Revealing Hidden Context in Camouflage Instance Segmentation

Thanh-Hai Phung and Hong-Han Shuai

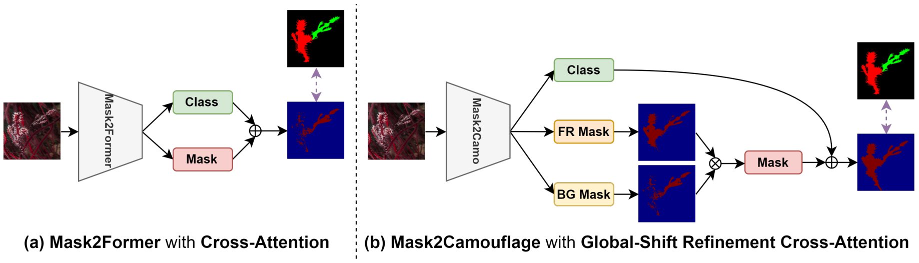
National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
{haipt.ee08,hhshuai}@nycu.edu.tw

Abstract. Predicting the instance-level masks of objects hidden in complex contexts is the goal of Camouflage Instance Segmentation (CIS), a task complicated by the striking similarities between camouflaged objects and their backgrounds. The diverse appearances of camouflage objects, including varying angles, partial visibilities, and ambiguous morphologies, further heighten this challenge. Prior works considered classifying pixels in a high uncertainty area without considering their contextual semantics, leading to numerous false positives. We proposed a novel method called Mask2Camouflage, which simultaneously enhances the modeling of contextual features and refines instance-level predicted maps. Mask2Camouflage leverages multi-scale features to integrate the extracted features from the backbone. Then, a Global Refinement Cross-Attention Module (GCA) is introduced to complement the foreground mask and background mask each other to reduce the false positive. Furthermore, by simulating a global shift clustering process, we present the Global-Shift Multi-Head Self-Attention (GSA), which enables the object query to capture not only information from earlier features but also their structural concepts, thereby reducing intra-class issues in the camouflage object detection task when validated with evaluated data. Compared with 15 state-of-the-art approaches, our Mask2Camouflage significantly improves the performance of camouflage instance segmentation. Our code is available at https://github.com/underlmao/Mask2Camouflage.

Keywords: Camouflage Instance Segmentation · Global-to-Local Refinement

1 Introduction

Camouflage is a strategy used by various organisms to evade predators or stealthily approach prey, often involving changes in pigmentation or illuminative adaptations to blend with their environment, a phenomenon well-documented in evolutionary biology [10, 24]. Humans have also leveraged camouflage to remain undetected in artistic and military contexts, offering a strategic advantage [9]. In computer vision, Camouflaged Object Detection (COD) and Camouflaged Instance Segmentation (CIS) have gained popularity, applied in areas ranging from art and medical imaging to search and rescue operations [9,13,15,48]. While



(a) **Mask2Former** with **Cross-Attention**

(b) **Mask2Camouflage** with **Global-Shift Refinement Cross-Attention**

Fig. 1: Comparison between Mask2Former [6] and our proposed Mask2Camouflage.

COD focuses on identifying camouflage objects within an image, CIS further delineates instance-level masks, crucial for handling complex scenes and detailed object identification.

In this paper, we focus on the Camouflage Instance Segmentation (CIS) problem, which presents significant challenges due to the diversity of concealment tactics employed by various object classes within the same scene. The complexity of CIS stems from the need to distinguish between multiple instances, each employing unique camouflage strategies. Traditional approaches to Camouflaged Instance Segmentation (CIS), such as those relying on established instance segmentation frameworks [25] or requiring hand-designed Non-Maximum Suppression (NMS) for effective segmentation of camouflage objects [38], often face limitations of adaptability and precision in diverse environments. Recently, [12] uses Transformers with multiple query representations to learn and share mask and boundary queries. Moreover, [32] proposes a de-camouflaging mechanism, employing Fourier Transformations to expose hidden characteristics. However, these methods still struggle with precise pixel-level detection and segmentation of camouflage objects, which vary greatly in scale, class, and appearance.

Figure 1(a) illustrates how background surroundings can significantly disrupt the performance of current CIS models, making it challenging to distinguish camouflage objects from their environments and accurately predict unclear regions. To effectively address these challenges, our work distills the CIS problem into two main questions: 1) *How can we precisely segment pixels of camouflage objects, particularly when these objects display subtle appearances and vary in size?* 2) *How can we enhance the detection of camouflage instances while minimizing the impact of background noise?* Inspired by human perceptual capabilities, we aim to propose a multi-scale modeling approach that mimics how humans assess changes in shape and appearance across different scales to identify camouflages or obscure instances in a scene accurately. By replicating this human-like behavior at the pixel level, we anticipate significant improvements in the accuracy and reliability of camouflage object detection.

Specifically, we introduce *Mask2Camouflage*, a masked-based network engineered to enhance the performance of camouflage instance segmentation significantly. To mitigate the risks associated with cascading feature reuse, a prevalent issue in this domain, our strategy involves the construction of a parallel unified framework that incorporates both global and local cues. Drawing inspiration from the human visual system, which is adept at processing images rich in information content, we classify high-resolution input images into two distinct categories: distant views that capture global information and close-up views that focus on local details. Our approach begins with developing a Feature Aggregation Module (FAM) designed to simultaneously process global semantic features and local details, leveraging the unique characteristics of the encoder. This method ensures the precise capture of objects' camouflage at the pixel level. By employing this architecture, we effectively circumvent the issue of feature hybridization that has plagued previous methods, thereby enhancing the robustness of our model.

