Generative One-shot Camouflage Instance Segmentation

Thanh-Danh Nguyen[1,2], Vinh-Tiep Nguyen[†1,2], and Tam V. Nguyen[3]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]University of Dayton, Dayton, OH 45469, United States
{*danhnt, tiepnv*}*@uit.edu.vn, tamnguyen@udayton.edu*, [†]*corresponding author*

***Abstract*—Identifying camouflaged instances is a critical yet underexplored problem in computer vision, where traditional segmentation models often fail due to extreme visual similarity between foreground and background. While recent advances have shown promise with deep learning models, they heavily depend on large annotated datasets, which are costly and impractical to collect in camouflage scenarios. In this work, we tackle this limitation by introducing a novel framework, dubbed CAMO-GenOS, that leverages one-shot annotated samples to drive a generative process for data enrichment. Our approach integrates prompt-guided and mask-conditioned generative mechanisms to synthesize diverse, high-fidelity camouflaged instances, thereby enhancing the learning capacity of segmentation models under minimal supervision. We demonstrate the effectiveness of our CAMO-GenOS by setting up a novel state-of-the-art baseline for one-shot camouflage instance segmentation research on the challenging CAMO-FS benchmark. Code can be found at https://github.com/danhntd/CAMO-GenOS.**

***Index Terms*—One-shot Instance Segmentation, Multiple-Conditional Instance Synthesis, Camouflage Object Detection, Camouflage Instance Segmentation.**

I. Introduction

Image segmentation is a fundamental task in computer vision aiming to partition an image into semantically meaningful regions. Among its many variants, instance segmentation has gained prominence due to its capability to delineate individual object instances at the pixel level. This is crucial for applications ranging from autonomous driving to medical imaging. In recent years, a specialized subdomain has emerged, known as camouflage instance segmentation (CIS), focusing on identifying instances that intentionally blend into their surroundings. CIS has direct implications for practical tasks such as military surveillance, wildlife monitoring, and search-and-rescue tasks, where detecting subtle, low-contrast targets is essential [1]–[7].

Despite rapid progress in deep learning-based segmentation, research in camouflage instance segmentation remains an extremely challenging vision task due to both general and domain-specific issues. From a general perspective, high-performing segmentation models require large-scale, well-annotated datasets to learn class features and contextual cues effectively. However, collecting and annotating camouflaged instances is particularly difficult because the objects are intentionally designed to evade visual detection, even by human annotators. Furthermore, camouflage scenes often contain low inter-class variance and high intra-class similarity, where the instance foregrounds and contextual background share similar appearance cues, leading to frequent misclassification [1], [6]–[9]. These difficulties are magnified under severe camouflage conditions, such as in cluttered natural scenes, low-contrast environments, or with highly deformable and irregular object shapes. In such settings, typical deep learning architectures may fail to identify consistent features, resulting in poor generalization. The scarcity of domain-specific data, coupled with the lack of explicit visual boundaries, makes camouflage segmentation at the instance level one of the ill-posed problems in current computer vision research.

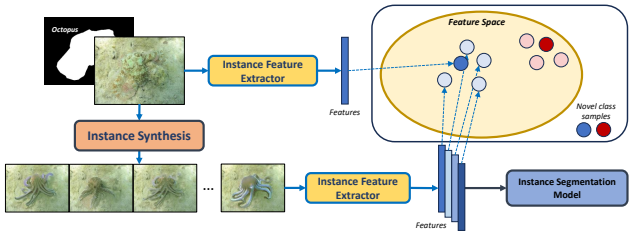

Fig. 1. Our CAMO-GenOS Concept Idea Presentation.

To address data scarcity and improve generalization under such conditions, few-shot learning has gained traction in segmentation research [10]–[12]. It enables models to recognize novel object categories using only a small number of labeled examples, often with the help of meta-learning or feature-matching paradigms. Among these, the restrictive form of one-shot learning pushes the model to learn robust representations from a single annotated instance per class [13]–[15]. This low-shot learning approach is particularly valuable for camouflage segmentation, where collecting and labeling diverse examples of camouflaged instances is not only labor-intensive but sometimes infeasible due to the rarity and variability of such patterns. However, there is currently a lack of research in one-shot learning for the camouflage domain, especially the instance-level segmentation task, which is our focus.

