Reference Prompted Model Adaptation for Referring Camouflaged Object Detection

Xuewei Liu[1,2*] Shaofei Huang[1,2*] Ruipu Wu[3] Hengyuan Zhao[3]
Duo Xu[3] Xiaoming Wei[4] Jizhong Han[1,2†] Si Liu[3,5]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Institute of Artificial Intelligence, Beihang University [4]Meituan
[5]Hangzhou Innovation Institute, Beihang University

***Abstract*—The goal of referring camouflaged object detection is to identify and segment the specified object hidden in the surroundings given text or images as references. The previous method still faces limitations in learning discriminative object features and comprehensively exploiting reference information due to coarse reference-image fusion upon disunified network components. In this paper, we propose a novel Reference Prompted Model Adaptation (RPMA) pipeline that employs rich and fine-grained semantic knowledge in a generic segmentation network to enhance the Ref-COD model's capability. Within RPMA, we design a Cross Reference Adapter (CRA) to integrate reference information into the generic segmentation network to prompt reference-relevant camouflaged image features, and also devise a Reference-guided Dynamic Convolution (RDC) for foreground-background segmentation via reference-generated kernels. Extensive experiments on the Ref-COD benchmark show that our method achieves new state-of-the-art performance.**

***Index Terms*—Referring Camouflaged Object Detection, Reference Prompting, Model Adaptation**

I. Introduction

Camouflaged Object Detection [1] (COD) aims to identify and segment objects that are naturally hidden in their surrounding environment, which is difficult even for humans to distinguish the camouflaged objects without explicit targets. Therefore, the task of Ref-COD [2] is recently proposed to guide the detection process with additional text or images as references (*e.g.*, class name or salient images belonging to the target class of camouflaged object), accomplishing more specified and applicable COD.

However, Ref-COD still presents substantial challenges. First, a significant representation gap exists between the references and camouflaged images. For instance, a reference image often depicts the salient appearances of the target object, differing remarkably from their camouflaged state. The modality discrepancy between reference text and camouflaged image also hinders the model's ability to comprehend textual cues for assisting object identification and segmentation. Additionally, it is also hard to learn discriminative features like structures and textures of camouflaged objects from limited training samples, which hampers the segmentation performance. As shown in Fig. 1, the previous method [2] conducts coarse fusion of references and image features upon disunified segmentation network components and only leverages limited training data of the Ref-COD dataset, which still faces limitations in comprehensively exploiting reference information and learning discriminative object features.

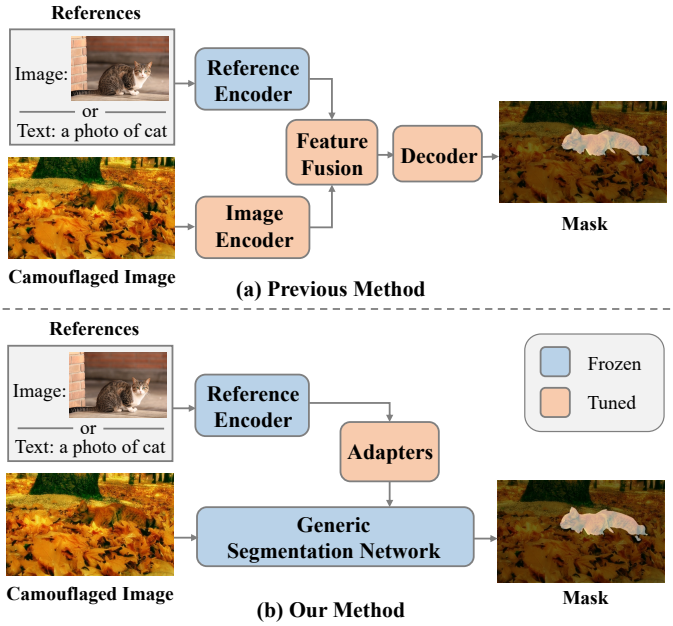

Fig. 1: Methodology comparison. (a) The previous method conducts coarse reference-image fusion upon disunified network components with limited reference exploitation and feature learning ability. (b) Our method prompts image features relevant to the reference information using reference-integrated adapters to leverage the knowledge of a generic segmentation network for building a better Ref-COD model.

