CAMO-InstSynth: Few-shot Camouflage Instance Segmentation with Multi-Conditional Background Synthesis and Generative Augmentation

Thanh-Danh Nguyen[1,2][0000−0001−6577−2122], Vinh-Tiep Nguyen[1,2*][0000−0003−4260−7874], Kunpeng Li[3][0000−0003−4291−7207], and Tam V. Nguyen[4][0000−0003−0236−7992]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Air Force Institute of Technology, Ohio, 45433, United States
[4] University of Dayton, Dayton, Ohio, 45469, United States
{danhnt,tiepnv}@uit.edu.vn, kunpeng.li@us.af.mil, tamnguyen@udayton.edu,
*corresponding author

Abstract. Camouflage object detection and instance segmentation remain a challenging frontier in computer vision due to the intrinsic high similarity between foreground objects and their background surroundings. Furthermore, the scarcity of annotated camouflage data exacerbates the difficulty of training robust models, particularly in data-sparse regimes. In this paper, we introduce CAMO-InstSynth, a novel data enhancement framework exploiting the background context understanding to address few-shot camouflage object detection and instance segmentation. Accordingly, we propose a Multi-Conditional Background Synthesis Module that utilizes diffusion models to generate diverse, high-fidelity backgrounds that maintain semantic consistency with camouflaged foregrounds. Unlike traditional augmentation techniques, which increase only the instance diversity, our method further conditions the synthesis process to simulate the camouflage environment. We validate our approach on the common CAMO-FS dataset over strong few-shot baselines. Our experiments demonstrate that CAMO-InstSynth significantly outperforms state-of-the-art methods, improving the iFS-RCNN baseline in both Segmentation and Detection tasks. Code can be found at https://github.com/danhntd/CAMO-InstSynth.

Keywords: Camouflaged Instance · Camouflaged Animal · Few-shot Learning · Object Detection · Instance Segmentation.

1 Introduction

Camouflage serves as a sophisticated biological strategy wherein animals minimize their visual salience to merge seamlessly with their habitat. The ability to autonomously perceive these concealed entities is pivotal for high-stakes applications, including search-and-rescue in complex environments [15], wildlife

population monitoring [15], and identifying tampered visual media [14]. While generic image segmentation has a long history, standard detectors often fail to distinguish camouflaged targets due to high foreground-background similarity [36]. Early research dedicated to camouflage detection utilized handcrafted low-level features; however, these methods lack generalization and typically falter in heterogeneous or cluttered scenes. Contemporary deep learning approaches have significantly advanced the field [15,2,21,20] yet they remain heavily dependent on large-scale, pixel-level annotated datasets, which poses a significant bottleneck for rare or elusive categories.

To mitigate the reliance on exhaustive annotations, Few-Shot Learning (FSL) has emerged as a promising paradigm, aiming to generalize to novel classes given only a handful of support samples [35,33,42]. However, applying FSL to camouflage presents a unique feature-contradiction paradox: while conventional FSL relies on extracting distinct prototypes to separate foreground from background, camouflaged objects are biologically evolved to suppress these exact features. The challenge has been concerned and addressed in the general domain by previous work [4,6,3,8,12,25,22]. However, under the context of few-shot learning for camouflage instance segmentation, there exists a few work [24,25,28]. The work of Nguyen *et al.* [28] introduced the CAMO-FS benchmark and demonstrated that standard few-shot baselines struggle significantly in cryptic scenarios due to the high frequency of texture overlap. While Nguyen *et al.* proposed the FS-CDIS framework to enhance instance discrimination via triplet loss, they highlighted that the scarcity of diverse environmental contexts in the support set remains a critical bottleneck. Consequently, models often overfit to the specific background patterns of the few available shots rather than learning the generalized concept of concealment, necessitating more robust data synthesis strategies.

To surmount the limitations of data scarcity and environmental monotony, leveraging generative models for data augmentation has emerged as a critical research direction [31,17,1]. Recent advances in diffusion models have enabled the synthesis of high-fidelity training samples, offering a potential remedy to the few-shot bottleneck by expanding the support set manifold. Notably, recent efforts [25,26] have primarily explored foreground synthesis, utilizing text-to-image models to generate new camouflaged instances into existing scenes. However, we argue that for cryptic detection, the specific biological texture of the organism is the most valuable signal; synthesizing the object runs the risk of generating generic patterns that lack the intricate, evolved concealment traits of real species. Addressing this, our work introduces a novel background synthesis paradigm. Instead of hallucinating the object, we preserve the real, complex camouflaged foreground anatomy and utilize multi-conditional diffusion models to generate diverse, semantically consistent environments around it. This strategy forces the detector to learn robust boundary features invariant to background shifts, directly countering the overfitting observed in the CAMO-FS benchmark [28]. To summarize, our contributions in this work are threefold:

– First, we propose a novel framework exploiting multi-conditional generative models to synthesize diverse and semantically consistent environments to

enhance the camouflage detection and instance segmentation under few-shot settings.
- Second, we propose a generative camouflage background synthesis process to augment the existing camouflage instances.
- Third, our extensive experiments on CAMO-FS benchmark have demonstrated the performance of the proposed CAMO-InstSynth over existing state-of-the-art methods in the approach.