Furthermore, the effectiveness of one-shot learning in the camouflage domain is also limited by the insufficient intra-class diversity that a single example can provide. To overcome this limitation, recent advances in generative modeling—such as generative adversarial networks (GANs) [16], variational autoencoders (VAEs), and diffusion models [17]–[20]—offer promising avenues. These models can be utilized to synthesize realistic, high-fidelity camouflage samples conditioned on limited supervision, thereby enriching the training distribution and allowing the segmentation model to learn better class-specific representations. Prior work [21] completed this idea in a fully supervised manner on an instance-level urban scene understanding task. In this work, by coupling generative synthesis with one-shot learning, it becomes feasible to synthesize complex camouflage instances given multiple strict

conditions to expand visual diversity and ultimately bridge the generalization gap posed by minimal data availability (as illustrated in Figure 1).

To summarize, we propose a novel generative approach to one-shot camouflage instance segmentation that integrates generative modeling to alleviate data scarcity and improve segmentation quality in severe camouflage scenarios. The key contributions of our work are as follows:

- Firstly, we introduce CAMO-GenOS - the pioneer generative one-shot camouflage instance segmentation framework that exploits and focuses on learning from one-shot annotated samples to segment camouflage instances.
- Secondly, we propose a generative approach to enrich the existing one-shot camouflage sample, strictly following multiple conditions of guided prompts, referenced images, and conditional masks to boost the performance of the instance segmentation models.
- Thirdly, we demonstrate the performance of our proposed CAMO-GenOS and setup a new state-of-the-art baseline for one-shot camouflage instance segmentation research on the recent challenging CAMO-FS benchmark [7].

II. Related work

A. Image Segmentation Research

Semantic Segmentation. Early approaches to semantic segmentation primarily employed convolutional neural networks (CNNs), formulating the task as a dense per-pixel classification problem [22]. These methods, while effective at capturing local context, often struggled with long-range dependencies and global scene understanding. The introduction of transformer-based architectures has significantly advanced the field by enabling rich contextual modeling. Works such as Segmenter [23], SegFormer [24], and SARFormer [25] build upon the Vision Transformer (ViT) [26], utilizing the self-attention mechanism originally introduced in NLP [27]. These models offer improved performance in parsing complex scenes. However, despite these advances, semantic segmentation falls short in scenarios that require distinguishing between multiple instances of the same class.

Instance Segmentation. To address the limitations of semantic segmentation, instance segmentation aims to simultaneously classify and delineate each object instance at the pixel level. This enables more detailed scene interpretation, especially in visually complex scenarios with overlapping or similar objects. Traditional solutions include two-stage architectures like the typical Mask R-CNN [28], PANet [29], and HTC [30], which first localize object proposals and subsequently refine segmentation masks. More recent research emphasizes one-stage or fully end-to-end designs for improved efficiency and scalability. Models such as YOLACT [31], Mask2Former [32], and FastInst [33] represent this trend by unifying detection and segmentation within a shared architecture. Recently, OneFormer [34] introduces a unified framework for panoptic, semantic, and instance segmentation using task-specific queries and a shared encoder-decoder structure. However, such aforementioned models required training with abundant annotated data to achieve significant results, which is limited under the concept of camouflage [1], [7], [8].

B. Camouflage Research

Camouflage Instance Segmentation. Camouflage object detection and semantic segmentation has emerged as a challenging sub-domain of computer vision, focusing on identifying objects that are visually similar to their background in texture, color, or shape. Unlike standard object segmentation tasks, camouflage scenarios involve low visual contrast, misleading boundaries, and contextual ambiguity. Early research efforts largely focused on camouflaged object detection (COD) [35], then camouflage object segmentation (COS) [1], [36], [37] which aims to produce a semantic segmentation map of the camouflaged region. These methods typically adapt saliency detection pipelines and integrate multi-level feature fusion, edge guidance, and attention mechanisms to emphasize subtle differences between foreground and background.

To extend COD and COS to instance-level understanding, recent work has introduced camouflaged instance segmentation (CIS). The SINet-V2 framework [38] and the more recent CamoFormer [39] attempt to address CIS by incorporating transformer-based encoding, boundary refinement modules, and multi-scale instance decoders. Such recent work [6], [8], [9], [40], [41] also pays attention to solving CIS via multiple approaches, ranging from the original CNNs to the recent transformer architectures. However, similarly to the general instance segmentation, CIS currently address challenges such as occlusion, multiple overlapping objects, or vague contours thanks to abundant annotated training data.