Our transformer decoder further incorporates a Global Refinement Cross-Attention module (GCA) to refine segmentation accuracy in regions with ambiguous boundaries. This module is specifically tailored to enhance the delineation between foreground instances and their backgrounds, improving the clarity and precision of the segmentation process. Additionally, the transformer decoder includes a Global-Shift Multi-Head Self-Attention (GSA) module, which adopts an information-shift clustering technique. This innovative module dynamically adjusts feature weights based on a rich dataset of inter- and intra-class camouflage information, facilitating more effective instance segmentation. Overall, our framework boosts segmentation accuracy and provides a robust and adaptable solution for addressing the complex challenges associated with camouflaged instance segmentation. As such, Mask2Camouflage effectively handles a wide range of scenarios, making it a versatile tool in the field of computer vision. Our main contributions can be summarized as follows:

- We identify the challenge of detecting camouflage instances and propose Mask2Camouflage. This innovative network harnesses mixed-scale features to adapt to various pixel-level camouflages adeptly and is designed to perform instance-level segmentation by effectively integrating global and local information. This approach enables the model to identify and segment camouflage objects across diverse scenarios accurately.
- We propose a Global Refinement Cross-Attention (GCA) module that enhances the distinction between foreground and background areas. This module uniquely focuses on each region, leveraging intra- and inter-class information shifts from the training phase to evaluation, improving the visibility and differentiation of camouflage instances.
- Experimental results demonstrate that our method outperforms the state-of-the-art camouflage instance segmentation approaches by at least **3.2%** in terms of AP_{75} on both COD10K and NC4K datasets.

2 Related Works

2.1 Camouflage Object Detection

COD aims to identify objects that blend into their backgrounds, a challenge that has intrigued both biologists and artists for decades [17, 40, 46]. Early research in this field relied on handcrafted features such as texture, boundaries, and intensity to distinguish camouflage objects [21, 30, 36, 39]. Recently, the advent of deep learning (DL) has brought significant advancements to COD, enabling more effective end-to-end learning approaches [13, 14]. Le *et al.* [26] introduced a binary classification predictor as an auxiliary task to enhance the primary task of camouflage object segmentation. Similarly, Lv *et al.* [33] developed a method that simultaneously localizes and segments camouflage objects while ranking their significance. Zhai *et al.* [47] proposed a differential model-based graph learning approach to simultaneously learn the mutual feature connections between the edge and region of an object in graph space. Bio-inspired methods have also emerged, utilizing multi-scale features from single or multiple views to enhance model performance [14, 34, 37, 49]. For instance, Pang *et al.* [37] argued that single-view input is insufficient for accurate camouflage detection, advocating for multi-view approaches. Zheng *et al.* [49] further exploited visual perception knowledge and semantic cues by aggregating complementary information from multiple views to improve detection accuracy. Inspired by [37, 49], which utilize multi-scale features to detect camouflage objects accurately, we introduce the Mask2Camouflage framework; our methods enhance the camouflage instance segmentation by reinforcing object mask queries with rich object characteristic features obtained through multi-view multi-scale information.

2.2 General Instance Segmentation

Instance segmentation is a complex task that predicts pixel-level and instance-level masks. Current methods can be categorized into two-stage and one-stage approaches [4, 6–8, 16, 18, 19, 23, 41]. In particular, early instance segmentation techniques primarily follow a two-stage process: first, generating ROIs by bounding boxes for object localization, and then refining instance masks within these boxes [3, 5, 19, 22]. Although the two-stage approaches perform well, these works are barely applicable for near-real-time applications due to their long inference times. Recently, the one-stage approaches have achieved similar performance as two-stage approaches while maintaining a faster inference times based on their simplified detection pipelines [1, 2, 4, 41, 43, 44]. For example, their works typically generate non-local prototype masks and predict a set of mask coefficients by grouping per-pixel embeddings into different instances within an input image. SOLO [43] relies on the classification and mask branches to perform instance prediction masks, eliminating the region proposal network and using a grid-based positive and negative sample allocation technique, while SOLOv2 [44] directly decouples the original mask prediction into kernel learning and feature learning to generate final instance segmentation results. Currently, the transformer-based methods [6, 7] utilize instance-specific prototypes that continuously interact with pixel features through attention mechanisms, attaining leading performance in instance segmentation. MaskFormer [7] proposed the cross-attention mechanism to learn per-pixel classification, bypassing the slow inference time after acquiring the region proposal by the FPN. Mask2Former [6] employs a transformer decoder

