IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 June 2024, accepted 17 July 2024, date of publication xx xxxxx 2024, date of current version 20 July 2024.

Digital Object Identifier xx.xxxx/ACCESS.xxxx.xxxxxxx

The Art of Camouflage: Few-shot Learning for Animal Detection and Segmentation

THANH-DANH NGUYEN[1,4,*], ANH-KHOA NGUYEN VU[1,4,*], NHAT-DUY NGUYEN[1,4,*], VINH-TIEP NGUYEN[1,4], THANH DUC NGO[1,4], THANH-TOAN DO[5], MINH-TRIET TRAN[2,3,4], TAM V. NGUYEN[6], (Senior Member, IEEE)

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]University of Science, Ho Chi Minh City, Vietnam
[3]John von Neumann Institute, VNU-HCM, Vietnam
[4]Vietnam National University, Ho Chi Minh City, Vietnam
[5]Monash University, Clayton, VIC 3800, Australia
[6]University of Dayton, Dayton, OH 45469, United States

Corresponding author: Tam V. Nguyen (e-mail: tamnguyen@udayton.edu), *equal contribution

ABSTRACT

Camouflaged object detection and segmentation is a new and challenging research topic in computer vision. There is a serious issue of lacking data on concealed objects such as camouflaged animals in natural scenes. In this paper, we address the problem of few-shot learning for camouflaged object detection and segmentation. To this end, we first collect a new dataset, CAMO-FS, for the benchmark. As camouflaged instances are challenging to recognize due to their similarity compared to the surroundings, we guide our models to obtain camouflaged features that highly distinguish the instances from the background. In this work, we propose FS-CDIS, a framework to efficiently detect and segment camouflaged instances via two loss functions contributing to the training process. Firstly, the instance triplet loss with the characteristic of differentiating the anchor, which is the mean of all camouflaged foreground points, and the background points are employed to work at the instance level. Secondly, to consolidate the generalization at the class level, we present instance memory storage with the scope of storing camouflaged features of the same category, allowing the model to capture further class-level information during the learning process. The extensive experiments demonstrated that our proposed method achieves state-of-the-art performance on the newly collected dataset. Code is available at https://github.com/danhntd/FS-CDIS.

INDEX TERMS Camouflaged animals, camouflaged instances, few-shot learning, object detection, and instance segmentation.

I. INTRODUCTION

Camouflage is a defense mechanism that animals use to conceal their appearance by blending in with their environment [1]. Autonomously detecting camouflaged animals is helpful in various applications, e.g., search-and-rescue missions [2]; wild species discovery and preservation activities [2]; and media forensics [3]–[6](manipulated image/video detection and segmentation [7]). By leveraging camouflage recognition at object detection or instance segmentation level autonomously, these practical applications can be done with minor efforts from humans while maintaining work performance. To be specific, utilizing a drone flying around the mountainous area to collect images and videos for a system to detect and segment the target object in danger is more effective than sending a group of humans manually scanning the zone. By this means, this process can support biological scientists in identifying and preserving endangered species effectively. Further related applications can be considered in different important fields including healthcare, agriculture, or military, where exist the concept of finding objects with camouflaged features. Indeed, camouflage detection and segmentation tasks can provide further applications such as assisting doctors in medical imaging understanding, supporting modern farmers in managing the vast fields of crops via visual information or even detecting hidden enemies in nature. These applications can be potential via the development of camouflaged research via image understanding at detailed levels of object detection and instance segmentation.

Although image segmentation methods have been proposed for a long time, general detectors cannot deal with cam-

TABLE 1: Statistics of camouflage datasets (without non-camo images).

Dataset	Year	Venue	Type	#Annot. Camo. Img.	#Meta-Cat.	#Obj. Cat.	#Ins. or #Obj. per Img.	Bbox. GT	Obj. Mask GT	Ins. Mask GT	Few-shot
CamouflagedAnimals [8]	2016	ECCV	Video	181	-	6	1.238	×	✓	✓	×
MoCA [9]	2020	ACCV	Video	7,617	-	67	1.000	✓	×	×	×
CHAMELEON [10]	2018	-	Image	76	-	-	1.000	×	✓	×	×
CAMO [2]	2019	CVIU	Image	1,250	2	8	1.000	×	✓	×	×
COD [11]	2020	CVPR	Image	5,066	5	69	1.171	✓	✓	✓	×
NC4K [12]	2021	CVPR	Image	4,121	5	69	1.171	✓	✓	✓	×
CAMO++ [13]	2022	TIP	Image	2,695	10	47	1.171	✓	✓	✓	×
CAMO-FS (Ours)	2024	IEEE ACCESS	Image	2,852	10	47	1.172	✓	✓	✓	✓