Low-shot Learning in CIS. To get rid of the dependency on a large amount of annotated data, low-shot learning is particularly crucial in camouflage instance segmentation. This approach allows the model to learn from a few limited annotated samples (few-shot) or even a single sample (one-shot). In general few-shot instance segmentation, we observe the development of many works published to address the task [10]–[12], [42]–[44], one-shot instance segmentation is limited [13]. However, such aforementioned work was proposed to serve the generic domain. In the specific case of CIS, it is noticed that the pioneer work of Nguyen *et al.* [7] was proposed to address COD, and CIS built on top of iMTFA [11] and iFS-RCNN [10] and the proposals of instance triplet loss and instance memory storage [7]. In this work, we are the pioneers in resolving even the most extreme case of CIS, which is one-shot CIS, where the number of training samples is limited to one single camouflage sample, which increases the hardness of the task and is yet underexplored.

C. Multi-conditional Image Synthesis in Low-shot CIS

Conditional image synthesis has emerged as a valuable strategy to augment training datasets, both in scale and diversity, thereby enhancing the performance of instance segmentation models [21], [45], [46]. Deep generative models, including

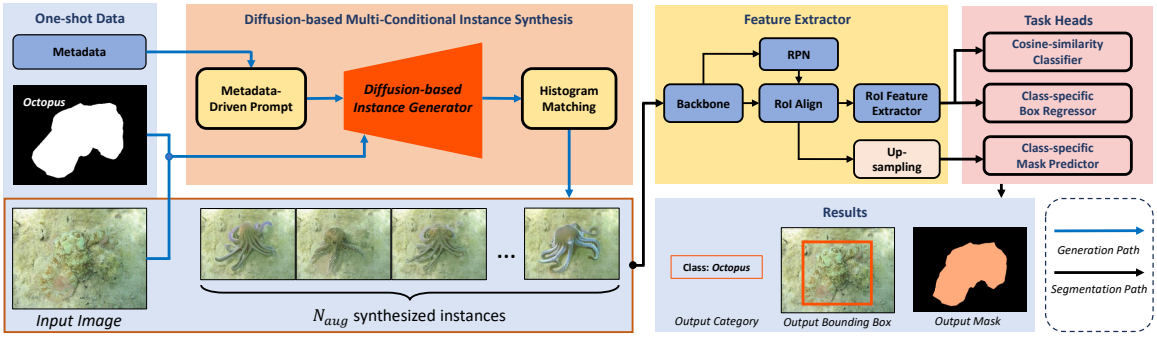

Fig. 2. Overview of our CAMO-GenOS framework leveraging generative models to enhance one-shot camouflage instance segmentation.

Generative Adversarial Networks (GANs) [16] and diffusion-based methods [47], have been extensively explored for this purpose. These techniques enable the creation of novel image samples that retain essential semantic attributes while introducing variability in appearance. Recent advancements in generative modeling, particularly with latent diffusion models [17], have shown promising results in producing high-quality, semantically meaningful images. A variety of derivative diffusion models, such as those in [17]–[19], [48]–[52], further demonstrate the effectiveness of these techniques in diverse image synthesis tasks. Inspired by the work of [21], [46], which has been a success in applying multiple conditional image synthesis to boost the instance segmentation prediction in a fully supervised manner, we propose a generative approach in low-shot CIS. Furthermore, we focus on the extreme case of one training sample (one-shot) to evaluate the effectiveness of the image generators for this intense task.

III. Proposed Method

A. Problem Definition

One-shot CIS Formulation. We define one-shot learning for the camouflage instance segmentation task as follows: Given a reference image showing one instance of a novel object category, find and segment all instances of objects belonging to the category in a separate query image of the test set, which shows an entire visual scene containing many objects [13]. Let us formulate one-shot CIS following the task of one-shot image classification [53]. The setting of *N-way K-shot* classification is considered to have a support set of K labeled samples for each of N classes [54], where $K = 1$ in one-shot configuration. Accordingly, we have one set of base classes denoted C_{base} with abundant annotated training data and one disjoint set of novel classes C_{novel} containing one sample per each class of novel data. The ultimate goal is to train a model to predict well on the novel classes $C_{test} = C_{novel}$ or on both base and novel data $C_{test} = C_{base} \cup C_{novel}$ [7]. However, going further than the classification task, one-shot CIS not only return classification results but also resolve the localization and segmentation prediction. Use an image I_q to query; one-shot CIS returns labels y_i, bounding boxes b_i, and segmentation masks M_i for all objects I_q that belong to the set of C_{test}.

