Supplementary Material for
Camouflaged Instance Segmentation via Explicit De-camouflaging

In the supplementary material, we first provide more details and experiments about fusion layer and reference attention mechanism. Then, we provide more visualization of Fourier spectrum and instance segmentation predictions.

1. Additional details

Here, we provide more details about fusion layer in the pixel-level camouflage decoupling module. Then we describe the additional components of reference attention in the instance-level camouflage suppression module and conduct additional experiments.

1.1. Fusion Layer

To acquire fine-grained target information for more accurate segmentation, we use multi-scale features $\{\mathbf{F}_i\}_{i=2}^{5}$ from different stage of the backbone, and generally they have strides of $\{2^i\}_{i=2}^{5}$ pixels with respect to input image x. To reduce the computational cost, we feed the last three feature maps $(\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5)$ to the difference attention mechanism to obtain de-camouflaged pixel features $\mathbf{F}_3^d, \mathbf{F}_4^d, \mathbf{F}_5^d$, respectively. In the fusion layer, following FPN [3], we gradually upsample the features in a top-down pathway from lowest-resolution features, meanwhile aggregate features with the same resolution by lateral connections, as shown in Figure 1. In this way, we can fusion the hierarchical features ($\mathbf{F}_5^d, \mathbf{F}_4^d, \mathbf{F}_3^d$ and $\mathbf{F}_2$) to generate the high-resolution pixel-level features $\mathbf{E}$ at 1/4 scale of input image, which is used for final mask prediction.

1.2. Reference Attention Mechanism

Despite the effectiveness of the reference attention mechanism, employing it only once is not enough to learn instance prototypes well. As shown in Figure 2, in our approach, we equip the reference attention with self-attention [6] and feed-forward network (FFN) to form a decoder layer, and stack multiple layers to implement continuous interactions between instance prototypes and pixel features.

To explore effect of the number of decoder layers T, we conduct experiments using DCNet (ResNet-50 [2] backbone) with different number of decoder layers on COD10K [1] and NC4K [5] in terms of AP metric. As shown in Table 1, when $T = 1$, the network cannot converge. When $T = 3$, the performance drops much compared to case when $T = 6$ due to insufficient interactions. When $T = 9$, the best performance on COD10K is achieved by a slight improvement (*i.e.*, 0.2%), however, with large increase of parameters (*i.e.*, 4.9M). Thus, we choose $T = 6$, considering the trade-off of performance and the computational cost.

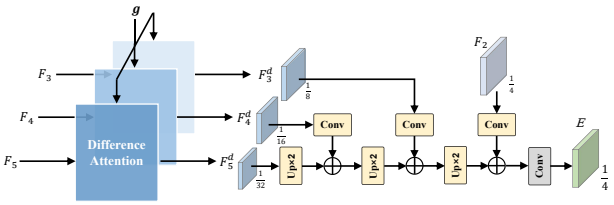

Figure 1. Illustration of the fusion layer, where $\boldsymbol{g}$ is camouflaged characteristics and $\mathbf{E}$ is the generated high-resolution pixel-level features.

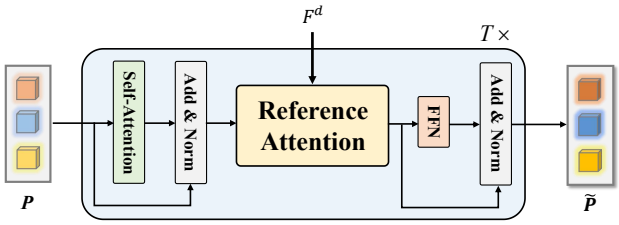

Figure 2. Illustration of the decoder layers, which contain the proposed reference attention mechanism, where $\tilde{\mathbf{P}}$ is updated instance prototypes.

Table 1. Comparisons of performance with different number of decoder layers on COD10K [1] and NC4K [5] in terms of AP metric.

T	COD10K	NC4K	Params(M)
1	-	-	45.1
3	35.2	40.6	48.4
6	45.3	**52.8**	53.4
9	**45.5**	52.3	58.3

2. Additional Visualization

In this section, we show the reconstructed image of amplitude and phase from Fourier spectrum, respectively. Then we represent more instance segmentation results of our DCNet.

2.1. Fourier Spectrum

As shown in Figure 3, we show some camouflage images and corresponding images reconstructed by amplitude and phase. In fact, the types of camouflage scenes are limited, such as *sand*, *jungle*, *lawn*, *etc*., and the texture or patterns of these scenes are relatively monotonous. Therefore, the low-level statistical information involved in the amplitude can well represent the camouflage characteristics. On the contrary, although phase images contain semantic information, they also have abundant pixel-level noise (in the background area), which is not conducive to de-camouflaging.

2.2. Instance Segmentation Prediction

We visualize several instance segmentation predictions of our DCNet model with Swin-S [4] backbone in Figure 4.