ouflaged animals [14]–[17] due to their specific camouflaged features. The detectors initially developed for camouflage detection [18]–[25], which use handcrafted low-level features, are effective only for images with a simple and uniform background. More recently developed deep learning-based detectors [2], [10]–[12], [26]–[30] for camouflaged object segmentation. Most previous methods are trained on large-scale datasets to perform computer vision tasks. Nevertheless, building such standard datasets for camouflaged objects is labor-intensive due to the ambiguity between the objects and their backgrounds, which leads to more time and cost in the labeling procedure. To address the problem, we consider the camouflaged object detection and instance segmentation under the few-shot learning which shows potential results when utilizing limited labeled samples to classify [31]–[38], detect [39]–[48], or even segment [49]–[53] new objects. However, to the best of our knowledge, there is a lack of camouflaged datasets supporting few-shot learning in camouflaged research. Therefore, we introduce a new benchmark, CAMO-FS, for camouflaged few-shot object detection and instance segmentation under the few-shot settings. The new benchmark is mainly reconstructed from the CAMO++ dataset [13] due to its diversity. The new benchmark consists of 197 camouflaged images for training and $2,655$ camouflaged images for performance evaluation as described in Table 1.

There are several approaches for few-shot learning. Beginning with few-shot classification (FSL), many works are based on meta-learning [31]–[34], [54] or transfer-learning [35]–[38] approaches to leverage a few labeled data to classify new objects and achieve incredible results. The former approaches compute the similarity between query and support images to pinpoint novel objects while the latter involves utilizing knowledge from the source domain to adapt a different but related target domain. Fueled by these successes, most existing works on few-shot object detection (FSOD) and few-shot segmentation (FSS) which are recently developed to tackle the problem through meta-learning [42]–[48], [51]–[53], [55]–[57] and transfer-learning [39]–[41], [49], [50], [58], [59] methods. Nonetheless, such methods focus on the general domain and thus fail to generate effective features for camouflaged objects due to the ambiguity between backgrounds and foregrounds.

To overcome the specific issue of camouflage objects, we propose a novel framework for few-shot camouflaged object detection and instance segmentation, dubbed FS-CDIS, which is based on the transfer-learning approach. The model is trained on two stages of processing: (1) one base phase training for the model to gain concept knowledge of general domains with abundant data, and then (2) performing a novel phase that can do the specific task on the few-shot data. To be specific, in the base training stage, we train our model on generic object detection and instance segmentation dataset (e.g. COCO [60]) and focus on improving the model in the novel fine-tuning stage. Because of the similarity between camouflaged objects and their surroundings, we aim to guide our few-shot models to obtain camouflaged features that highly distinguish the instances from the background. To achieve that target, we introduce two loss functions contributing directly to the novel fine-tuning process. Firstly, the instance triplet loss with the characteristic of differentiating the anchor, which is the mean of all camouflaged foreground points, and the background points are employed to work at the instance level. Secondly, to consolidate the generalization at the class level, we present instance memory storage with the scope of storing camouflaged features of the same category, allowing the model to capture further class-level information during the learning process.

To summarize, our contributions in this work are two-fold:

- First, we build a new benchmark dataset, CAMO-FS, which is among the first datasets to support detection and instance segmentation on camouflaged instances in nature under the few-shot concept.
- Second, we propose a framework to detect and segment camouflaged instances efficiently, named after FS-CDIS, given a small shot of training data for novel classes utilizing the idea of instance triplet loss and instance memory storage.

The remainder of this paper is organized as follows. Section II summarizes related work. Next, Section III introduces the newly constructed CAMO-FS dataset and presents our proposed framework for few-shot camouflaged object detection and segmentation. Section IV presents the experimental results and comparison among baselines and our proposals on the newly constructed dataset. Finally, Section V summarizes the key points and mentions future work.