B. The proposed CAMO-GenOS framework

General framework. Originating from FS-CDIS [7], which is the pioneer in the research of low-shot learning for camouflage object detection and instance segmentation, our CAMO-GenOS employs a generative approach [21], [46] to exploit and enrich the one-shot setting of the CIS task. In this work, we adopt the FS-CDIS architecture [7], which is built on top of iMTFA [11], a two-stage training and fine-tuning scheme (as illustrated in Figure 2). The initial training phase is conducted on 80 categories from the COCO dataset, producing foundational model weights, known as base weights. These base weights are subsequently refined in the second fine-tuning phase, where the model is fine-tuned to recognize novel camouflaged object categories in the CAMO-FS dataset given limited samples. In our case, this one-shot fine-tuning relies on a single annotated example per class, enriched by our proposed generative instance synthesis method.

One-shot fine-tuning. Following FS-CDIS [7] and iMTFA [11], the input query images are fed into a feature extractor F consisting of backbone B, RoI Align, RoI feature extractor modules, and a region proposal network. Our model includes three heads specifying three tasks that this scheme supports: a classification head C to predict labels, a box regression head R to locate objects, and a mask prediction head M to distinguish semantic pixels. In the first stage, the network is trained on the base classes C_{base} with abundant data of the COCO dataset. Then in the second stage, the backbone network B of the feature extractor F is frozen, and we only train the three prediction heads. Therefore, only the RoI classifier C, box regressor R, and mask predictor M are updated in the second novel fine-tuning stage. In our framework, we strictly follow the aforementioned procedure to train the proposed CAMO-GenOS framework with the number of shot $K = 1$ (one-shot).

Furthermore, we also rely on FS-CDIS [7] to extend their proposal on instance triplet loss (ITL) and instance memory storage (IMS) to generalize our proposed generative methods.

C. Diffusion-based Multi-Conditional Instance Synthesis

Inspired by InstSynth [21], [46] that succeeded in leveraging generative diffusion-based models to enhance the segmentation task under a fully supervised manner, we expand the contribution to our one-shot CIS task. Consider each novel training sample C_{novel}, the diffusion-based generative model $G(.)$ strictly takes the multiple conditions of the referenced query image I_q, ground truth mask M_q, and guided text prompt P_q to return N_{aug} samples.

TABLE I
SOTA COMPARISON OF OUR CAMO-GENOS EVALUATED ON CAMO-FS [7]. THE UTILIZED BACKBONES ARE COCO-80 FPN-RESNET-101.

		Instance Segmentation										
Method	**Synthesis Base**	**nAP**		**nAP50**	**nAP75**	**nAPs**	**nAPm**	**nAPl**	**nAR1**	**nAR10**	**nARs**	**nARm**	**nARl**
Mask-RCNN [28]		2.99		5.73	3.26	20.68	3.06	2.74	12.45	13.81	21.85	8.34	13.74
iMTFA [11]		3.66		5.37	4.09	22.42	4.35	2.01	11.30	13.58	25.97	12.96	12.53
iFS-RCNN [10]		4.27		5.98	4.75	21.57	5.71	4.87	11.70	13.51	23.35	11.75	14.28
FS-CDIS [7]		4.46	-	7.34	4.84	25.50	5.60	3.48	14.77	17.26	27.20	13.51	17.11
CAMO-GenOS (ours)	BlendedDiff [19]	4.80	+0.34	7.79	5.37	28.59	5.67	3.32	17.85	19.53	29.00	13.45	20.65
	DiffInpainting [17]	4.91	+0.45	7.84	5.47	26.54	5.06	4.02	17.18	18.72	27.70	9.75	19.23
	GLIGEN [18]	4.74	+0.28	7.53	5.31	28.10	4.79	5.28	17.65	19.38	29.33	12.29	20.42
		Object Detection											
Method	**Synthesis Base**	**nAP**		**nAP50**	**nAP75**	**nAPs**	**nAPm**	**nAPl**	**nAR1**	**nAR10**	**nARs**	**nARm**	**nARl**
Mask-RCNN [28]		3.74		6.15	4.33	26.60	5.95	4.37	16.83	18.44	27.57	11.85	19.66
iMTFA [11]		2.93		5.86	2.20	20.95	4.18	2.03	9.25	10.84	21.74	11.49	8.77
iFS-RCNN [10]		3.79		5.92	4.46	20.95	5.17	4.55	10.04	11.67	21.15	10.60	13.01
FS-CDIS [7]		3.88	-	7.71	3.21	22.38	6.40	3.32	12.66	14.85	22.67	11.89	15.36
CAMO-GenOS (ours)	BlendedDiff [19]	4.90	+1.02	8.09	4.78	29.12	7.49	3.61	17.70	19.34	29.13	15.24	20.42
	DiffInpainting [17]	5.00	+1.12	8.33	5.26	27.90	6.57	4.05	18.04	19.60	28.20	9.67	20.51
	GLIGEN [18]	4.83	+0.95	7.94	4.85	29.23	6.28	3.97	18.46	20.59	29.64	12.92	21.81

**The increased values in blue are compared to the SoTA baseline FS-CDIS [7].*

Fig. 3. Exemplary histogram matching results on the synthesized instances.