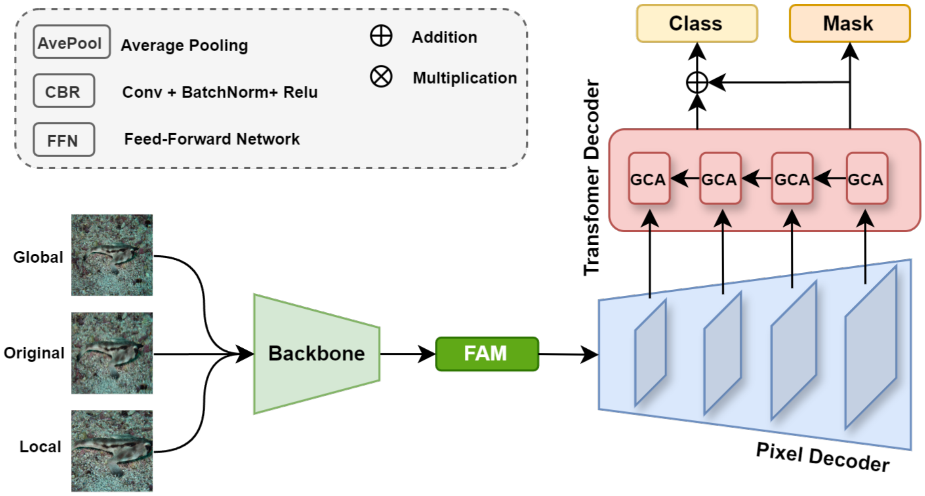


Fig. 2: The overall architecture of our Mask2Camouflage is based on a backbone consisting of four shared feature layers. From the features extracted by this backbone, we first reduce the channel size and use the Feature Aggregation Module (FAM) to identify object locations through a cascading approach. Next, we extract low-level features using the Pixel Decoder to enhance texture representation with high-resolution, per-pixel embeddings for more accurate instance segmentation. Finally, the foreground and background masks are combined and refined through the Transformer Decoder, predicting the final refined segmentation.

with masked attention to learn object queries and localize features within predicted mask regions. Unfortunately, the camouflage objects are highly similar to the background, so the general instance segmentation methods can not directly apply to the camouflage instance segmentation task.

2.3 Camouflage Instance Segmentation

Compared with Camouflage Object Detection, Camouflage Instance Segmentation remains an under-explored area. OSFormer [38] is the pioneering one-stage framework for camouflage instance segmentation. It tackles the camouflage instance segmentation task by utilizing a localized sensing transformer module with a coarse-to-fine refinement strategy. Unlike OSFormer, UQFormer [12] sees the segmentation of camouflage instances from the standpoint of query learning by fully integrating and interacting between the queries of the boundary of the object and the region of interest in the object to improve the representation of the query features. In addition to these approaches, DCNet [32] contends that a single perspective on input images is insufficient for handling CIS tasks. To address this, they introduce a dependable Fourier transform based on reference points to measure similarity accurately and robustly deceptive backgrounds for camouflage characteristic detection. However, although current works can successfully target the object query during training, it still can not handle the intra-class variation due to the camouflage dataset distribution shift during inference. In this paper, we handle the CIS from the perspectives of both object query learning and feature enhancement by viewing the input images in multi-scale to strengthen the coarse feature, then a global refinement masked attention is proposed to refine the coarse feature map into finer instance makes by complement object queries both in the foreground and background masks.

3 Mask2Camouflage

3.1 Overall Architecture

Figure 2 illustrates the architecture of Mask2Camouflage, which consists of three modules: i) the backbone for extracting features, ii) the pixel decoder that interpolates the extracted feature from the backbone to produce the high-resolution feature per pixel embeddings, and iii) the transformer decoder to associate the predicted maps with class score embedding to provide the final instance masks. Specifically, the image features $F_i \in \mathbb{R}^{C\times \frac{H}{2^{(i+1)}} \times \frac{W}{2^{(i+1)}}}$ are extracted from an RGB image input $I \in \mathbb{D}^{3\times H\times W}$ by a backbone encoder (*e.g.,* ResNet-50 [20]), where $i \in [1, 2, 3, 4]$ and C, H, and W respectively represent the channel, height, and width. We feed these three sets of feature maps from the multi-scale inputs into the Feature Aggregation Module (FAM) to aggregate the contextual information. These features are then projected to a 256-channel representation using a 1×1 convolutional layer. Next, the aggregated features are fed into the pixel decoder, aided by the Feature Pyramid Network (FPN) [27], to extract fine-grained target information and generate high-resolution pixel-level features for more precise instance segmentation. Finally, the decoder utilizes these high-resolution pixel-level feature maps to update the feature maps by refining the foreground and background masks associated with the object queries in global and local contexts, predicting the instance mask at each decoder layer. The final instance mask $M \in \mathbb{R}^{N\times(H\times W)}$ obtained by multiplying the mask embeddings with their per-pixel embeddings. These masks are trained with focal loss [28] and dice loss [35] for the class masks and cross-entropy loss with the class scores.

3.2 Features Aggregation Module

To acquire fine-grained target information for more accurate segmentation, we screen through different scale features to combine scale-specific information after getting multi-scale features from different views. Before scale integration, the global features F_i^{glo} and the local features F_i^{loc}, generated by scaling factors of 1.5 and 0.5 from the input images, are first resized to be consistent with the original scale resolution. Specifically, a hybrid structure consisting of *Max-pooling* and *Average-pooling* is employed as a down-sampling function to preserve efficacy and diversity of responses for camouflage object characteristics in high-resolution features. The designs aim to selectively aggregate the scale-specific information to explore subtle but critical semantic cues at different scales.