2 Related work

2.1 Camouflage Research

Given any specified region, such as a bounding box or polygon mask, associated with an object of interest (e.g., animals or man-made objects) in an image, if that region is classified as background, the content within it can be regarded as a camouflaged object. Accordingly, a camouflaged object is defined as a collection of bounding boxes or camouflaged pixels in an image, without requiring additional information such as object count or semantic labels [15,28,26]. Although camouflaged animal-related tasks arise in many practical applications, this research area remains relatively underexplored, particularly in the context of few-shot learning, which is well-suited to scenarios involving limited data, as is common for camouflaged animals.

Camouflage Datasets Early camouflage datasets such as CamouflagedAnimals [30] and CHAMELEON [34] provided pixel-level annotations but were too small to support deep learning. Subsequent efforts introduced larger datasets, including CAMO [15] and COD [2], though these remain limited in scale or annotation richness, while MoCA [11] is restricted to bounding-box labels. To overcome these limitations, CAMO++ [13] proposed a comprehensive benchmark for camouflaged instance segmentation with rich instance-level annotations across 10 meta-categories. Building upon CAMO++, CAMO-FS [28] was recently introduced as the first benchmark for few-shot camouflage object detection and instance segmentation, featuring $2,852$ images from 47 semantic categories with carefully designed few-shot evaluation protocols.

2.2 Camouflage Semantic Segmentation

Early studies on camouflage detection, prior to the widespread adoption of deep neural networks, primarily relied on handcrafted or low-level visual features to distinguish camouflaged regions from their backgrounds. These methods exploited subtle discrepancies in external characteristics such as color, intensity, texture, shape, orientation, and edges, even when foreground textures were highly similar to the background. While effective in controlled settings, such approaches generally failed in complex scenes due to the high visual similarity between foreground and background regions, limiting their applicability to real-world camouflage scenarios.

With the availability of binary ground-truth datasets for camouflaged objects [2,15,34], recent research has largely focused on binary camouflage semantic segmentation, where camouflaged regions are separated from the background without considering object instances or semantic categories. Representative deep-learning-based methods include ANet [15], which integrates classification and segmentation branches, and SINet [2], inspired by predator hunting strategies to sequentially search and identify camouflaged targets. Subsequent works further improved feature discrimination through dual-stream architectures (e.g., MirrorNet [40]), texture-aware refinement (e.g., TINet [43]), and joint modeling of localization, segmentation, and ranking [20]. Although these methods effectively reveal the presence of camouflaged objects at the pixel level, they remain limited to binary foreground–background separation and provide no instance-level or semantic differentiation.

2.3 Camouflage Instance Segmentation

Camouflage instance segmentation aims to identify and delineate individual camouflaged objects with separate masks, posing a significantly more challenging problem due to the lack of clear object boundaries and strong background similarity. Compared with semantic-level approaches, instance-level understanding remains relatively underexplored. To address this gap, Le *et al.* [16] introduced a practical framework that integrates multiple state-of-the-art instance segmentation methods, along with a user-interactive tool to refine segmentation masks, facilitating training and evaluation on this challenging task. Recognizing that single-model predictions are often insufficient, Le *et al.* [13] further proposed Camouflage Fusion Learning (CFL), which adaptively fuses complementary strengths of different instance segmentation models by leveraging image context. These efforts represent early but important steps toward fine-grained understanding of camouflaged objects beyond binary segmentation.

2.4 Few-shot Learning for Camouflage Instance Segmentation

Few-shot learning provides a natural framework for camouflage instance segmentation, where dense annotations are expensive, and objects exhibit strong foreground-background similarity. Early works extended detection-based frameworks to the few-shot setting, with Meta R-CNN [41] pioneering joint few-shot detection and instance segmentation by adapting predictor heads from limited support annotations. Subsequent research shifted toward segmentation-centric approaches, particularly prototype-based learning, which has become central to few-shot segmentation. Representative methods include prototype learning, prototype alignment [38], cross-reference networks with mask refinement [19], and later dynamic, context-aware, and generative extensions [32,37].

However, most few-shot segmentation methods are designed for generic objects with clear boundaries, limiting their effectiveness under camouflage conditions. Recent studies have therefore begun to specifically address few-shot camouflage instance segmentation. FS-CDIS [28] introduced instance-level discrim-

ination and memory mechanisms to separate camouflaged objects from background, achieving strong performance on CAMO-FS. Follow-up work further explored frequency-domain representations [24] and generative one-shot learning [26] to enhance robustness under extreme data scarcity. Together, these efforts establish an emerging research direction tailored to the unique ambiguity of camouflage instance segmentation.

2.5 Generative Approach in Few-shot Learning

Recent advances in data synthesis have introduced diffusion models as powerful tools for segmentation data augmentation. Nguyen *et al.* [23] proposed *Dataset Diffusion*, one of the first frameworks to leverage diffusion models for generating large-scale, pixel-level annotated data for semantic segmentation, substantially reducing the need for manual labeling. In parallel, instance-level synthesis methods such as InstSynth [27] and CAMUL [29] have explored instance-aware data generation by synthesizing diverse foreground objects under varying styles and contexts. However, these approaches largely assume a clear separation between foreground and background, an assumption that breaks down in camouflage scenarios where object appearance is tightly coupled with surrounding context.