Metadata-Driven Conditional Text Prompt. We empirically define the text prompt P_q following the structure "a photo of a/an [*size*] [*meta-class*] [*instance class*]" (e.g., "a photo of a medium aquatic shrimp"). Currently, we assume the sizes, which includes *small, medium,* and *large*, cover the diversity of camouflage instances and randomly select the size in the synthesizing procedure. This structural prompt allows us to exploit the provided meta-class information in CAMO-FS, enhancing the contextual information of the prompts. Utilizing latest image-generation models of GLIGEN [18], DiffInpainting [17], and BlendedDiff [19] with Stable Diffusion XL (SDXL) [55] for $G(.)$, we create N_{aug} new instances together with the original version to train CAMO-GenOS.

Histogram Matching Post-processing. After generating N_{aug} synthesized images, we apply histogram matching to align their color distribution with that of the original image. This post-processing step ensures visual consistency between synthetic and real images, which guarantees the camouflage representation. Specifically, histogram matching adjusts the pixel intensity values of the synthesized image so that its cumulative distribution function closely matches that of the original. This is done by mapping the intensity values in the synthesized image to new values that result in a histogram resembling the reference image. As a result, the synthesized image adopts similar lighting and color tone characteristics, improving both the realism and utility of the synthetic data in training and evaluation (as illustrated in Figure 3).

IV. EXPERIMENTAL RESULTS

A. Configurations Setup

Settings. We follow the procedure published in FSOD and FSIS methods [7], [11], [56], [57]. We employed the one-shot configuration of FS-CDIS baseline [7] implemented using the Detectron2 framework [58]. The backbone is ResNet-101 [59] with Feature Pyramid Network [60]. The models are trained in two stages: base training and the novel fine-tuning stage. In the base phase, we train our model with abundant data from 80 classes with $118K$ images in the training set of the COCO dataset. The training hyperparameters of the base phase are set according to Detectron2 settings [58]. In the fine-tuning phase, we evaluate the performance of having $K = 1$ shot for each novel class. Accordingly, we train our novel detector on 47 camouflage classes of CAMO-FS [7]. The novel phase is trained with a learning rate $lr = 0.00125$ inferred from the iMTFA configurations. The balance parameters are $\alpha = 1 \times 10^{-1}$ and $\beta = 1 \times 10^{-2}$ when we inherit instance triplet loss and instance memory storage, respectively. Other training hyper-parameters of the novel phase are set following FS-CDIS [7] settings. Then, the novel models are assessed in a test set including $2,655$ images with $3,107$ instances of 47 camouflage classes to obtain the final results. Please visit [7] or [58] for more details on other parameters of both the training and testing phases. Our models are trained and tested on $4\times$ GeForce RTX 2080 Ti GPUs. Regarding the instance generation process, we strictly follow the published settings of [17]–[19] to create $N_{aug} = 4$ samples.

Dataset. We utilize CAMO-FS [7] to evaluate our proposal in one-shot CIS. To the best of our knowledge, CAMO-FS is the pioneer dataset in the camouflage domain to support multiple vision tasks and is formatted to serve few-shot learning. CAMO-FS includes $2,852$ images in total, which provides 197 images (with 235 instances) in the training set and the remaining $2,655$ images (with $3,107$ instances) in the test set. In our work, we utilize the one-shot configuration of

TABLE II
ABLATION STUDY OF OUR CAMO-GENOS ON MULTIPLE INSTANCE GENERATION-BASED METHODS EVALUATED ON CAMO-FS [7].