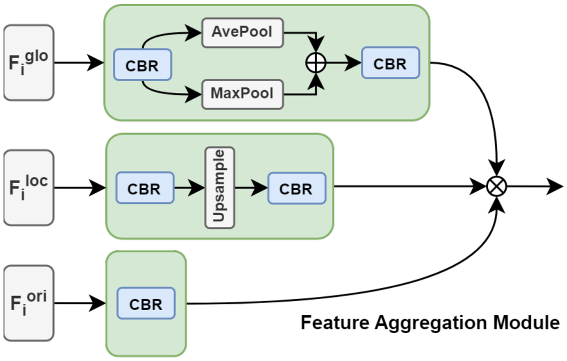


Fig. 3: Illustration of the Feature Aggregation Module (FAM).

First, for F_i^{glo}, a parallel structure of *Max-pooling* and *Average-pooling* is utilized to obtain the global camouflage characteristic as follows:

$$F_{glo}^{out} = CBR(MaxPool(CBR(F_{glo})) \oplus AvePool(CBR(F_{glo}))), \tag{1}$$

where $\oplus$ represents addition function, and CBR denotes a sequence of $Conv + BatchNorm + Relu$ to normalize output features layer after the computation. Then, for the F_i^{loc}, a bi-linear interpolation is applied to up-sample with the original feature size as follows:

$$F_{loc}^{out} = CBR(Interpolate(CBR(F_{loc}))). \tag{2}$$

Finally, the features with different scales are concatenated as follows:

$$F^{out} = concat(F_{glo}^{out}, F_{ori}, F_{loc}^{out}). \tag{3}$$

3.3 Global Refinement Cross-Attention Module

A key factor in achieving state-of-the-art segmentation results with Mask2Former [6] is replacing the cross-attention (CA) layer in the transformer decoder with masked attention (MA). Masked attention targets only the pixels within the foreground region of the predicted mask for each query, operating on the premise that local features are adequate for updating the query object features. The output of the T-th masked-attention layer can be expressed as follows:

$$softmax(M_l^F \oplus QK^T) \otimes V \oplus X_{in}, \tag{4}$$

where $X_{in} \in \mathbb{R}^{N \times C}$ is the N C-dim query features from the previous decoder layer. The input queries $Q \in \mathbb{R}^{N \times C}$ are obtained by linearly transforming the query features with a learnable transformation. In contrast, the keys and values $K,\ V$ are the image features under learnable linear transformation $w_k(.)$ and

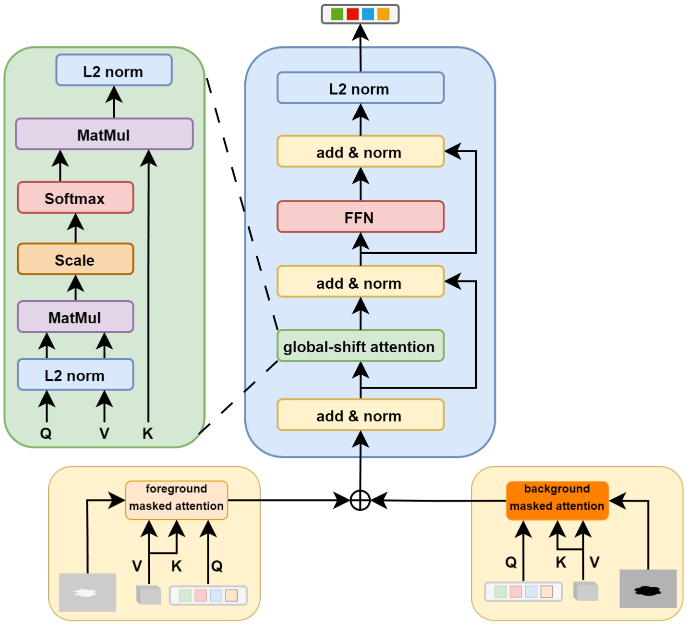


Fig. 4: The overall framework of our Global Refinement Cross-Attention Module (GCA) is built upon the cross-attention module as in [6]. We further introduce our foreground and background-masked attention to refine the high-resolution pixel-level feature from the Pixel Decoder. Finally, we proposed the Global-Shift Multi-Head Self-Attention (GSA), which adopts an information-shift clustering technique to facilitate more effective instance segmentation.

$w_v()$. Finally, M_t^F is the predicted foreground attention mask that at each pixel location *(i,j)* as follows:

$$M_t^F(i,j) = \begin{cases} 0 & \text{if } M_{t-1}(i,j) \geq 0.5, \\ -\infty & otherwise, \end{cases} \tag{5}$$

where M_{t-1} is the output mask of the previous layer. By focusing solely on foreground objects, masked attention enables faster convergence and improved instance segmentation performance compared to cross-attention. However, focusing only on the foreground region constitutes a challenge for camouflage segmentation, as camouflage objects can also appear in background regions. Omitting background information can result in failure cases where camouflage objects in the background are completely overlooked. To ameliorate camouflage detection in the high uncertainty areas, we extend the masked attention with an additional term focusing on the background region.

$$X_{out} = softmax(M_t^F \oplus QK^T) \otimes V \oplus softmax(M_t^B \oplus QK^T) \otimes V \oplus X_{in}, \tag{6}$$

where M_t^B is the additional background attention mask that complements the foreground object mask M_t^F, and it is defined at the pixel coordinates *(i,j)* as:

$$M_t^B(i,j) = \begin{cases} 0 & \text{if } M_{t-1}(i,j) < 0.5, \\ -\infty & \quad otherwise. \end{cases} \tag{7}$$

The global masked attention in Equation (7) differs from the masked attention by additionally attending to the background mask region. Yet, it retains the benefits of faster convergence w.r.t. the cross-attention.