Motivated by this limitation, recent trends emphasize background-aware and context-driven synthesis, shifting augmentation from foreground manipulation to background generation. Such strategies better model the intrinsic foreground-background entanglement of camouflage, enabling more faithful simulation of real-world camouflage phenomena and offering a promising direction for improving few-shot camouflage instance segmentation.

3 CAMO-InstSynth: Multi-Conditional Background Synthesis for Few-shot Camouflage Instance Segmentation

3.1 Few-shot instance segmentation formulation

Few-shot learning typically considers two disjoint class sets: a set of base classes C_{base}, for which abundant annotated training data are available, and a set of novel classes C_{novel}, where only a few labeled samples per class are provided. The objective is to learn a model that can generalize effectively to the novel classes, either by evaluating exclusively on $C_{test} = C_{novel}$ [35] or jointly on both base and novel classes, i.e., $C_{test} = C_{base} \cup C_{novel}$ [7]. In few-shot classification, episodic training is a widely adopted strategy. Training is organized into a sequence of episodes $E_i = (I_q, S_i)$, where S_i denotes a support set composed of N classes sampled from $C_{train} = C_{base} \cup C_{novel}$, with K labeled examples per class, forming an N-way K-shot task. Given a query image I_q, the model is trained to predict its class among those defined in S_i. By repeatedly solving diverse episodic tasks, the model learns transferable representations that improve generalization to unseen classes in C_{novel}. This episodic learning paradigm has

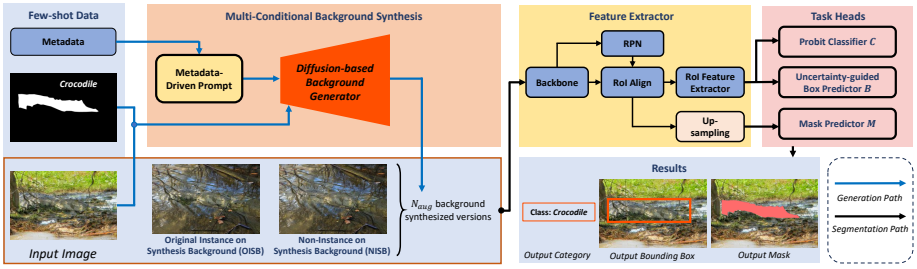

Fig. 1. Overview of the proposed CAMO-InstSynth framework.

been extended to few-shot object detection (FSOD) and few-shot instance segmentation (FSIS) [10,3,41,4,28,22,5,25]. In these settings, all object instances within a query image are treated as queries, while a single support set is shared per image rather than per instance. Compared to classification, FSIS introduces additional challenges beyond label prediction, as the model must also localize and delineate object instances. Specifically, given a query image I_q, FSIS aims to predict the class labels y_i, bounding boxes b_i, and segmentation masks M_i for all object instances belonging to the target class set C_{test}.

3.2 Framework Overview

Building on top of FS-CDIS [28], a pioneering work in few-shot camouflage object detection and instance segmentation, our proposed CAMO-InstSynth framework exploits the multi-conditional background synthesis in an instance-aware manner inspired by [27,29]. Instead of hallucinating the objects, we preserve the real, complex camouflaged foreground anatomy and generate diverse, semantically consistent environments around it. We adopt the FS-CDIS architecture [28] as our baseline, specifically its variant built upon iFS-RCNN [22], which follows a two-stage training and fine-tuning paradigm (as illustrated in Fig. 1). In the base-training phase, the model is trained on the COCO dataset [18], leveraging abundant annotations from 80 object categories to learn robust base representations. Following the configuration reported in [28], we employ base weights obtained from a ResNet-101 backbone trained on all 80 COCO classes, which were shown to deliver the strongest performance among different settings. These pretrained weights are subsequently used to initialize the fine-tuning stage, where the model is adapted to novel camouflage categories from the CAMO-FS dataset [28] under few-shot regimes with $K \in \{1, 2, 3, 5\}$ samples per class. In the novel-finetuning phase, we employ Blended Diffusion [1] to perform the background synthesis. The synthesis stage is performed as an offline augmentation procedure to guarantee the stability of the training and inference time of the utilized baselines [22].

3.3 Few-Shot Fine-Tuning

Following the design principles of FS-CDIS [28] and iFS-RCNN [22], query images are first processed by a feature extractor F, which consists of a backbone network B, a Region Proposal Network (RPN), RoIAlign, and RoI feature extraction modules. On top of these shared features, the model employs three task-specific heads for classification C, bounding box regression R, and instance mask prediction M. Specifically, we adopt the probit-based classifier for the classification head C and the uncertainty-guided bounding box regression head R from iFS-RCNN [22]. During the base-training phase on the base category set C_{base} from COCO, all network components are jointly optimized. In the subsequent fine-tuning stage on the novel camouflage categories of CAMO-FS, the backbone B is frozen to retain the learned generic representations, while only the prediction heads C, R, and M are updated to adapt the model to the few-shot setting along with the synthesis versions. Our CAMO-InstSynth framework strictly follows this two-stage training and fine-tuning protocol.