II. RELATED WORK

A. CAMOUFLAGE RESEARCH

Given any region (i.e. bounding boxes or polygon masks) presented for an object of interest (i.e. animals or artificial

objects) in an image and then they tend to be classified as background, contents in that region can be qualified as camouflaged objects. Thus, a camouflaged object is defined as a set of bounding boxes or camouflaged pixels in an image without any further detailed information such as the number of objects or the semantic meaning [2]. Although tasks related to camouflaged animals can be performed in a wide range of applications such as security systems [3], [4], pollution detection [5], watermark detection [6], this research field has not been well explored in the literature, especially few-shot learning which is practically suitable to the context of scare data as camouflaged animals.

Binary camouflage segmentation. Prior to the advancement of deep neural networks, most of the work exploits identical regions between camouflaged regions and the background by handcrafted or low-level features, specifically based on external characteristics (e.g., color, shape, orientation, and brightness). Particularly, early camouflage detection works had attention on the foreground region even when some of its texture was similar to the background [18]–[20], [61]. The foreground was distinguishable from the background via simple features, such as color, intensity, shape, orientation, and edge [61]–[65]. A few methods [21]–[25], [66] based on handcrafted low-level features have been proposed for tackling the problem of camouflage detection. However, they are effective only for images with a simple and uniform background. Thus, their performances are unsatisfactory in camouflaged object segmentation due to the substantial similarity between the foreground and the background.

Until now, the convention of binary prefers binary ground truth camouflaged object datasets [2], [10], [11]. Existing methods for camouflaged objects [2], [11], [12], [26]–[30] based on binary ground truth are considered as the binary camouflage segmentation. For example, Le *et al.* [2] proposed an end-to-end Anabranch Network, dubbed ANet which includes two streams of classification and segmentation. The outputs of both streams are fused to improve the segmentation performance of camouflaged objects. This proposed network was also flexibly applied to any fully convolutional networks. Similarly, motivated by the way of hunting strategies of predators, Fan *et al.* [11] designed Search Identification Network (SINet) with two main modules to simulate this hunting behavior, namely a search module searching for targets and an identification module identifying the existence of targets then catching them. Yan *et al.* [29] recently introduced MirrorNet, a dual-stream network comprising a mainstream and a mirror stream. This mirror stream aimed to capture instinct information by horizontally flipping camouflaged objects to break their camouflaged nature and make them more distinguishable. Zhu *et al.* [30] presented the TINet, which interactively refines multi-level texture and segmentation features and thereby gradually enhances the segmentation of camouflaged objects. Lv *et al.* [12] simultaneously worked on ranking and localization to well-present camouflaged objects. As a result, they formed a triplet task with localizing, segmenting, and ranking the camouflaged objects. Besides, the authors also introduced the NC4K dataset for camouflaged segmentation. Such methods reveal the presence of the camouflaged objects with the high level of bounding boxes and contain corresponding pixel-wise ground truth belonging to camouflage. Further understanding of the camouflage level may help us to give comparative analyses, finding evidence for links between camouflage and other defensive strategies with aspects of habitat and life-history [67].

Camouflage instance segmentation. Although several works have been proposed, there is still a difficulty in efficiently exploring the information of camouflage animals, especially at the instance level with more challenging detailed masks. Therefore, for ease of training methods with the challenging task of camouflaged instance segmentation, Le *et al.* [68] introduced a framework with several state-of-the-art methods and proposed a tool with user interactive cues to tune the segmentation mask on a website. Realizing that the semantic level is not detailed enough, Le *et al.* [13] introduced a camouflage fusion learning (CFL) to utilize the strength of different instance segmentation methods by fusing various models via learning image contexts.

Camouflage datasets. CamouflagedAnimals [8] and CHAMELEON [10] were the first two camouflage datasets with mask annotations. The two datasets do not contain enough images to train deep learning methods. Le *et al.* [2] created the CAMO dataset, the first camouflage dataset with more than $1,000$ annotated images. It contains $1,250$ annotated images, which is a limited number of samples to train and evaluate deep learning methods. Then, Fan *et al.* [11] collected the COD dataset, which comprises $10,000$ images (both camouflage and non-camouflage) divided into 5 meta-categories. However, they annotated only $5,066$ camouflage images. Lamdouar *et al.* [9] recently developed the MoCA dataset for the camouflage object detection task; it contains only bounding box ground truths. Hence, these datasets limit their annotations at binary ground truth datasets which have a shortage of intensive annotations for multi-task camouflage problems. CAMO++ [13] is different from the aforementioned datasets providing a benchmark for camouflaged instance segmentation with more comprehensive annotations and diverse meta-categories of 10. The dataset comprises $5,500$ images with superiority over other datasets on instances including $32,756$ instances for both camo and non-camo objects. Different from the existing work, we address camouflage research under the few-shot learning concept to detect objects and segment camouflaged instances. Therefore, we introduce a new benchmark, dubbed CAMO-FS, to support the evaluation process of this specific task. Accordingly, our CAMO-FS comprises $2,852$ images as a result of the inheritance from CAMO++ [13] and our further collection. This specific dataset serves $93,1\%$ of the annotated images for the evaluation process while the rest few samples are provided for training (i.e. 197 images for training and $2,655$ images for testing).

