$ as Eq. (7)
8: $G_{\text{dis}}^{(i+1)} \leftarrow \text{Sampler}(G_{\text{dis}}^{(i)}, D^{(i)})$ as Eq. (7)
9: // Reiterate step
10: Initialize the next stage $\text{Net}^{(i+1)} \leftarrow \text{Net}^{(i)}$
11: end for

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 $352 \times 352$, 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
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 ($S_o$) [6], adaptive E-measure ($\alpha E$) [10], and weight F-measure ($wF$) [26]. $M$ is defined as element-wise difference between prediction map and binary ground truth. $S_o$ is defined as $S_o = \alpha S_o + (1 - \alpha)S_r$, where $S_o$ denotes object-aware structural similarity and $S_r$ denotes region-aware structural similarity. $\alpha E$ 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-R [31], SINet [8], LSR [25], PFNet [28], $C^2F$-Net [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 $\alpha E$, 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
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 $S_{\alpha}$, 16.8% on $\alpha E$, 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]
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.

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.

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.

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 $S_a$, $\alpha E$, and $w F$, 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

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


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