3.4 Multi-Conditional Background Synthesis

The core of our framework is a background-oriented synthesis module designed to explicitly model the intrinsic entanglement between camouflaged objects and their surrounding context. Given a real image $I_{\text{real}} \in \mathbb{R}^{H\times W\times 3}$ containing a camouflaged object and its corresponding binary instance mask $M \in \{0,1\}^{H\times W}$, where $M = 1$ denotes foreground pixels and $M = 0$ denotes background pixels, our objective is to synthesize new training samples that preserve camouflage characteristics while increasing contextual diversity.

We formulate the image as a composition of foreground and background regions:

$$I_{\text{real}} = I_{\text{fg}} \odot M \; + \; I_{\text{bg}} \odot (1 - M), \tag{1}$$

where $\odot$ denotes element-wise multiplication. Our synthesis strategy focuses on regenerating background content under multi-conditional guidance while carefully controlling the foreground–background interaction.

Multi-Conditional Guidance Background generation is driven by two complementary conditions. First, the instance mask M constrains the diffusion process, ensuring that foreground geometry, pose, and appearance are either strictly preserved or explicitly excluded, depending on the selected synthesis variant. Second, a textual condition in the form of a prompt T, describing plausible environmental contexts, guides the diffusion model to generate semantically coherent background textures that maintain the camouflage property. Conditioned jointly on (M, T), we employ an inpainting diffusion backbone to selectively resample background pixels, thereby increasing contextual diversity while preventing the model from overfitting to spurious background cues in the few-shot regime.

Original Instance on Synthesis Background (OISB) The first synthesis strategy, termed *Original Instance on Synthesis Background* (OISB), preserves the original foreground object while synthesizing a new background. Formally, the synthetic image I_{OISB} is defined as:

$$I_{\text{OISB}} = I_{\text{fg}} \odot M \; + \; \hat{I}_{\text{bg}} \odot (1 - M), \tag{2}$$

where $\hat{I}_{\text{bg}} = \mathcal{D}(I_{\text{real}}, M, T)$ is the background generated by the diffusion model $\mathcal{D}$ under mask and textual conditions. This strategy enforces instance consistency across real and synthetic samples, allowing the model to observe the same camouflaged object embedded in diverse yet semantically coherent environments. In our experiments, each real image contributes one original sample and one OISB-generated sample to the training set in few-shot settings.

Non-Instance on Synthesis Background (NISB) While OISB improves instance-level robustness by exposing the same camouflaged object to diverse backgrounds, the training process can still remain biased toward foreground appearance. To further emphasize contextual understanding, we introduce *Non-Instance on Synthesis Background* (NISB), which explicitly incorporates background-only synthesis into the training pipeline. In the NISB setting, each training group consists of three samples: one original image, one OISB-generated image that preserves the foreground instance, and one background-only synthesized image in which the foreground object is entirely removed. The background-only sample is formulated as

$$I_{\text{NISB}} = \hat{I}_{\text{bg}} \odot (1 - M), \tag{3}$$

where foreground pixels are masked out and excluded from supervision. By jointly training on the original image, the OISB sample, and the background-only synthesis, the model is encouraged to simultaneously learn instance-aware discrimination and robust background suppression. Unlike conventional negative samples, the synthesized background-only images are semantically consistent with camouflage environments, enabling the model to better recognize and reject background patterns that closely resemble camouflaged objects, thereby reducing false positives.

Sampling Strategy In this work, we adopt controlled sampling strategies tailored to each synthesis variant. For OISB, we employ a balanced ratio of *1 Real : 1 Synthetic*, where each real few-shot support image is paired with one OISB-generated sample. For NISB, the training set is expanded to include *1 OISB + 1 non-instance background-only synthesis* sample per real image, explicitly reinforcing background learning alongside instance preservation. These carefully designed ratios introduce sufficient contextual diversity while avoiding excessive dependence on synthetic data that could lead to domain bias. Together, OISB and NISB provide complementary perspectives for camouflage learning: OISB enhances instance discrimination across diverse contexts, whereas NISB explicitly regularizes background understanding, both of which are essential for effective few-shot camouflage instance segmentation.

Table 1. State-of-the-art comparison on camouflage few-shot object detection instance segmentation established on CAMO-FS benchmark.