B. FEW-SHOT LEARNING

Few-shot object detection (FSOD). When having some available samples of given classes with their corresponding bounding boxes, FSOD aims to learn from these limited data in order to help models adapt to the new classes. To date, several works [39], [42]–[44] have been proposed to deal with FSOD. Early works [42], [43] mainly prefer to overcome the difficulties of the data scarcity of FSOD via meta-learning approaches by combining supportive information from meta-based streams with their main streams. Particularly, Bingyi [42] proposed a Feature Reweighting framework that leverages the free-proposal approach of a well-known one-stage framework such as YOLO [69] to boost FSOD performance. The network integrated a meta-model that aims to generate reweighting vectors from support samples for highlighting the attention to features from the YOLO network. Conversely, Meta RCNN [43] based on the two-stage proposal approach as Mask RCNN [70] and fed available annotations such as bounding boxes and segmented masks to train a meta-network called Predictor-head Remodeling Network for inferring attention features. Fan *et al.* [44] recently proposed to take advantage of support images from a massive FSOD dataset to generate significant results combined with their proposed network called Attention-RPN, Multi-Relation Detectors. The Attention-RPN directed the trained model to look at the image for the task of object detection. Differently, Wang *et al.* [39] simply adopted Faster RCNN with two-stage finetuning to transfer massive knowledge from abundant data in the base model to fine-tune the novel one by freezing the whole network except for the fully connected layer for object classification. Through this simple straightforward mechanism, this model significantly improved few-shot performance without a complex pipeline of training the model. Further, such works [48], [56], [71]–[73] presented advanced methods by applying class max-margin, multiple scale proposals, or feature alignment in FSOD. Other ones were based on transformed inputs [57], [74], transformer approaches [45], [46], contrastive method [55], or kernels design [75]. Other methods [39]–[41], [76] relied only on query images to deal with FSOD via extra text data [41], unlabeled image [77], generated samples [76], gradient scaling [40].

Few-shot object segmentation (FSS). Recently, the field of few-shot segmentation gained attention from the community. As mentioned above, the first work Meta RCNN originated from Mask RCNN, therefore, Meta RCNN simultaneously performed detection and segmentation. Liu *et al.* [78] utilized a cross-reference network for generic image segmentation. The authors proposed a cross-reference mechanism and a mask refinement module to specifically support the task of segmentation. Before, Dong *et al.* [79] proposed a prototype learning component in a framework of semantic segmentation that learned to take discriminative information from features to help segment objects better. Also, Wang *et al.* [80] introduced a prototype align method that learns class-specific prototype representations from a few image samples

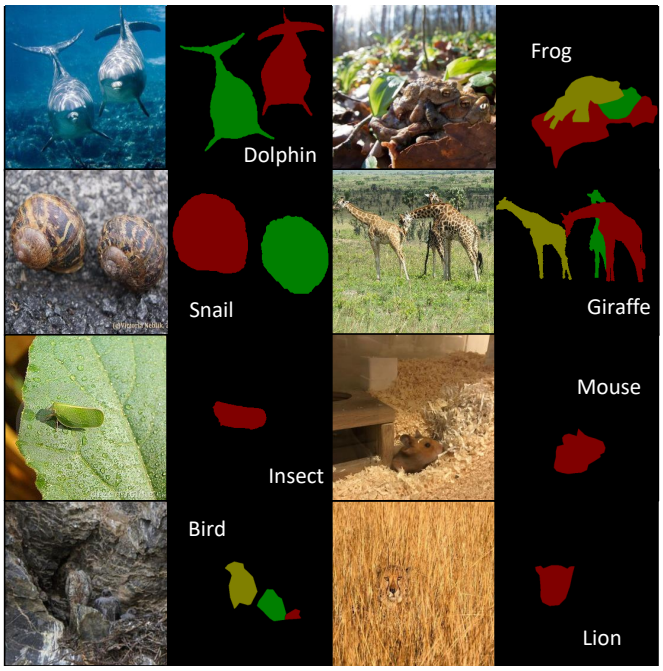


FIGURE 1: Exemplary images with instance-level mask annotations from our proposed CAMO-FS dataset.