Despite not having encountered extensive data of certain objects in camouflaged scenarios, humans can still make judgments using their rich prior knowledge and familiarity with the common appearances of these objects. In light of this insight, we propose to utilize a generic segmentation network trained on large-scale data as the foundation model, whose

*Equal contribution (liuxuewei@iie.ac.cn, nowherespyfly@gmail.com)
†Corresponding author

rich and fine-grained semantic knowledge can be leveraged to enhance the Ref-COD model's capability in identifying and segmenting camouflaged objects. Concretely, we propose a Reference Prompted Model Adaptation pipeline, abbreviated as RPMA, to adapt a generic segmentation network to the Ref-COD task. Within our RPMA pipeline, to fully understand the reference information and reduce the discrepancy with the camouflaged image, a Cross Reference Adapter (CRA) is designed in the segmentation encoder. Utilizing the cross-attention mechanism, reference information is integrated into the segmentation network in a residual form by CRA, aiding in prompting camouflaged image features relevant to the reference information. In the segmentation decoder, we also devise a Reference-guided Dynamic Convolution (RDC) to better capture the contours of the camouflaged object. RDC employs reference information to generate dynamic convolution kernels for foreground-background segmentation of the camouflaged image, effectively distinguishing relevant areas as foreground and irrelevant ones as background, thereby achieving improved segmentation results.

The contributions of our paper are summarized as follows:

- We propose a novel Reference Prompted Model Adaptation (RPMA) pipeline, leveraging the rich, fine-grained semantic knowledge in a generic segmentation network to enhance the Ref-COD model's capability.
- Within RPMA, we design a Cross Reference Adapter (CRA) to prompt camouflaged image features relevant to the reference information, and also develop a Reference-guided Dynamic Convolution (RDC) for foreground-background segmentation via reference-generated kernels.
- Extensive experiments show that our method achieves new state-of-the-art performance on the Ref-COD benchmark.

II. Related Work

A. Camouflaged Object Detection

The goal of Camouflaged Object Detection (COD) [1], [3] is to identify and segment objects that are naturally concealed in their surrounding environment, which is challenging due to the inherent property of camouflaged objects to visually merge with the background. To solve this task, techniques such as multi-scale feature fusion [4], [5], multi-source aggregation [6], and uncertainty learning [7] have been proposed to improve segmentation performance. With the great success of Transformer [8] in vision tasks, its application in COD tasks has also been widely explored due to its superiority in capturing long-range dependencies. For example, UGTR [9] focuses on learning the uncertainty of camouflaged objects with multi-head self-attention. FSPNet [10] designs a feature shrinkage pyramid Transformer to hierarchically decode transformer features through progressive shrinking for camouflaged object detection. In this paper, we focus on the problem of Referring Camouflaged Object Detection (Ref-COD) [2], which guides the detection process with additional text or images as references. Ref-COD prevents searching aimlessly for different regions, thus realizing more specified camouflaged object detection. A baseline named R2CNet [2] is also proposed to fuse the reference feature with the camouflaged image feature to segment reference-relevant camouflaged objects. Compared with R2CNet, our proposed RPMA pipeline designs reference-integrated adapters and reference-generated dynamic convolutions to adapt a pre-trained generic segmentation network, which utilizes its rich and fine-grained knowledge for better Ref-COD performance.

B. Model Adaptation

The progress of large pre-trained models such as language models [11], [12] has spurred increased exploration to adapt these powerful pre-trained models to specific downstream tasks. Earlier adaptation methods typically adopt a full-model finetuning strategy, which is time-consuming and requires numerous high-quality datasets and computational resources. To overcome this limitation, several effective model adaptation methods have been proposed for efficient downstream model adaptation, such as Adapters [13], LoRA [14], and Prompt Tuning [15]. These methods either introduce a subset of learnable parameters for funetuning or learn to update attention weights by training low-rank matrices. For vision tasks, with the rise of large-scale pre-trained visual models [16], [17], model adaptation methods have also been explored across a wild range, such as Visual Prompt Tuning (VPT) [18] and AdaptFomer [19]. In this paper, we leverage reference information through adapters to prompt reference-relevant image features from a pre-trained generic segmentation network for adapting it to the Ref-COD task, thus integrating both reference information and semantic knowledge effectively.