Model			Novel AP									
Method	Year	Baseline	Instance Segmentation					Object Detection				
			1	2	3	5	Avg.	1	2	3	5	Avg.
MTFA [4]	2021	ResNet-50	2.48	6.67	5.81	6.40	5.34	1.98	6.47	5.82	6.17	5.11
M-RCNN† [9]	2017		4.08	6.79	6.90	8.29	6.52	2.82	5.09	5.46	6.18	4.89
iFS-RCNN [22]	2022		4.17	6.26	5.73	6.38	5.64	3.92	6.06	5.47	6.60	5.51
MTFA [4]	2021	ResNet-101	3.66	6.21	6.16	5.95	5.50	2.93	5.90	5.84	5.84	5.13
M-RCNN† [9]	2017		4.39	7.69	7.94	10.09	7.53	3.03	5.80	6.20	7.79	5.71
iFS-RCNN [22]	2022		4.27	6.55	6.07	7.80	6.17	3.79	6.28	6.01	8.08	6.04
FS-CDIS-ITL [28]	2024		4.46	5.57	6.41	8.48	6.23	4.04	7.28	7.49	9.76	7.14
FS-CDIS-IMS [28]	2024		5.46	6.95	7.36	9.61	7.35	4.50	6.95	7.55	10.36	7.34
CAMO-Freq [24]	2025		5.71	-	-	8.31	7.01	5.56	-	-	8.89	7.23
CAMO-GenOS [26]	2025		4.91	-	-	-	4.91	5.00	-	-	-	5.00
Our performance												
CAMO-InstSynth†	2026	ResNet-101	6.17	8.51	8.81	9.52	**8.25**	5.91	8.32	8.68	8.10	**7.75**
CAMO-InstSynth††	2026		6.12	8.62	9.16	10.48	**8.60**	5.79	8.55	8.63	8.33	**7.83**

M-RCNN† is Mask R-CNN [9] with sigmoid classifier.

CAMO-InstSynth† utilizes OISB; CAMO-InstSynth†† utilizes NISB; Both are built on top of iFS-RCNN [22].

4 Experiments

4.1 Experimental Setup

Settings Following the published experimental protocols of representative FSOD and FSIS methods [4,10,41,6,22,28], we configure our experiments to evaluate the proposed CAMO-InstSynth framework. We adopt FS-CDIS [28], built upon iFS-RCNN [22], as the primary baseline for all experiments. The implementation is based on the Detectron2 framework [39], using a ResNet-101 backbone equipped with a Feature Pyramid Network (FPN). All models are trained following a two-stage paradigm consisting of a base-training phase and a novel fine-tuning phase. In the base phase, the network is trained on the COCO dataset [18], leveraging abundant annotations from 80 semantic categories and over 118K training images. This stage strictly follows the default training configurations recommended by Detectron2 [39]. In the novel phase, the pretrained model is fine-tuned on the CAMO-FS dataset under few-shot settings with $K \in \{1, 2, 3, 5\}$ samples per novel class. The learning rate is set to $lr = 0.01$ and the batch size to $bz = 16$, consistent with the iFS-RCNN configuration [22]. Other hyperparameters in this phase follow the settings reported in FS-CDIS [28]. After training, the model is evaluated on the CAMO-FS test set, which contains $2,655$ images and $3,107$ annotated instances across 47 camouflage categories, to obtain the final performance. For additional implementation details and hyperparameter configurations in both training and testing stages, we refer readers to [28,22] and the Detectron2 documentation [39]. All experiments are conducted on a single NVIDIA GeForce RTX 3090 Ti GPU with CUDA version 12.4.

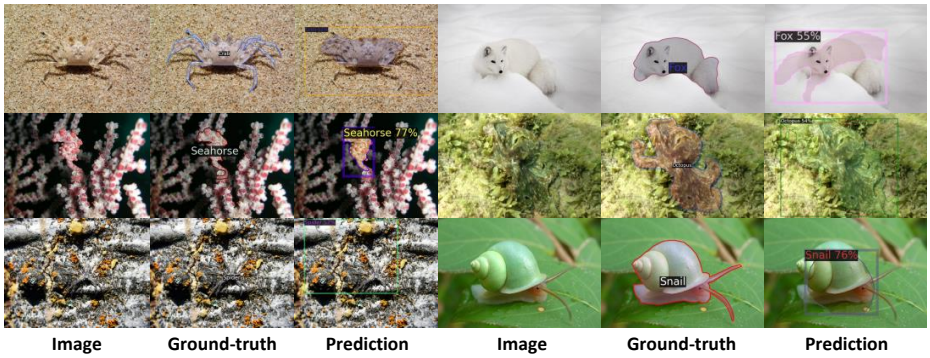

Fig. 2. Qualitative results of our CAMO-InstSynth reported on the CAMO-FS benchmark test set. The results are visualized under the configuration of the 5-shot setting.

Evaluation metrics To report our results on detection and instance segmentation tasks, we use average precision (AP) and average recall (AR). To be detailed, we report AP@50 and AP@75, along with AR@10. Besides, we also report AP and AR at small, medium, and large scales of the instances to further understand the model performance. For more details, readers can visit the homepage of the COCO dataset for detection and segmentation evaluation metrics[5].

4.2 State-of-the-art Comparison

We compared the performance of the vanilla MTFA [4], Mask RCNN† [9], iFS-RCNN [22], FS-CDIS [28] against our CAMO-InstSynth. Such methods are selected due to their published performance under the camouflage scenario. The quantitative results are presented in Table 1. The results clearly demonstrate the effectiveness of the proposed CAMO-InstSynth approach across both tasks and under varying shot settings. For the instance segmentation task, CAMO-InstSynth with the iFS-RCNN [22] baseline achieves an average Novel AP of 8.25%, representing a notable improvement over existing methods such as FS-CDIS-ITL [28] and FS-CDIS-IMS [28], whose best average performance remains below this level. The gains are consistent across $K = \{1, 2, 3, 5\}$ shot settings, indicating that the proposed synthesis-based strategy effectively enhances the model's ability to generalize to novel camouflage instances with limited supervision. Even when compared against strong multi-branch and transfer-learning-based baselines, CAMO-InstSynth maintains a clear performance margin, highlighting its robustness in learning discriminative representations for camouflaged objects.