#Instances	Ratio (%)	#Images
1	85.55	2440
2	8.31	237
3	3.44	98
3+	2.70	77

TABLE 2: Number of instances per image of CAMO-FS.

to perform segmentation over the query images. Lately, Liu *et al.* proposed a dynamic prototype convolution network to address few-shot semantic segmentation. The work of [81] proposed context-aware prototype learning. [82] introduced generative models approach for this task. Recently, Nguyen *et al.* [83] came up with iFS-RCNN, an instance segmenter via an incremental approach. Gao *et al.* [84] proposed the DCFS framework, an effective decoupling classifier that boosted the performance of object detection and segmentation heads. Han *et al.* [85] suggested a reference twice transformer-based framework (RefT) to enhance features in segmentation tasks. Also in the transformer approach, Wang *et al.* [86] introduced DTN to directly segment the target object instances from arbitrary categories given reference images.

In common, these aforementioned methods of the two approaches including FSOD and FSS mostly focus on generic objects, which cannot create effective distinguished features and fail to recognize camouflaged objects. In our case, our proposed methods aim at highlighting the differences between backgrounds and foregrounds which we considered as the key feature to detect or segment camouflaged objects. Furthermore, our proposed approaches contribute directly to the training process of such models via loss functions.

TABLE 3: Extra collected number of images and instances in CAMO-FS dataset.

Classes	Bat	Bear	Camel	Dolphin	Elephant	Horse	Kangaroo	Monkey	Penguin	Rhino	Squirrel	Total
#images	12	14	14	13	14	16	22	16	11	14	17	163
#instances	12	14	15	19	14	17	25	20	14	14	17	181

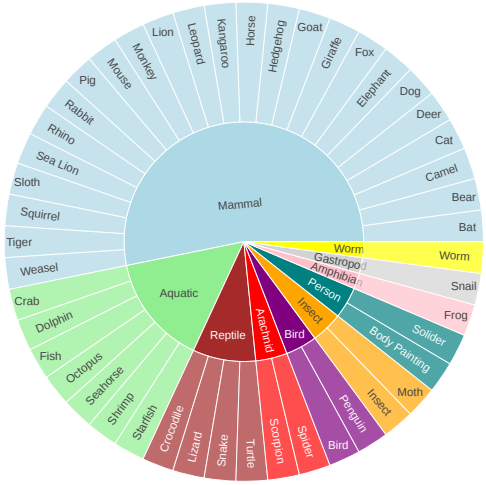


FIGURE 2: Hierarchical taxonomic structure of our CAMO-FS dataset.

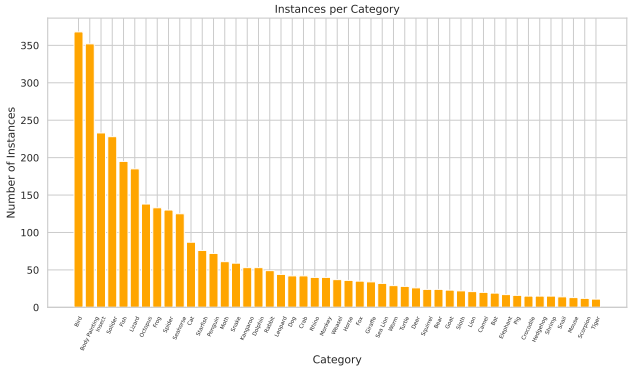


FIGURE 3: The distribution of CAMO-FS dataset. The categories are sorted.