3.4 Global-Shift Multi-Head Self-Attention

Here, we introduced our Global-Shift Multi-Head Self-Attention (GSA) and its difference compared to the Scaled Dot-Product Attention (SA) [42]. For the SA, the formula is defined as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V = \frac{exp\left(\frac{QK^T}{\sqrt{d_k}}\right)}{C_n} \times V, \tag{8}$$

where $Q \in \mathbb{R}^{N \times D_k}$, $K \in \mathbb{R}^{M \times D_k}$, and $V \in \mathbb{R}^{M \times D_v}$ denotes the N D_k-dim query vectors, M D_k-dim key vectors, and M D_v-dim values vectors, respectively. This function computes the affinity between the query Q and key K, determining the attention weights assigned to the values V to produce the weighted sum of the values. However, traditional segmentation frameworks that use masked attention often struggle with their reliance on local features for updating queries, depending heavily on multi-head self-attention (MHSA) for context gathering. Our GSA module addresses this challenge by introducing object queries as cluster centers within the feature embedding space of each decoder layer. Utilizing the Von Mises-Fisher clustering algorithm [45], these cluster centers are refined by employing the attention mask to guide cross-attention towards the local regions surrounding them in each iteration. When the object queries converge to their maxima, they are transformed into mask embeddings. These mask embeddings then generate pixel similarities through multiplication with pixel embeddings, forming the final object masks from pixels exhibiting positive similarities. The update for the cluster center is performed as follows:

$$\omega_{t+1} = \frac{\sum_i^N \rho_i \times exp(\gamma \omega_t^T \rho_i)}{\left\| \sum_i^N \rho_i \times exp(\gamma \omega_t^T \rho_i) \right\|}, \tag{9}$$

where ρ_i is the i-th sampling point from a dataset, the density function is defined as $P(\rho;\omega,\gamma) = C_d(\gamma) exp(\gamma \omega^T \rho)$, $C_d(\gamma)$ is a normalization constant, ω is a unit vector representing the mean direction of the distribution and γ is the concentration parameter that controls the concentration of the distribution around the mean direction. The SA and the updated clustering algorithm in Equation (8) and Equation (9) represent the similarity that scaled by the factor

(γ and $\frac{1}{\sqrt{d_k}}$) before normalized. By incorporating this factor into the scaled dot-product attention, we can discern the differences between various sampling points provided as object queries. The formula for Global-Shift Multi-Head Self-Attention can then be expressed as follows:

$$GSAttention(Q, K, V) = g(softmax(\gamma g(\tilde{Q}) \times g(\tilde{K}^T)) \times V), \tag{10}$$

where $\tilde{Q}$ and $\tilde{K}$ are the unit vectors that normalized by the L_2 function from these vectors. As demonstrated in Equation (10), the process begins by calculating the dot products of the normalized query vectors and key vectors, treating these as cosine similarities between Q and K. These similarities are then scaled by a factor of γ. Following this, the weights for the value vectors are computed using the *softmax* function. The output is subsequently obtained by normalizing the weighted sum of the value vectors.

4 Experiments

4.1 Experiment Setup

Datasets. Two datasets are used to train and evaluate the performance of the proposed approach. (1) COD10K [14] contains 5066 images, including 3040 images for training and 2026 images for evaluation. (2) NC4K [33], which comprised 4,121 images collected from the internet with full annotations. We follow the settings in previous works [12, 32, 38] to adopt the COD10K-Train set for model training and COD10K-Test and NC4K for evaluating the generalization of the trained model.

Evaluation Metrics. For the task of CIS, we utilize AP, AP_{50}, and AP_{75} as the evaluation metrics. These metrics comprehensively assess the model's performance by measuring its precision and recall under different conditions. While the AP_{50} and AP_{75} consider a prediction positive if the IoU between the predicted segmentation and the ground truth is at least 50% and 75%, respectively. AP combines metrics over multiple IoU thresholds, typically ranging from 50% to 95% in 5% increments.

Implementation Details. For fair comparisons, we follow previous works to utilize the ResNet-50 [20] pre-trained on ImageNet [11] as the encoder to extract features. We train our model with Adam optimizer [31] with batch size 2 for 90K iterations. In the transformer decode module, we followed DCNet [32] to set the number of object queries to 10 while keeping all the hyper-parameters and configurations the same as Mask2Former [6].

4.2 Comparison with State-of-the-arts

Quantitative Results. For quantitative comparison, we evaluated our method against various general instance segmentation approaches using ResNet-50 [20] as the encoder for a fair comparison. As demonstrated in Table 1, our proposed