Regarding the object detection task, CAMO-InstSynth further establishes state-of-the-art performance by achieving an average Novel AP of 7.75%, outperforming all competing approaches listed in the table. Notably, it delivers the highest detection accuracy in the extremely low-shot settings ($K = 1$ and

[5] https://cocodataset.org/#detection-eval

Table 2. The breakdown improvement of our proposed CAMO-InstSynth over two different approaches on top of baseline iFS-RCNN [22]. # denotes the Number of shots.

#	Method	AP	AP50	AP75	APs	APm	APl	AR1	AR10	AR100	ARs	ARm	ARl
Instance Segmentation													
1	**Baseline**	5.49	8.14	6.04	25.86	5.48	4.40	23.49	28.26	28.39	36.05	20.01	28.69
	iFS-RCNN + OISB	6.17	8.96	7.00	27.63	6.34	6.55	24.41	29.76	29.80	38.71	19.93	29.71
	iFS-RCNN + NISB	6.12	8.87	6.81	26.64	5.57	5.31	24.28	30.33	30.35	37.85	21.51	30.07
2	**Baseline**	8.07	12.08	9.27	32.59	6.19	6.40	27.47	36.62	36.86	41.49	23.89	36.86
	iFS-RCNN + OISB	8.51	12.62	10.10	34.45	6.88	6.51	28.79	36.97	37.19	42.87	24.83	36.40
	iFS-RCNN + NISB	8.62	12.53	9.98	34.29	7.99	5.56	29.48	37.17	37.41	41.63	26.16	36.75
3	**Baseline**	8.26	12.23	9.37	36.68	6.98	6.92	29.83	39.92	40.43	44.37	25.00	39.55
	iFS-RCNN + OISB	8.81	12.78	9.97	38.29	7.12	6.08	30.41	39.76	40.06	47.48	27.19	38.64
	iFS-RCNN + NISB	9.16	13.22	10.24	36.78	7.24	8.69	30.15	39.37	39.61	44.43	26.81	38.30
5	**Baseline**	8.76	13.41	10.04	38.73	6.02	8.36	28.80	40.51	41.40	43.38	27.06	41.03
	iFS-RCNN + OISB	9.52	14.39	10.77	39.02	6.73	7.48	29.71	41.21	41.69	43.67	27.75	40.90
	iFS-RCNN + NISB	10.48	15.84	11.54	38.18	7.19	9.03	29.06	41.14	41.76	42.83	29.41	41.55
Object Detection													
1	**Baseline**	5.25	8.16	5.90	29.74	7.06	4.17	20.60	25.53	25.72	34.33	19.34	26.95
	iFS-RCNN + OISB	5.91	9.11	6.75	30.05	7.91	6.47	22.56	27.65	27.72	34.91	19.50	27.89
	iFS-RCNN + NISB	5.79	8.93	6.63	29.44	7.59	5.03	22.73	28.13	28.23	35.04	20.86	29.08
2	**Baseline**	7.76	12.13	9.15	32.56	8.44	6.34	24.93	32.83	32.98	38.27	22.65	33.82
	iFS-RCNN + OISB	8.32	12.55	9.60	34.56	10.17	6.26	27.70	34.93	35.08	40.53	23.69	34.50
	iFS-RCNN + NISB	8.55	12.40	9.68	35.00	10.48	5.18	28.13	35.38	35.65	39.86	24.65	35.35
3	**Baseline**	7.77	12.32	8.47	37.18	8.52	6.33	26.36	35.28	35.70	41.37	22.84	36.22
	iFS-RCNN + OISB	8.68	12.64	9.95	39.36	12.29	6.36	29.36	38.32	38.68	45.55	25.99	38.36
	iFS-RCNN + NISB	8.63	13.13	9.29	38.08	11.50	8.08	28.34	36.62	36.88	42.02	25.66	36.36
5	**Baseline**	7.20	13.87	7.14	33.45	7.27	6.81	24.32	33.67	34.52	36.09	21.76	34.69
	iFS-RCNN + OISB	8.10	14.52	7.99	33.10	10.58	6.89	25.58	34.00	34.98	34.97	22.81	34.95
	iFS-RCNN + NISB	8.33	16.21	8.14	32.70	10.43	7.66	24.18	35.14	35.87	35.21	23.17	36.10

$K = 2$), where conventional few-shot detection methods typically suffer from unstable learning and poor localization. This consistent improvement across different shot numbers suggests that CAMO-InstSynth effectively mitigates the domain gap between base and novel classes in camouflage scenarios.