III. Method

In this section, we first introduce the overview of our RPMA pipeline in Section III-A. Then, we elaborate on the proposed CRA and RDC in section III-B and section III-C respectively.

A. Reference Prompted Model Adaptation

As shown in Fig. 2, the inputs to our RPMA pipeline include a camouflaged image $\boldsymbol{I}_c$ and references, which could be reference images $\boldsymbol{I}_r$ or reference text $\boldsymbol{T}_r$. Under the guidance of these references, the desired output of our pipeline is the segmentation mask $\boldsymbol{M}$ of the camouflaged objects. We adopt a pre-trained SegFormer [17] as the generic segmentation network to process the camouflaged image $\boldsymbol{I}_c$, which consists of a multi-layer Transformer encoder and a lightweight MLP decoder. Its encoder is composed of four stages with L blocks per stage. We denote the obtained visual features of $\boldsymbol{I}_c$ from the four stages as $\{\boldsymbol{F}_c^i\}_{i=1}^4 \in \mathbb{R}^{H^i \times W^i \times C^i}$, where H^i and W^i equals the height H and width W of $\boldsymbol{I}_c$ downsampled for 2^{i+1} times, C^i is the number of channels. For the references, we utilize the image encoder and text encoder of a pre-trained CLIP model [20] to process the reference images $\boldsymbol{I}_r$ and reference text $\boldsymbol{T}_r$, respectively. Since $\boldsymbol{I}_r$ and $\boldsymbol{T}_r$ are both tokenized before the encoders, we use $\boldsymbol{F}_r \in \mathbb{R}^{N \times C_r}$ to denote

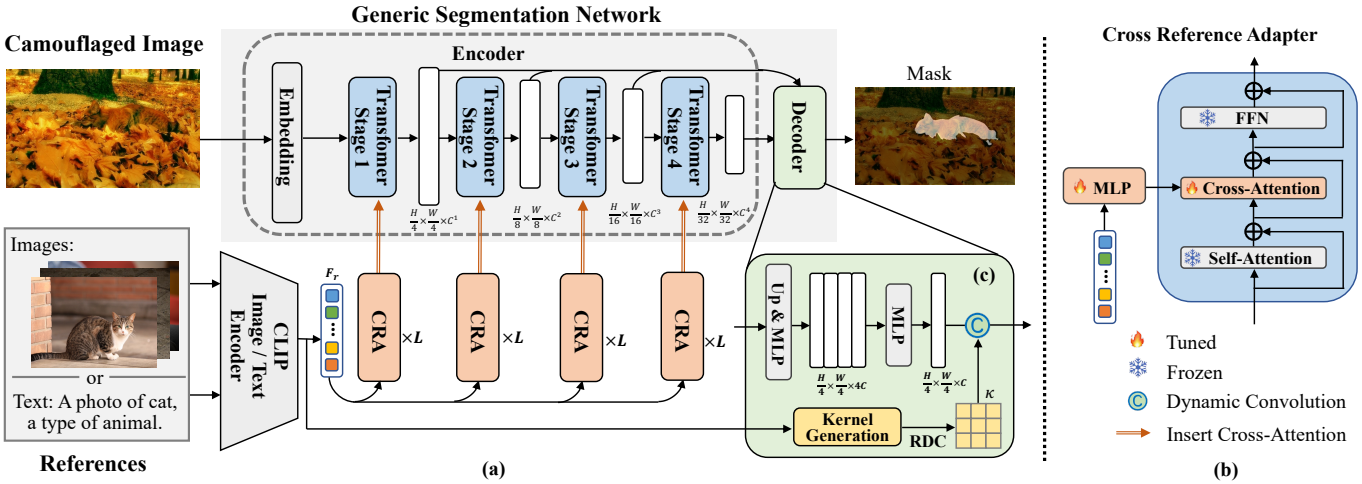

Fig. 2: (a) Overall architecture of our proposed Reference Prompted Model Adaptation (RPMA) pipeline. In the encoder, the Cross Reference Adapter (CRA) is densely integrated into Transformer blocks to prompt image features relevant to the reference information. In the decoder, a Reference-guided Dynamic Convolution (RDC) is incorporated for foreground-background segmentation. (b) The detailed structure of our proposed Cross Reference Adapter. (c) The decoder of the segmentation network, where RDC is utilized to enhance the prediction head.