Method	Instance Segmentation				Object Detection			
	nAP		nAP50	nAP75	nAP		nAP50	nAP75
FS-CDIS [7]	4.46		7.34	4.84	3.88		7.71	3.21
+ ITL	4.55		7.52	4.94	3.99		7.92	3.47
+ IMS	3.94		7.44	3.64	4.01		8.05	3.44
+ Both	4.10		7.40	4.15	3.99		7.82	3.40
CAMO-GenOS (ours)								
w/ BlendedDiff [19]	4.80	+0.34	7.79	5.37	4.90	+1.02	8.09	4.78
+ ITL	5.16	+0.61	8.25	5.73	4.97	+0.98	8.54	5.08
+ IMS	4.19	+0.25	7.98	4.54	4.75	+0.74	8.38	5.16
+ Both	4.25	+0.15	7.36	4.71	4.79	+0.80	7.71	4.52
w/ DiffInpainting [17]	4.91	+0.45	7.84	5.47	5.00	+1.12	8.33	5.26
+ ITL	4.80	+0.25	7.90	5.32	4.97	+0.98	8.29	4.61
+ IMS	4.04	+0.10	7.21	4.34	4.68	+0.69	7.84	4.84
+ Both	4.29	+0.19	7.30	4.60	4.70	+0.71	7.83	4.86
w/ GLIGEN [18]	4.74	+0.28	7.53	5.31	4.83	+0.95	7.94	4.85
+ ITL	5.30	+0.75	8.26	6.02	5.23	+1.24	8.63	5.61
+ IMS	4.39	+0.45	7.28	4.86	4.52	+0.51	7.87	4.52
+ Both	4.33	+0.23	7.28	4.74	4.75	+0.76	7.62	5.52

**The increased values in blue are compared to the SoTA baseline FS-CDIS [7] with the corresponding ITL, IMS, and both of them.*

the training set for training CAMO-GenOS and the test set for the evaluation procedure.

Evaluation metrics. To report our results on detection and instance segmentation, we use the common average precision (AP) and average recall (AR). In detail, we report AP@50 and AP@75, along with AR@10. We also provide AP and AR at small, medium, and large scales of the instances to further analyze the performance of our model. For more details, please visit the official COCO dataset site at https://cocodataset.org/#detection-eval.

B. Results and Discussion

State-of-the-art CIS Comparison. To prove the effectiveness of our CAMO-GenOS framework, we utilize CAMO-FS [7] as a benchmark and compare our results with other state-of-the-art (SoTA) methods in this approach including Mask R-CNN [28], iMTFA [11], iFS-RCNN [10], and the baseline FS-CDIS [7]. We tested the results of $K = 1$ shot and leveraged the published source code of the aforementioned methods to report their results. Table I presents the evaluation of our CAMO-GenOS among SoTA methods. The details performance of the instance triplet loss and instance memory storage [7] are declared in the ablation section. Regarding the instance segmentation task, we improved over SoTA FS-CDIS [7] thanks to our generative approach based on diffusion models and achieved $4.80\%, 4.91\%, 4.74\%$ via GLIGEN [18], DiffInpainting [17], and BlendedDiff [19], respectively. The corresponding numbers in the object detection task are $4.90\%, 5.00\%, 4.83\%$, respectively. Despite the limited results, we defeated the early models on detection and instance segmentation tasks on one-shot CIS.

Ablation on the Contrastive Learning Components. Inspired by the hypothesis that contrastive learning helps distinguish foreground and background in CIS [7], we extend our proposal on the instance triplet loss (ITL) and instance memory storage (IMS) in the ablation study (reported in Table II). Consequently, our CAMO-GenOS not only improves the baseline FS-CDIS [7], but also improves all of the configurations related to the contrastive learning components. The highest value reported in nAP is 5.30% for instance segmentation, and 5.23% for object detection, which belong to CAMO-GenOS with GLIGEN [18] generation-based implementation with ITL, utilizing the backbone of COCO-80 FPN-ResNet-101. GLIGEN [18], in this case, demonstrates their effectiveness in synthesizing visually coherent and semantically meaningful contents by accurately filling in the masked regions with appropriate high-resolution instances.

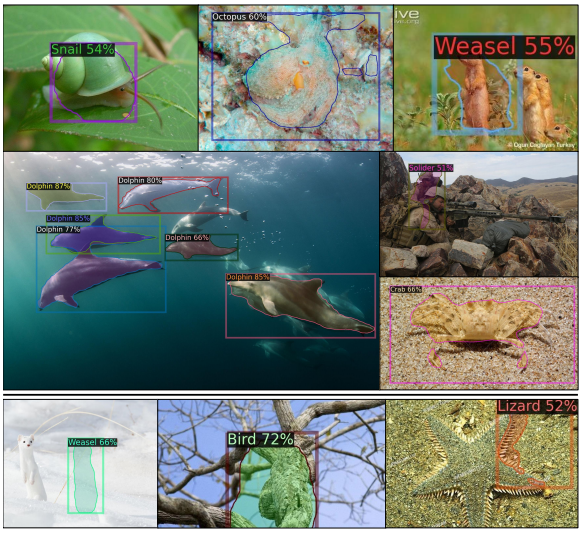

Fig. 4. Visualization results of our CAMO-GenOS on the CAMO-FS [7]. The results are visualized under the configuration of GLIGEN [18] generation-based model with ITL. Several failed cases are listed in the last row.