Methods	**Years**	**COD10K-Test**			**NC4K**			**Params(M)**	**GFLOPs**
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}		
Two-stages									
Mask R-CNN [19]	$ICCV_{17}$	25.0	55.5	20.4	27.7	58.6	22.7	43.9	186.3
MS R-CNN [22]	$CVPR_{19}$	30.1	57.2	28.7	31.0	58.7	29.4	60.0	198.5
Cascade R-CNN [3]	$TPAMI_{19}$	25.3	56.1	21.3	29.5	60.8	24.8	71.7	334.1
HTC [5]	$CVPR_{19}$	28.1	56.3	25.1	29.8	59.0	26.6	76.9	331.7
One-stage									
YOLACT [2]	$ICCV_{19}$	24.3	53.3	19.7	32.1	65.3	27.9	-	-
BlendMask [4]	$CVPR_{20}$	28.7	56.3	26.4	29.4	56.7	27.2	35.8	233.8
CondInst [41]	$ECCV_{20}$	30.6	63.6	26.1	33.4	67.4	29.4	34.1	200.1
SOLOv2 [43]	$NIPS_{20}$	32.5	63.2	29.9	34.4	65.9	31.9	46.2	318.7
QueryInst [16]	$ICCV_{21}$	28.5	60.1	23.1	33.0	66.7	29.4	-	-
SOTR [18]	$ICCV_{21}$	27.9	58.7	24.1	29.3	61.0	25.6	63.1	476.7
MaskFormer-based									
MaskFormer [7]	$NIPS_{21}$	38.2	65.1	37.9	44.6	71.9	45.8	45.0	174.2
OSFormer [38]	$ECCV_{22}$	41.0	71.1	40.8	42.5	72.5	42.3	46.6	324.7
UQFormer [12]	MM_{23}	45.2	71.6	46.6	47.2	74.2	49.2	37.5	221.0
Mask2Former-based									
Mask2Former [6]	$CVPR_{22}$	39.4	67.7	38.5	45.8	73.6	47.5	43.9	241.0
DCNet [32]	$CVPR_{23}$	45.3	70.7	47.5	52.8	77.1	56.5	53.4	207.0
Ours		**46.8**	**72.5**	**49.0**	**53.8**	**77.6**	**58.3**	65.5	221.0

Table 1: Comparison of state-of-the-arts based COD10K-Test and NC4K datasets. The best and second best results are respectively highlighted in **bold** and underline, respectively, while "-" means "not available".

Mask2Camouflage consistently outperforms state-of-the-art methods by a significant margin on both datasets. On the COD10-Test dataset [14], Mask2Camouflage surpasses the previous best method, DCNet [32], by approximately **3.0%** in both AP_{50} and AP_{75} metrics, even without using additional Non-maximum Suppression (NMS) to eliminate redundant predictions, it reflects the superiority of our designed module FAM in case of feature enhancement and GCA for feature map refinement. For the NC4K dataset [33], Our model yields an accuracy of 58.3% in AP_{75}, making an obvious performance improvement by **3.2%** in AP_{75} compared to the best-performing method DCNet [32]. Since the model is trained with the COD10K-Train dataset, the high performance on NC4K indicates that our methods also have better generalization ability. In addition, our GFLOP is slightly higher than the previous state-of-the-art methods, which is a trade-off for enhancing instance map prediction with different scale features.

Qualitative Results. As shown in Figure 5, our proposed method is capable of improving the instance mask in different cases of subjects by utilizing the multi-scale extracted features from different scale cascade features at the pixel

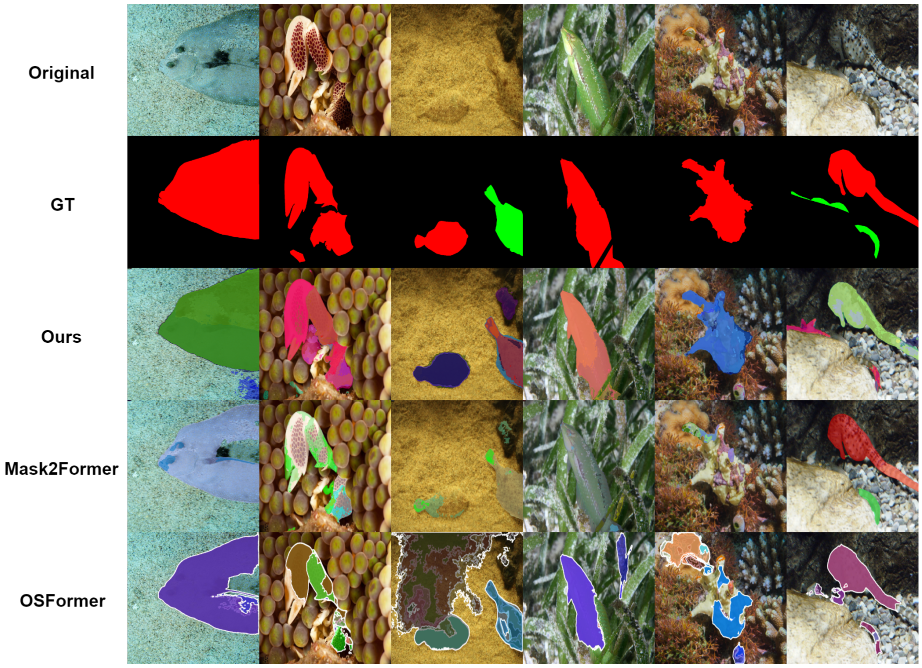


Fig. 5: Visual comparison of our prediction maps with state-of-the-art methods. Our proposed method shows natural improvement compared to other methods.

level and also able to identify the accurate locations of target camouflage objects at the instance level. Compared to the previous methods, our Mask2Camouflage performs better at whole object instances of camouflage objects, suppresses distracting background regions, and distinguishes multiple instances (the 1-5, the 2-6, and the 3-4 columns in Figure 5 from left to right). Note that the qualitative comparisons with DCNet [32] and UQFormer [12] are not provided since their predicted instance segmentation maps are unavailable.

4.3 Ablation Studies

Here, we conduct comprehensive ablation studies on COD10K and NC4K to verify the effectiveness of each module in the proposed Mask2Camouflage.