the reference features uniformly, where N is the number of image patch tokens or text tokens.

To adapt SegFormer to the Ref-COD task, enabling it to segment camouflaged images based on references, we design a Cross Reference Adapter (CRA) and densely insert it into every Transformer layer of the SegFormer encoder to integrate reference information into the pipeline. To enhance the model's adaptability to different camouflaged objects, a Reference-guided Dynamic Convolution (RDC) is incorporated in the decoder to facilitate more precise foreground-background segmentation. Apart from the parameters of our CRA and RDC, all other parameters remain frozen.

B. Cross Reference Adapter

The purpose of our Cross Reference Adapter is to integrate the multimodal reference information into the generic segmentation network, thereby prompting image features relevant to the camouflaged object. As illustrated in Fig. 2(b), we achieve this by inserting a cross-attention layer into the Transformer blocks of the SegFormer encoder. Taking the l^{th} Transformer block in the i^{th} stage as an example, we denote the input features of the camouflaged image as $\boldsymbol{F}_c^{i,l-1}$. For presentation simplicity, the stage index i is omitted in the following notations as $\boldsymbol{F}_c^{l-1}$. In the original Transformer block, $\boldsymbol{F}_c^{l-1}$ is processed by a self-attention layer followed by an FFN (Feed-Forward Network) layer. To preserve the original feature information of SegFormer as much as possible, we insert the cross-attention layer in a residual form between the self-attention layer and the FFN layer. The output feature $\bar{\boldsymbol{F}}_c^{l-1}$ from self-attention is adopted as the query embedding of the cross-attention layer. For the key and value embedding, we first use an MLP layer to transform the reference feature $\boldsymbol{F}_r$ into the feature space of the camouflaged image as $\bar{\boldsymbol{F}}_r \in \mathbb{R}^{N\times C^i}$ and then apply two linear projections on it to obtain the embeddings. Concretely, the cross-attention operation is implemented in a multi-head manner and we take one head as an example to illustrate its calculation as follows:

$$\boldsymbol{F}_a^{l,m} = \text{Softmax}\left(\frac{(\bar{\boldsymbol{F}}_c^{l-1}\boldsymbol{W}_q)(\bar{\boldsymbol{F}}_r\boldsymbol{W}_k)^T}{\sqrt{C^i}}\right)(\bar{\boldsymbol{F}}_r\boldsymbol{W}_v), \quad (1)$$

$$\boldsymbol{F}_a^{l} = \text{Concat}\left(\boldsymbol{F}_a^{l,1}, \cdots, \boldsymbol{F}_a^{l,M}\right), \quad (2)$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$ denote parameters of query, key, and value projection layers, M denotes the number of attention heads, and $\text{Concat}(\cdot)$ denotes concatenating features along channel dimensions. $\boldsymbol{F}_a^l$ is added to the output of the self-attention layer as a residual and then fed into the FFN layer to obtain the output of the l^{th} Transformer block $\boldsymbol{F}_c^l$, which serves as the input to the subsequent block. It is important to note that within each block, the parameters of the FFN and self-attention layers remain unchanged, with only the parameters of cross-attention layers trainable.