Discussion. We provide Figure 4 to qualitatively visualize the results of the best model. Despite the improvement in nAP, the final predictions are still modest. Besides good cases where the model correctly classified, located, and segmented the instances, there exist cases where the phenomenon of missing instances, over-segmentation, or misclassification occurs. To this end, we can conclude the effective support of the generative approach in synthesizing diverse samples for the limited one-shot camouflage instance. Yet there still remains the question of how to better distinguish the ambiguous foreground and the background of the camouflage instances.

V. CONCLUSION

In this paper, we propose CAMO-GenOS—a pioneer framework addressing one-shot camouflage instance segmentation utilizing a generative approach to enrich the training sample. The framework exploits the diffusion-based generative models to enhance the diversity of the existing annotated one-shot sample to boost the performance of the camouflage instance segmentation models. Our proposals resolve the limitation in annotated training samples via multiple conditional image generation at the instance level, specifically applied to camouflage research. To this end, we set a novel state-of-the-art baseline in one-shot CIS, serving the community interested in this field. Finally, we demonstrate the effectiveness of our proposals on the challenging CAMO-FS dataset via the extensive experiments and ablation investigations. In the future, we plan to generalize our proposals to the general domain and automate the multiple conditional image generation procedure.

VI. Acknowledgement

Thanh-Danh Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.TS.068.

References

[1] T.-N. Le, T. V. Nguyen, Z. Nie *et al.*, "Anabranch network for camouflaged object segmentation," *CVIU*, vol. 184, pp. 45–56, 2019.

[2] Z. Zhou, B. Zhang, and X. Yu, "Immune coordination deep network for hand heat trace extraction," *Infrared Physics & Technology*, vol. 127, p. 104400, 2022.

[3] X. Yu, X. Liang, Z. Zhou, B. Zhang, and H. Xue, "Deep soft threshold feature separation network for infrared handprint identity recognition and time estimation," *Infrared Physics & Technology*, p. 105223, 2024.

[4] X. Yu, X. Ye, and S. Zhang, "Floating pollutant image target extraction algorithm based on immune extremum region," *Digital Signal Processing*, vol. 123, p. 103442, 2022.

[5] X. Wang *et al.*, "Non-linear statistical image watermark detector," *Applied Intelligence*, vol. 53, no. 23, pp. 29 242–29 266, 2023.

[6] T.-D. Nguyen, D.-T. Luu, V.-T. Nguyen, and T. D. Ngo, "Ce-ost: Contour emphasis for one-stage transformer-based camouflage instance segmentation," in *MAPR*. IEEE, 2023, pp. 1–6.

[7] T.-D. Nguyen, A.-K. N. Vu, N.-D. Nguyen, V.-T. Nguyen, T. D. Ngo, T.-T. Do, M.-T. Tran, and T. V. Nguyen, "The art of camouflage: Few-shot learning for animal detection and segmentation," *IEEE Access*, 2024.

[8] T.-N. Le, Y. Cao *et al.*, "Camouflaged inst. seg. in-the-wild: Dataset, method, and benchmark suite," *IEEE TIP*, vol. 31, pp. 287–300, 2021.

[9] M.-Q. Le, M.-T. Tran, T.-N. Le, T. V. Nguyen, and T.-T. Do, "Camofa: A learnable fourier-based augmentation for camouflage segmentation," in *WACV*. IEEE, 2025, pp. 3427–3436.

[10] K. Nguyen and S. Todorovic, "ifs-rcnn: An incremental few-shot instance segmenter," in *CVPR*, 2022, pp. 7010–7019.

[11] D. A. Ganea, B. Boom, and R. Poppe, "Incremental few-shot instance segmentation," in *Proceedings of (CVPR)*, June 2021, pp. 1185–1194.

[12] B.-B. Gao, X. Chen *et al.*, "Decoupling classifier for boosting few-shot object detection and instance segmentation," *NeurIPS*, vol. 35, pp. 18 640–18 652, 2022.

[13] C. Michaelis, I. Ustyuzhaninov, M. Bethge, and A. S. Ecker, "One-shot instance segmentation," *arXiv preprint arXiv:1811.11507*, 2018.

[14] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[15] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[16] I. Goodfellow, J. Pouget-Abadie *et al.*, "Generative adversarial networks," *Com. of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[18] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023.