III. PROPOSED METHOD

A. CAMO-FS BENCHMARK DATASET

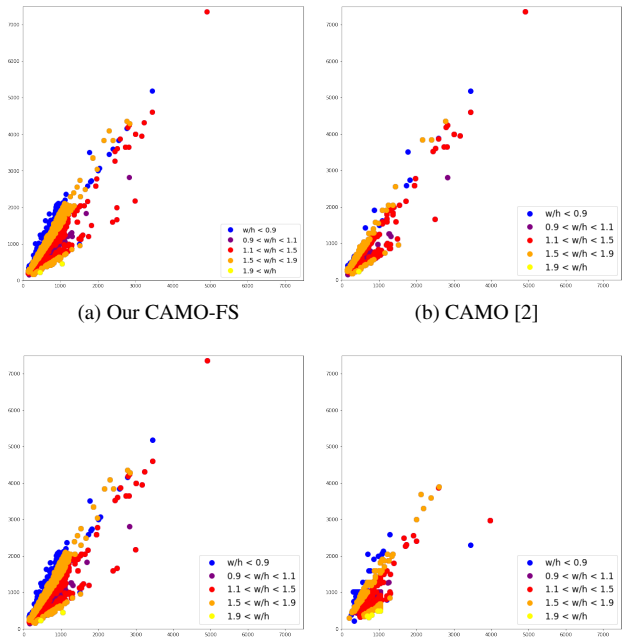
Camouflaged data tends to be more difficult to collect in the real world rather than non-camouflaged ones. Generating intensive annotations with multi-task or hierarchical labels for camouflaged objects is also costly and complicated, especially with the pixel level as polygon masks. Particularly, the visual characteristics of a camouflaged object are extremely identical to the background. The external appearances (i.e. the intensity, color, and textures) are close to their surrounding environment, the boundary between camouflaged objects and the background or other identical-type camouflaged objects in case of being nearly or partly overlapped. Thus, it is really tough to provide the concurrence between annotators due to ambiguity in verifying camouflaged regions blended in surroundings. For ease of data preparation such as collections and annotations, one of the most common ways is to inherit existing camouflaged datasets and CAMO++ [13] is our selected dataset since it is a high-diversity dataset with a variety of camouflaged object categories. Furthermore, the key to few-shot learning lies in the generalization ability of the pertinent model when presented with a few available samples. The context of camouflaged objects inherently matches this understanding because the number of camouflaged images is often scarce in practice.

CAMO++ Dataset. CAMO++ generally contains camouflaged and non-camouflaged images with a total of 5,500 images corresponding to 32,756 instances [13]. The dataset contains 93 fine-grained classes assigned to 13 coarse-grained classes. However, in the case of camouflaged objects, there are 47 fine-grained classes designed with a hierarchical structure and assigned into 10 coarse-grained classes. In detail, CAMO++ contributes 2,695 camouflage images including 1,250 existing camouflage images in the previous CAMO dataset with 1,450 newly collected camouflage images for CAMO++. In this scope of our paper, 2,800 remaining non-camouflage images are ignored. CAMO++ especially provides common ground truths such as bounding boxes, object masks, and instance masks which are suitable for many tasks of camouflage research.

CAMO-FS Dataset. In general, there are three steps to construct the CAMO-FS dataset: **inheritance**, **collection**, **data splitting**.

In the **inheritance** step, we leverage the available CAMO++ to build our CAMO-FS dataset. In this way, we inherit the biology taxonomic and vision taxonomic structure of CAMO++ which helps us to reduce the burden of data collection. Table 1 provides an overview of previous works done on camouflage, which is mentioned in the related work, and our proposed CAMO-FS in terms of main characteristics. We exploit the diversity of CAMO++ by its 10 meta-categories to build up the few-shot concept for instance segmentation. To this end, our CAMO-FS not only keeps a good ratio of instances per image of 1.172 but also contributes as the very first dataset specific for few-shot research on camouflaged animals. Note that the large amount of images in some datasets does not mean they are all camouflaged images.