C. Reference-guided Dynamic Convolution

To facilitate the decoder's ability to accurately capture the contours of the camouflaged object based on reference information, we devise a Reference-guided Dynamic Convolution (RDC) to enhance the prediction head in the decoder. The kernel parameter of RDC are generated from the reference feature $\boldsymbol{F}_r$, so that reference-relevant areas can be identified as foreground from irrelevant ones, thus realizing improved segmentation performances. To obtain the kernel parameters, we first obtain an affinity map from $\boldsymbol{F}_r$:

$$\boldsymbol{A} = \text{Softmax}(\boldsymbol{F}_r\boldsymbol{W}_a), \quad (3)$$

where $\boldsymbol{W}_a \in \mathbb{R}^{C_r\times k^2}$ denotes the parameter of projection layer, k denotes the size of dynamic convolution kernel which

Models	Overall				Single-obj				Multi-obj			
	S-measure ↑	α E-measure↑	w F-measure ↑	MAE↓	S-measure ↑	α E-measure↑	w F-measure ↑	MAE↓	S-measure ↑	α E-measure↑	w F-measure ↑	MAE↓
PreyNet-RefS [21]	0.817	0.900	0.704	0.032	0.822	0.900	0.709	0.032	0.763	0.898	0.645	0.041
PreyNet-RefT [21]	0.816	0.901	0.705	0.033	0.821	0.900	0.710	0.032	0.759	0.902	0.648	0.041
DGNet-RefS [22]	0.821	0.891	0.696	0.032	0.827	0.890	0.703	0.031	0.748	0.879	0.607	0.045
DGNet-RefT [22]	0.824	0.891	0.701	0.032	0.830	0.892	0.709	0.031	0.745	0.873	0.596	0.046
SINetV2-RefS [3]	0.823	0.888	0.700	0.033	0.828	0.889	0.705	0.032	0.771	0.874	0.634	0.043
SINetV2-RefT [3]	0.822	0.887	0.696	0.033	0.827	0.888	0.702	0.032	0.766	0.866	0.629	0.043
ZoomNet-RefS [5]	0.834	0.886	0.720	0.029	0.839	0.887	0.726	0.029	0.781	0.876	0.652	0.038
ZoomNet-RefT [5]	0.835	0.897	0.725	0.029	0.839	0.897	0.731	0.028	0.783	0.889	0.661	0.038
BSANet-RefS [23]	0.830	0.912	0.727	0.030	0.827	0.913	0.733	0.030	0.774	0.895	0.655	0.039
BSANet-RefT [23]	0.830	0.914	0.730	0.030	0.834	0.915	0.734	0.029	0.784	0.898	0.674	0.036
BGNet-RefS [24]	0.840	0.909	0.738	0.029	0.844	0.910	0.742	0.029	0.792	0.887	0.679	0.036
BGNet-RefT [24]	0.840	0.912	0.739	0.029	0.844	0.914	0.745	0.028	0.791	0.888	0.677	0.038
RPMA-RefS(ours)	**0.862**	**0.930**	**0.784**	**0.023**	**0.867**	**0.934**	**0.791**	**0.023**	**0.806**	**0.894**	**0.718**	**0.033**
RPMA-RefT(ours)	**0.861**	**0.928**	**0.783**	**0.024**	**0.867**	**0.931**	**0.789**	**0.023**	**0.802**	**0.890**	**0.717**	**0.034**

TABLE I: Comparison with previous state-of-the-art Ref-COD methods. "-RefS": Methods with reference images. "-RefT": Methods with reference text. "Single-obj": Scenes of a single camouflaged object. "Multi-obj": Scenes of multiple camouflaged objects. "Overall": All scenes containing camouflaged objects. "↑": The higher the better. "↓": The lower the better.

is empirically set to 3, and the Softmax operation is conducted along the first dimension to obtain the normalized affinity map $\boldsymbol{A} \in \mathbb{R}^{N\times k^2}$. The kernel parameter is thus generated by weighted summation of reference feature:

$$\boldsymbol{\kappa} = \boldsymbol{A}^T(\boldsymbol{F}_r\boldsymbol{W}_1)\boldsymbol{W}_2, \tag{4}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are both parameters of projection layers. The generated kernel parameter $\boldsymbol{\kappa}$ is then reshaped to $k \times k \times C \times 1$, with the first two dimensions denoting kernel size, and the last two denoting channel numbers of input and output respectively. As shown in Fig. 2(c), the dynamic convolution kernel $\boldsymbol{\kappa}$ is applied to the fused multi-scale image features to obtain the final segmentation mask.