[19] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*, 2022.

[20] A. B. Yildirim, V. Baday, E. Erdem, A. Erdem, and A. Dundar, "Inst-inpaint: Instructing to remove objects with diffusion models," 2023.

[21] T.-D. Nguyen, B.-N. Pham *et al.*, "Instsynth: Instance-wise prompt-guided style masked conditional data synthesis for scene understanding," in *MAPR*. IEEE, 2024, pp. 1–6.

[22] B. Cheng, L.-C. Chen *et al.*, "Spgnet: Semantic prediction guidance for scene parsing," in *CVPR*, 2019.

[23] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *ICCV*, 2021.

[24] E. Xie, W. Wang *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *NeurIPS*, 2021.

[25] L. Zhang, W. Huang, and B. Fan, "Sarformer: Segmenting anything guided transformer for semantic segmentation," *Neurocomputing*, vol. 635, p. 129915, 2025.

[26] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[27] A. Vaswani, N. Shazeer *et al.*, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988.

[29] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *ICCV*, 2019, pp. 9197–9206.

[30] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018.

[31] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *ICCV*, 2019.

[32] B. Cheng, I. Misra *et al.*, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[33] J. He, P. Li, Y. Geng, and X. Xie, "Fastinst: A simple query-based model for real-time instance segmentation," in *CVPR*, 2023, pp. 23 663–23 672.

[34] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *CVPR*, 2023.

[35] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, 2020, pp. 2777–2787.

[36] H. Mei, G.-P. Ji *et al.*, "Camouflaged object segmentation with distraction mining," in *CVPR*, 2021, pp. 8772–8781.

[37] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen, "Mirrornet: Bio-inspired camouflaged object segmentation," *IEEE Access*, vol. 9, pp. 43 290–43 300, 2021.

[38] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE TPAMI*, vol. 44, no. 10, pp. 6024–6042, 2021.

[39] B. Yin, X. Zhang *et al.*, "Camoformer: Masked separable attention for camouflaged object detection," *IEEE TPAMI*, 2024.

[40] N. Luo, Y. Pan, R. Sun *et al.*, "Camouflaged instance segmentation via explicit de-camouflaging," in *CVPR*, 2023, pp. 17 918–17 927.

[41] J. Pei, T. Cheng, D.-P. Fan, H. Tang, C. Chen, and L. Van Gool, "Osformer: One-stage camouflaged instance segmentation with transformers," in *ECCV*. Springer, 2022, pp. 19–37.

[42] Y. Han, J. Zhang, Z. Xue, C. Xu, X. Shen, Y. Wang, C. Wang, Y. Liu, and X. Li, "Reference twice: A simple and unified baseline for few-shot instance segmentation," *arXiv preprint arXiv:2301.01156*, 2023.

[43] H. Wang, J. Liu, Y. Liu, S. Maji *et al.*, "Dynamic transformer for few-shot instance segmentation," in *ACMMM*, 2022, pp. 2969–2977.

[44] W. Gao, C. Shi, R. Wang, A. Cai, C. Duan, and M. Liu, "Incremental few-shot instance segmentation via feature enhancement and prototype calibration," *CVIU*, p. 104317, 2025.

[45] Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," *NeurIPS*, vol. 36, pp. 76 872–76 892, 2023.

[46] T.-D. Nguyen, V.-T. Nguyen, and T. V. Nguyen, "A generative approach at the instance-level for image segmentation under limited training data conditions (stu. abs.)," in *AAAI*, vol. 39, no. 28, 2025, pp. 29 451–29 452.

[47] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[48] A. Hertz, R. Mokady, J. Tenenbaum *et al.*, "Prompt-to-prompt image editing with cross attention control," *ICLR*, 2023.

[49] R. Mokady, A. Hertz *et al.*, "Null-text inversion for editing real images using guided diffusion models," in *CVPR*, 2023.

[50] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023, pp. 18 392–18 402.

[51] C. Meng, Y. He, Y. Song, J. Song *et al.*, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in *ICLR*, 2022.

[52] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, "Controlnet++: Improving conditional controls with efficient consistency feedback," in *ECCV*. Springer, 2024, pp. 129–147.

[53] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICMLW*, vol. 2, 2015, pp. 1–30.

[54] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," *NeurIPS*, vol. 32, 2019.

[55] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint*, 2023.

[56] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *ICCV*, 2019.

[57] X. Yan, Z. Chen *et al.*, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *ICCV*, 2019.

[58] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[60] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.