Effectiveness of the Feature Aggregation Module. We implement different scale input features integrated with the original input image to evaluate the performance in Table 2 and Figure 6. Among (1), (2), and (3), the approach that cascades the original input with the local information yields better performance, showing the significance of the local information in terms of AP_{75}. This is because local information is more important in cases where masks overlap between prediction and the ground truth.

Effectiveness of the Global Refinement Cross-Attention Module. As shown in (4) and (5) of Table 2 and Figure 6, the performance of the GCA in all

Baseline	Local	Global	GCA	GSA	COD10K-Test			NC4K		
					AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
(1)			✓	✓	45.3	70.3	47.4	51.8	75.3	55.7
(2)	✓		✓	✓	46.3	71.4	**49.3**	53.4	77.3	57.2
(3)		✓	✓	✓	45.8	70.9	47.9	52.1	75.9	56.0
(4)	✓	✓			45.0	70.5	47.1	51.9	76.1	55.8
(5)	✓	✓		✓	44.8	70.0	46.7	52.4	76.6	56.3
(6)	✓	✓	✓		46.1	71.3	48.3	53.0	76.6	57.3
Ours	✓	✓	✓	✓	**46.8**	**72.5**	49.0	**53.8**	**77.6**	**58.3**

Table 2: Comparison of different ablation experiments to reflect the performance of our Mask2Camouflage. The best performance is highlighted in **bold**.

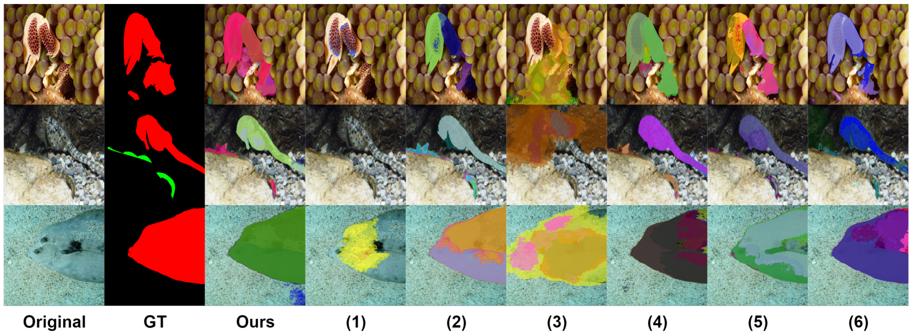


Fig. 6: Visual comparison of different ablation studies. Our proposed method shows natural improvement compared to other ablation experiments. The number in the bottom belongs to the experiments as shown in Table 2.

metrics during the two evaluated datasets showing the importance of the global refinement cross-attention module, especially the foreground and background masked complement to each other to refine the coarse pixel-level mask from the pixel decoder.

Effectiveness of the Global-Shift Multi-head Self-Attention. For the reflection of our GSA, we can observe the decrease all over the metric in both datasets COD10K-Test and NC4K (tasks (4) and (6) in Table 2 and Figure 6), showing not only how well our model deals with the data distribution shift from the COD10K-Train to the COD10K-Test but also the generalization in new circumstance for the NC4K dataset.

Effectiveness of different backbones. Table 3 shows the results using different encoders as the backbone. When using the ResNet backbone, it achieves 46.1% on the COD10K-Test and 53.0% on the NC4K dataset for the convolution depth of 50 and increases 0.7% and 0.5% when applied the convolution depth equals to 101, respectively. Furthermore, Swin-Transformer [29] showed their ca-

Methods	Backbone	COD10K-Test	NC4K
DCNet	ResNet-50	45.3	52.8
Ours		**46.8**	**53.8**
DCNet	ResNet-101	46.8	53.5
Ours		**47.7**	**54.0**
DCNet	Swin-T	50.3	56.3
Ours		**52.9**	**60.2**
DCNet	Swin-S	52.3	58.4
Ours		**54.0**	**61.1**

Table 3: Comparison of our method in the metric AP with the DCNet [32], the state-of-the-art method in different backbones.

pability when utilized as the encoder. Indeed, the AP metric improves by around **7%** in both evaluated datasets.

5 Conclusions

In this paper, we present Mask2Camouflage, a novel mask-based network designed to enhance the performance of camouflage instance segmentation. To overcome the challenges of cascading feature reuse, we propose a parallel unified framework that seamlessly integrates global and local cues. Mask2Camouflage incorporates a feature aggregation module to enrich pixel-level decoding and introduces the Global Refinement Cross-Attention Module (GCA) to improve mask refinement using foreground and background attention masks. Additionally, the Global-Shift Multi-Head Self-Attention (GSA) adjusts feature weights dynamically, optimizing instance segmentation by leveraging inter- and intra-class camouflage information. Our framework significantly improves segmentation accuracy and robustness when handling the complexities of camouflage instances, as demonstrated through extensive qualitative and quantitative evaluations. While multi-scale input learning further enhances our model's performance, it comes with a slight increase in training cost. Moving forward, we plan to develop a more streamlined and efficient approach to camouflage instance segmentation.

Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan under Grants NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3. This work was also financially supported in part (project number: 112UA10019-2) by the Co-creation Platform of the Industry-Academia Innovation School, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE) and industry partners in Taiwan.