IV. Experiments

A. Experimental Setup

Dataset. We conduct experiments on the R2C7K dataset [2], which is composed of the Camo and Ref subsets. The Camo subset consists of images with camouflaged objects, covering 64 animal categories and counting to a total of 5,015 images. While the Ref subset includes images with salient objects, which provides 25 salient images for each animal category, totaling 1,600 masked images.

Metrics. Following R2CNet [2], we report evaluation results on four metrics, including structural-measure (S_m), adaptive E-measure (αE), weighted F-measure (ωF), and Mean Absolute Error (MAE).

- **Structural measure metric** [25] (S_m) evaluates the structural similarity between corresponding regions and targets of predicted and true values by leveraging a region-aware structural similarity measure S_r and object-aware structural similarity measure S_o with a weight $\alpha \in [0, 1]$:

$$S = \alpha * S_o + (1 - \alpha) * S_r. \tag{5}$$

- **Adaptive E-measure metric** [26] (αE) evaluates the element-level and image-level similarity by leveraging an enhanced alignment matrix ϕ which is generated from the predicted foreground map and the binary ground-truth:

$$E = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\phi(x, y). \tag{6}$$

- **Weighted F-measure metric** [27] (ωF) evaluates the *weighted Recall* and *weighted Precision* between predicted and true values with a non-negative weight β^2 as follows:

$$F_\beta^\omega = \frac{(1+\beta^2)Precision^\omega \times Recall^\omega}{\beta^2 Precision^\omega + Recall^\omega}. \tag{7}$$

Implementation Details. We use SegFormer-B4 [17] as the generic segmentation network of our pipeline. The default number of reference images is set to 10. During the training phase, the batch size is set to 4, and the learning rate is set to 2×10^{-4}. We use AdamW optimizer for optimization. The model undergoes 50 epochs of training on a single NVIDIA TITAN RTX GPU, using a cosine annealing method to gradually decrease the learning rate. The camouflaged image is reshaped to 352×352. We use BCE loss and IoU loss as supervision. All experiments are conducted using PyTorch.

B. Comparison with State-of-the-art Methods

To validate the effectiveness of our approach, we compare it with previous state-of-the-art Ref-COD methods following [2]. These compared methods are based on COD models, which are implemented by combining with R2CNet [2] to enable the fusion of reference features. The comparison results are presented in Table I where both image reference and text reference settings are reported. It can be observed that our method outperforms previous methods by large margins on all four metrics for the overall testing set, especially on the weighted F-measure (4.6% for image reference and 4.4% fofir text reference), indicating that our RPMA pipeline can adequately leverage the rich and fine-grained semantic knowledge of the generic segmentation network by effective reference-based feature prompting and dynamic convolution. Besides,