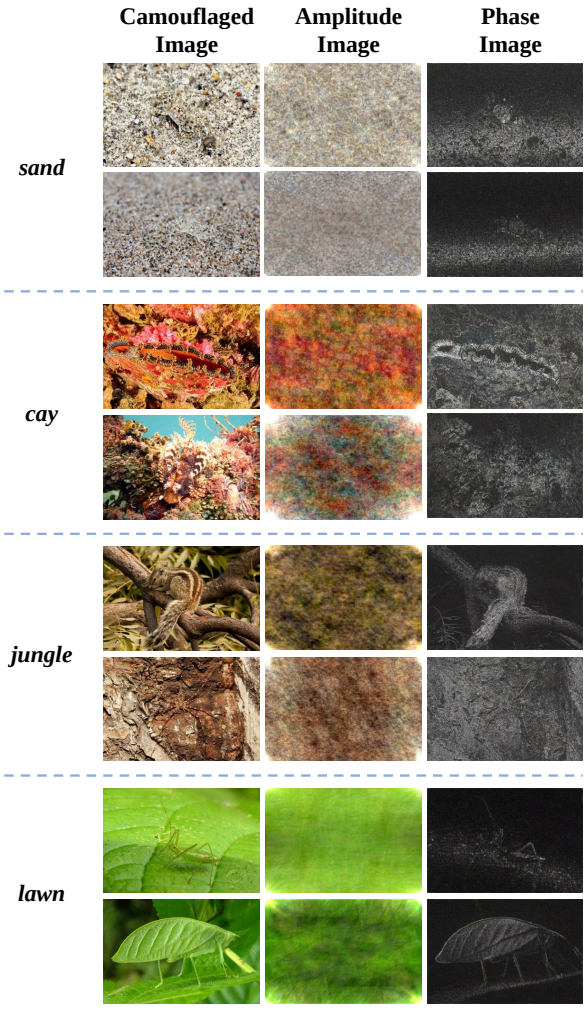

Figure 3. Visualization of the amplitude images and phase images.

References

[1] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11601, 2021. 1

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

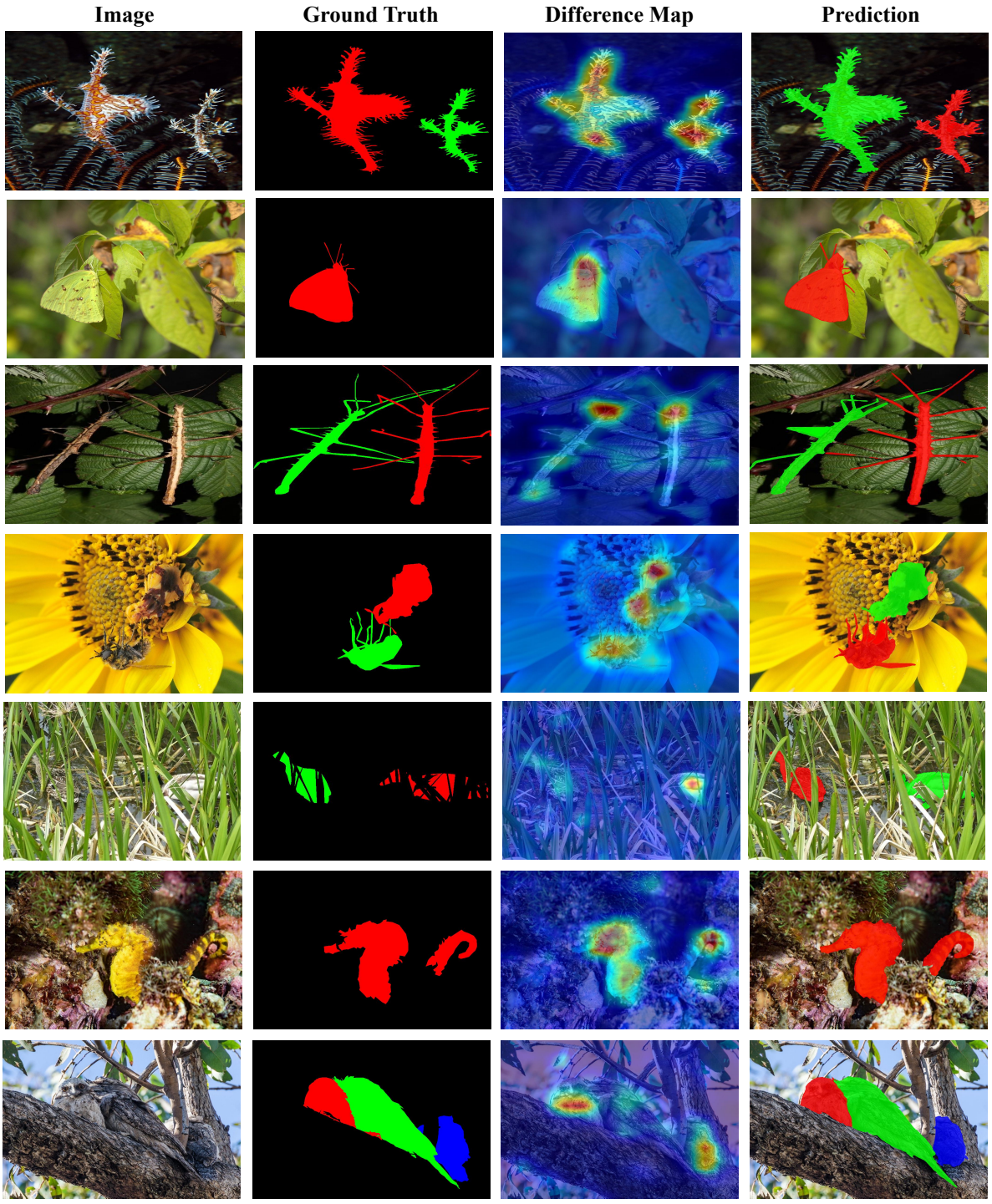

Figure 4. Visualization of DCNet with Swin-S [4] backbone, where different colors refer to different instances.