The results show that CAMO-InstSynth significantly advances the state-of-the-art on the CAMO-FS benchmark. By jointly improving instance segmentation and object detection performance under few-shot conditions, the proposed method demonstrates strong generalization capability and practical value for real-world camouflage understanding tasks, where annotated data are inherently scarce and object appearance is highly ambiguous. The visualization results are shown in Figure 2.

4.3 Ablation Study and Discussion

Table 2 reports the ablation results of the proposed CAMO-InstSynth framework on top of the baseline iFS-RCNN across different few-shot settings ($K = \{1, 2, 3, 5\}$), evaluated for both instance segmentation and object detection. Overall, the results consistently demonstrate that incorporating synthetic data via CAMO-InstSynth leads to clear and stable performance improvements over the baseline. In particular, both OISB and NISB variants outperform the vanilla

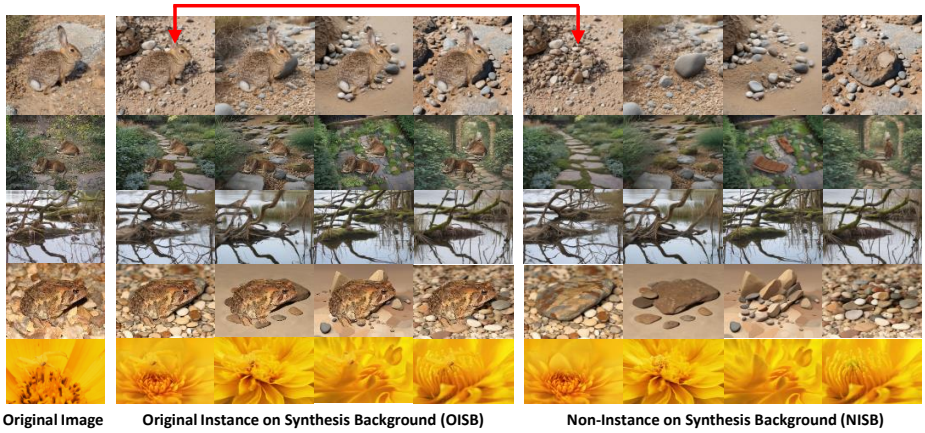

Fig. 3. Exemplary background synthesis results between the two variants of OISB and NISB.

iFS-RCNN in almost all metrics, indicating that background-level synthesis is beneficial for few-shot learning under camouflage conditions. The gains are especially evident in low-shot regimes ($K = 1$ and $K = 2$), where data scarcity is most severe, confirming that synthetic augmentation plays a crucial role in improving generalization.

Comparing the two variants, iFS-RCNN + NISB consistently achieves the best performance across most metrics and shot settings. For instance segmentation, NISB yields notable improvements in AP, AP50, and AR-based metrics compared to both the baseline and OISB, particularly at $K = 3$ and $K = 5$, suggesting that background-aware synthesis better captures the intrinsic camouflage characteristics where foreground and background are strongly entangled. Similar trends are observed in object detection, where NISB surpasses OISB in most cases, especially in higher-shot settings, indicating more robust localization and recall. These results validate our design choice of emphasizing background-oriented synthesis rather than purely foreground manipulation, and highlight the effectiveness of CAMO-InstSynth in enhancing few-shot camouflage instance segmentation and detection. Figure 3 shows the background synthesis results from both OISB and NISB.

5 Conclusion and Future Work

In this paper, we presented CAMO-InstSynth, a method to alleviate the data scarcity problem in few-shot camouflage instance segmentation. By integrating a diffusion-based background synthesis module, we are able to generate realistic, challenging training samples that respect the semantic context of camouflaged objects. Our empirical results on CAMO-FS show a consistent improvement over the state-of-the-art iFS-RCNN baseline. We conclude that generative back-

ground synthesis is a promising direction for robust camouflage understanding, offering a way to simulate infinite environmental variations from limited data.

In the future, we will focus on extending our CAMO-InstSynth to handle more complex camouflage scenarios, including multi-object, multi-scale, and cluttered environments, which are common in real-world applications. We also plan to investigate task-adaptive generative frameworks that jointly optimize background synthesis and instance segmentation in an end-to-end manner, allowing the generative process to be guided by downstream learning objectives. In addition, we will explore the generalization of the proposed diffusion-based synthesis strategy to other challenging domains characterized by severe foreground–background ambiguity and limited annotations, such as video-based camouflage understanding, infrared imagery, underwater scenes, and medical image segmentation.

Acknowledgments. This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS.C2025-26-08.