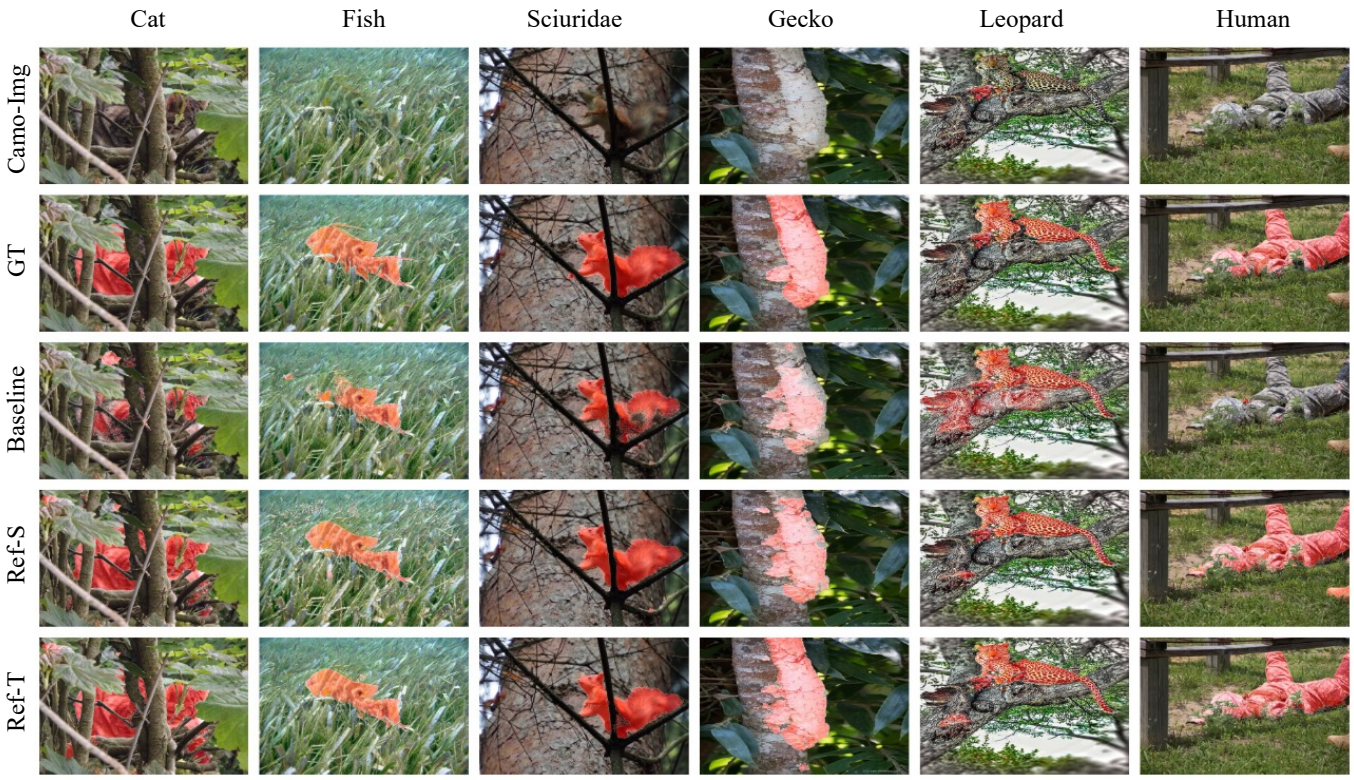

Fig. 3: Qualitative comparison of predictions between our text-model, image-model, and standard COD model (Baseline). '-S': Results with reference images. '-T': Results with reference text. The segmentation masks are shown in red.

our method achieves high performances on both single-object and multi-object testing sets, showing our method can well tackle complex camouflaged scenes.

TABLE II: Component analysis. * denotes inserting CRA before self-attention.

Method	$S_m \uparrow$	$\alpha E \uparrow$	$\omega F \uparrow$	MAE $\downarrow$
Baseline	0.843	0.911	0.749	0.028
+CRA	0.860	0.926	0.783	0.024
+CRA*	0.859	0.923	0.781	0.024
+RDC	**0.862**	**0.930**	**0.784**	**0.023**

TABLE III: Number of reference images.

Image Number	$S_m \uparrow$	$\alpha E \uparrow$	$\omega F \uparrow$	MAE $\downarrow$
1	0.859	0.926	0.781	0.024
5	0.861	0.929	0.782	0.024
10	**0.862**	**0.930**	**0.784**	**0.023**
15	0.860	0.928	0.783	0.024
20	0.859	0.926	0.782	0.023

C. Ablation Studies

Component Analysis. We conduct ablation studies to verify the effectiveness of each component. As shown in the 2nd row of Table II, our CRA brings significant performance improvement compared with the baseline where the generic segmentation model is finetuned on the Ref-COD dataset without well-designed model adaptation modules. The above results demonstrate the superiority of our CRA when adapting a generic segmentation network to the Ref-COD task. When adjusting the insertion position of CRA to before the self-attention layers, we observe a slight decrease in performance. By incorporating RDC into the segmentation decoder, further improvement is achieved, showing the effectiveness of utilizing reference information as dynamic convolution.

TABLE IV: Application stages of CRA.