References

1. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact++ better real-time instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 1108–1121 (2019), https://api.semanticscholar.org/CorpusID:209370542 4
2. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9157–9166 (2019) 4, 11
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6154–6162 (2018) 4, 11
4. Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y.: Blendmask: Top-down meets bottom-up for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8573–8581 (2020) 4, 11
5. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019) 4, 11
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022) 2, 4, 7, 8, 10, 11
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems **34**, 17864–17875 (2021) 4, 11
8. Cheng, T., Wang, X., Chen, S., Zhang, W., Zhang, Q., Huang, C., Zhang, Z., Liu, W.: Sparse instance activation for real-time instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4433–4442 (2022) 4
9. Chu, H.K., Hsu, W.H., Mitra, N.J., Cohen-Or, D., Wong, T.T., Lee, T.Y.: Camouflage images. ACM Trans. Graph. **29**(4), 51–1 (2010) 1
10. Darwin, C., Wallace, A.R.: Evolution by natural selection. (1958) 1
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009) 10
12. Dong, B., Pei, J., Gao, R., Xiang, T.Z., Wang, S., Xiong, H.: A unified query-based paradigm for camouflaged instance segmentation. In: Proceedings of The 31st ACM International Conference on Multimedia. pp. 2131–2138 (2023) 2, 5, 10, 11, 12
13. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1, 4
14. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2777–2787 (2020) 4, 10, 11
15. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE Transactions on Medical Imaging **39**(8), 2626–2637 (2020) 1
16. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6910–6919 (2021) 4, 11

17. Galun, Sharon, Basri, Brandt: Texture segmentation by multiscale aggregation of filter responses and shape elements. In: Proceedings Ninth IEEE International Conference on Computer Vision. pp. 716–723. IEEE (2003) 3
18. Guo, R., Niu, D., Qu, L., Li, Z.: Sotr: Segmenting objects with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7157–7166 (2021) 4, 11
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017) 4, 11
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 6, 10
21. Hou, J.Y.Y.H.W., Li, J.: Detection of the mobile object with camouflage color under dynamic background based on optical flow. Procedia Engineering **15**, 2201–2205 (2011) 4
22. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6409–6418 (2019) 4, 11
23. Ke, L., Danelljan, M., Li, X., Tai, Y.W., Tang, C.K., Yu, F.: Mask transfiner for high-quality instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4412–4421 (2022) 4
24. Kutschera, U.: Darwin–wallace principle of natural selection. Nature **453**(7191), 27–27 (2008) 1
25. Le, T.N., Cao, Y., Nguyen, T.C., Le, M.Q., Nguyen, K.D., Do, T.T., Tran, M.T., Nguyen, T.V.: Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. IEEE Transactions on Image Processing **31**, 287–300 (2021) 2
26. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Computer Vision and Image Understanding **184**, 45–56 (2019) 4
27. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017) 6
28. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017) 6
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) 13
30. Liu, Z., Huang, K., Tan, T.: Foreground object detection using top-down information based on em framework. IEEE Transactions on Image Processing **21**(9), 4204–4217 (2012) 4
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
32. Luo, N., Pan, Y., Sun, R., Zhang, T., Xiong, Z., Wu, F.: Camouflaged instance segmentation via explicit de-camouflaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17927 (2023) 2, 5, 10, 11, 12, 14
33. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11591–11601 (2021) 4, 10, 11

34. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8772–8781 (2021) 4
35. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. Ieee (2016) 6
36. Pan, Y., Chen, Y., Fu, Q., Zhang, P., Xu, X.: Study on the camouflaged target detection method based on 3d convexity. Modern Applied Science **5**(4), 152 (2011) 4
37. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2160–2170 (2022) 4
38. Pei, J., Cheng, T., Fan, D.P., Tang, H., Chen, C., Van Gool, L.: Osformer: One-stage camouflaged instance segmentation with transformers. In: European Conference on Computer Vision. pp. 19–37. Springer (2022) 2, 5, 10, 11
39. Sengottuvelan, P., Wahi, A., Shanmugam, A.: Performance of decamouflaging through exploratory image analysis. In: 2008 First International Conference on Emerging Trends in Engineering and Technology. pp. 6–10. IEEE (2008) 4
40. Song, L., Geng, W.: A new camouflage texture evaluation method based on wssim and nature image features. In: 2010 International Conference on Multimedia Technology. pp. 1–4. IEEE (2010) 3
41. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 282–298. Springer (2020) 4, 11
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017) 9
43. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: Solo: Segmenting objects by locations. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 649–665. Springer (2020) 4, 11
44. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. Advances in Neural Information Processing Systems **33**, 17721–17732 (2020) 4
45. Xiang, Y., Xie, C., Mousavian, A., Fox, D.: Learning rgb-d feature embeddings for unseen object instance segmentation. In: Conference on Robot Learning. pp. 461–470. PMLR (2021) 9
46. Xue, F., Yong, C., Xu, S., Dong, H., Luo, Y., Jia, W.: Camouflage performance analysis and evaluation framework based on features fusion. Multimedia Tools and Applications **75**(7), 4065–4082 (2016) 3
47. Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., Fan, D.P.: Mutual graph learning for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12997–13007 (2021) 4
48. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 120–130. Springer (2021) 1

49. Zheng, D., Zheng, X., Yang, L.T., Gao, Y., Zhu, C., Ruan, Y.: Mffn: Multi-view feature fusion network for camouflaged object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6232–6242 (2023) 4