Disclaimer. The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, the United States Air Force or the United States Space Force.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR (2022)
2. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR. pp. 2777–2787 (2020)
3. Fan, Z., Yu, J.G., Liang, Z., Ou, J., Gao, C., Xia, G.S., Li, Y.: Fgn: Fully guided network for few-shot instance segmentation. In: CVPR. pp. 9172–9181 (2020)
4. Ganea, D.A., Boom, B., Poppe, R.: Incremental few-shot instance segmentation. In: CVPR. pp. 1185–1194 (2021)
5. Gao, B.B., Chen, X., Huang, Z., Nie, C., Liu, J., Lai, J., Jiang, G., Wang, X., Wang, C.: Decoupling classifier for boosting few-shot object detection and instance segmentation. In: NeurIPS (2022)
6. Gao, B.B., Chen, X., Huang, Z., et al.: Decoupling classifier for boosting few-shot object detection and instance segmentation. NeurIPS **35**, 18640–18652 (2022)
7. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: CVPR. pp. 4367–4375 (2018)
8. Han, Y., Zhang, J., et al.: Reference twice: A simple and unified baseline for few-shot instance segmentation. TPAMI **46**(12), 9221–9238 (2024)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2980–2988 (2017)
10. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: ICCV. pp. – (2019)
11. Lamdouar, H., Yang, C., Xie, W., Zisserman, A.: Betrayed by motion: Camouflaged object discovery via motion segmentation. In: ACCV. pp. – (November 2020)

12. Le, M.Q., Nguyen, T.V., Le, T.N., Do, T.T., Do, M.N., Tran, M.T.: Maskdiff: Modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation. In: AAAI. vol. 38, pp. 2874–2881 (2024)
13. Le, T.N., Cao, Y., et al.: Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. TIP **31**, 287–300 (2022)
14. Le, T.N., Nguyen, H.H., Yamagishi, J., Echizen, I.: Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In: ICCV. pp. – (2021)
15. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. CVIU **184**, 45–56 (2019)
16. Le, T.N., Nguyen, V., et al.: Camoufinder: Finding camouflaged instances in images. In: AAAI. vol. 35, pp. 16071–16074 (2021)
17. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
18. Lin, T.Y., Maire, M., et al.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
19. Liu, W., Zhang, C., Lin, G., Liu, F.: Crnet: Cross-reference networks for few-shot segmentation. In: CVPR. pp. – (June 2020)
20. Lv, Y., Zhang, J., Dai, Y., et al.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR. pp. 11591–11601 (2021)
21. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: CVPR. pp. 8772–8781 (2021)
22. Nguyen, K., Todorovic, S.: ifs-rcnn: An incremental few-shot instance segmenter. In: CVPR. pp. 7010–7019 (2022)
23. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. NeurIPS **36**, 76872–76892 (2023)
24. Nguyen, T.D., Cao, H.P., Ngo, T.D., Nguyen, V.T., Nguyen, T.V.: Few-shot instance segmentation: An exploration in the frequency domain for camouflage instances. In: MAPR. pp. 1–6. IEEE (2025)
25. Nguyen, T.D., Nguyen, V.T., Nguyen, T.V.: A generative approach at the instance-level for image segmentation under limited training data conditions (student abstract). In: AAAI. vol. 39, pp. 29451–29452 (2025)
26. Nguyen, T.D., Nguyen, V.T., Nguyen, T.V.: Generative one-shot camouflage instance segmentation. In: MAPR. pp. 1–6. IEEE (2025)
27. Nguyen, T.D., Pham, B.N., Vu, T.T.D., Nguyen, V.T., Ngo, T.D., Nguyen, T.V.: Instsynth: Instance-wise prompt-guided style masked conditional data synthesis for scene understanding. In: MAPR. pp. 1–6. IEEE (2024)
28. Nguyen, T.D., Vu, A.K.N., Nguyen, N.D., Nguyen, V.T., Ngo, T.D., Do, T.T., Tran, M.T., Nguyen, T.V.: The art of camouflage: Few-shot learning for animal detection and segmentation. IEEE Access (2024)
29. Nguyen, T.D., Vu, T.T.D., et al.: Camul: Context-aware multi-conditional instance synthesis for image segmentation. IEEE MultiMedia (2025)
30. Pia Bideau, E.L.M.: It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In: ECCV. pp. – (2016)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
32. Saha, O., Cheng, Z., Maji, S.: Ganorcon: Are generative models useful for few-shot segmentation? In: CVPR. pp. 9991–10000 (June 2022)
33. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: BMVC (2017)

34. Skurowski, P., Abdulameer, H., Baszczyk, J., Depta, T., Kornacki, A., Kozie, P.: Animal camouflage analysis: Chameleon database. - (2018)
35. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS. vol. 30 (2017)
36. Sulimowicz, L., Ahmad, I., Aved, A.: Superpixel-enhanced pairwise conditional random field for semantic segmentation. In: ICIP. pp. 271–275 (2018)
37. Tian, Z., Lai, X., Jiang, L., Liu, S., Shu, M., Zhao, H., Jia, J.: Generalized few-shot semantic segmentation. In: CVPR. pp. 11563–11572 (June 2022)
38. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: ICCV. pp. – (October 2019)
39. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
40. Yan, J., Le, T.N., et al.: Mirrornet: Bio-inspired camouflaged object segmentation. IEEE Access **9**, 43290–43300 (2021)
41. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: ICCV. pp. – (2019)
42. Zhang, C., Lin, G., et al.: Canet: Class-agnostic segmentation networks with iterative refinement and attention. In: CVPR. pp. 5217–5226 (2019)
43. Zhu, J., Zhang, X., Zhang, S., Liu, J.: Inferring camouflage objects by texture-aware interactive guidance network. In: AAAI. pp. – (2021)