Stages	$S_m \uparrow$	$\alpha E \uparrow$	$\omega F \uparrow$	MAE $\downarrow$
{S4}	0.845	0.911	0.752	0.027
{S3, S4}	0.857	0.926	0.0775	0.024
{S2, S3, S4}	0.859	0.926	0.777	0.024
{S1, S2, S3, S4}	**0.862**	**0.930**	**0.784**	**0.023**

Number of Reference Images. We also evaluate different numbers of reference images in Table III. We randomly select n images in each iteration during training to feed to the CLIP image encoder for feature extraction and the averaged feature is then utilized in CRA and RDC for model adaptation. We observe that using more reference images does not necessarily lead to better results, possibly because a larger number of

images might blur the features of the target object, making it difficult to extract discriminative image features. The optimal outcome was achieved with 10 reference images, which is adopted as the default implementation of our method.

Application Stages of CRA. We conducted ablation experiments on the application stages of CRA. In Table IV, we present the results of applying the CRA to 1-4 stages of the Transformer encoder, in the order from the 4th stage to the 1st stage. We observe that segmentation performance continuously improved with CRA's application in more stages, and the impact is more pronounced when CRA is applied in the deep stages, indicating that deep model adaptation can adjust the image features more sufficiently to align with the reference information for improved segmentation results.

D. Qualitative Results

In Fig. 3, we compare the qualitative results with the baseline model. Through the effective utilization of reference information, our method shows significant improvements in both the identification of camouflaged objects and the segmentation accuracy. For instance, in the last column, our method successfully identifies a person hidden in the bushes, which the baseline fails to discern. In the 5th column, while the baseline segments part of the tree trunk with the leopard as the foreground region, our method accurately segments the leopard alone. These results effectively demonstrate the superiority of our method for the Ref-COD task.

V. Conclusion

In this paper, we propose a Reference Prompted Model Adaptation (RPMA) pipeline to tackle the referring camouflaged object detection task. Given that the previous method suffers from limited feature learning and reference use, our RPMA leverages the rich and fine-grained semantic knowledge from a generic segmentation network to improve the Ref-COD model's capability. A Cross Reference Adapter (CRA) is designed to prompt reference-relevant camouflaged image features by densely integrating reference information into the segmentation network. Additionally, we also devise a Reference-guided Dynamic Convolution (RDC) for more accurate foreground-background segmentation with reference-generated kernels. Extensive experiments on the Ref-COD benchmark demonstrate that our method achieves new state-of-the-art performance.

Acknowledgement. This research is supported by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (NO.62122010), Zhejiang Provincial Natural Science Foundation of China under Grant No.LDT23F02022F02, Key Research and Development Program of Zhejiang Province under Grant 2022C01082.

References

[1] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, 2020.

[2] X. Zhang, B. Yin, Z. Lin, Q. Hou, D.-P. Fan, and M.-M. Cheng, "Referring camouflaged object detection," *arXiv preprint arXiv:2306.07532*, 2023.

[3] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *TPAMI*, 2021.

[4] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *CVPR*, 2022.

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

[6] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen, "Mirrornet: Bio-inspired camouflaged object segmentation," *Access*, 2021.

[7] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *CVPR*, 2021.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[9] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *CVPR*, 2021.

[10] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, "Feature shrinkage pyramid for camouflaged object detection with transformers," in *CVPR*, 2023.

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[12] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia *et al.*, "GLM-130B: an open bilingual pre-trained model," *arXiv preprint arXiv:2210.02414*, 2022.

[13] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*, 2019.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[15] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *CVPR*, 2021.

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *NeurIPS*, 2021.

[18] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *ECCV*, 2022.

[19] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "AdaptFormer: Adapting vision transformers for scalable visual recognition," *NeurIPS*, 2022.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[21] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, "Preynet: Preying on camouflaged objects," in *ACM MM*, 2022.

[22] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, "Deep gradient learning for efficient camouflaged object detection," *MIR*, 2023.

[23] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin, "I can find you! boundary-guided separated attention network for camouflaged object detection," in *AAAI*, 2022.

[24] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," *IJCAI*, 2022.

[25] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017.

[26] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

[27] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *CVPR*, 2014.