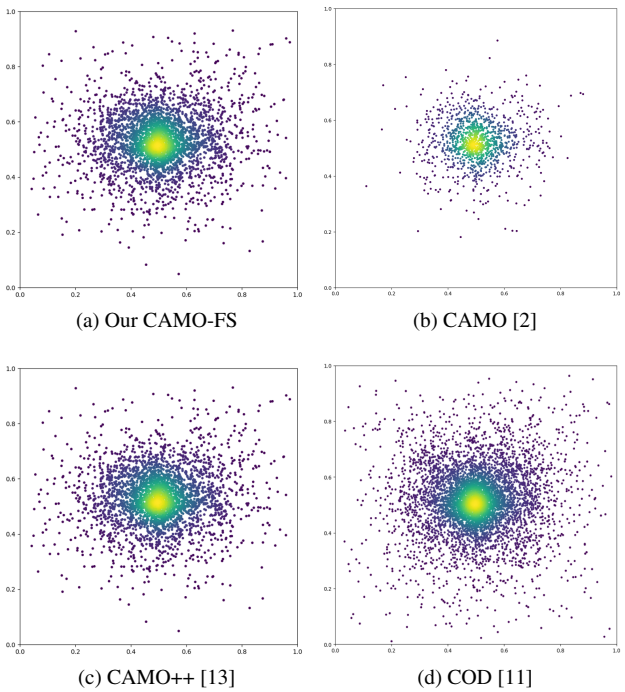
In the **collection** step, as CAMO++ faces issues such as imbalanced data and a shortage of the number of images of some classes, which cause evaluation problems for few-shot tasks. Particularly, there are 11 classes (*e.g. Camel, Dolphin,*



(a) Our CAMO-FS (b) CAMO [2]

(c) CAMO++ [13] (d) COD [11]

FIGURE 4: Distribution of camouflage image resolution. Best viewed online in color and zoomed in.



(a) Our CAMO-FS (b) CAMO [2]

(c) CAMO++ [13] (d) COD [11]

FIGURE 5: Instance center bias camouflaged datasets. Best viewed online with color and zoomed in.

Elephant, Horse, Kangaroo, Monkey, Penguin, Bat, Bear, Squirrel, and Rhino) having a shortage of images that are needed to train a few-shot model. Hence, we hardly perform training or testing on these classes. As a result, we manually collect more data for these classes with 163 total images corresponding to 181 instances (an average of 15-16 instances per class) on Google Image Search Engine with their class names as the search query. We also remove images with mistakes in the original dataset. The statistics of collected data are shown in Table 3. By gathering more camouflaged animals and combining them with the CAMO++ dataset, we conduct our CAMO-FS dataset for few-shot camouflaged animal detection and segmentation with $2,852$ total images corresponding to $3,342$ instances. Figure 2 shows the vision taxonomic structure of coarse-grained and corresponding fine-grained classes and illustrates the ratios of 10 coarse-grained classes in our proposed CAMO-FS dataset. We also show the distribution of 47 camouflaged classes in Figure 3, which indicates that CAMO-FS is a diverse and long-tailed dataset. Figure 1 shows exemplary images with mask annotations from our proposed CAMO-FS.

In Table 2, we report the aggregated number of instances per image. The number of instances per image ranges from 1 to 25 and commonly falls into 1, then 2 and 3 while the remaining is beyond 3 instances. As can be seen, the number of images that contain 1 to 3 instances takes up a large proportion of the entire dataset. This also illustrates the problem of data imbalance between the number of instances and the ratio of images in the dataset, which reflects a problem that the presence of camouflaged animals captured in photos is often limited, i.e. mostly one animal per image. Additionally, although being claimed in [13] that camouflaged objects in CAMO++ were localized over the entire image, after removing non-camouflage objects and adding new camouflaged images, we have the distributions of object centers in normalized image coordinates over all images in the CAMO-FS dataset as in Figure 5-a. This means camouflaged animals tend to be located in the center of images. Indeed, to capture images of camouflaged animals in the wild, photographers need to carefully focus on the animals, which leads to the central layout of collected images. Also in Figure 5, we illustrate the center bias of camouflaged images in other CAMO [2] and COD [11] datasets for better visual comparison. In Figure 4, we present the image resolution among camouflage datasets. As we only consider camouflaged images of CAMO++ [13] and COD [11], the density of our CAMO-FS is slightly higher than CAMO++ as a result of our extra collection of images presented in Table 3. In comparison with the previous COD [11] and CAMO [2], our CAMO-FS image resolution distribution is more satisfying in diversity.

In the **data splitting** step, to effectively create the data for the few-shot problem, we randomly get M instances for each camouflaged class from the CAMO-FS dataset to create training sets. In our setup, M equals 5 for 47 classes and thus there are 197 training images containing 235 camouflaged instances and the remaining $2,655$ images with $3,107$ instances are used for testing. We only remove some objects of the higher-level training set if they exist to create the other few-

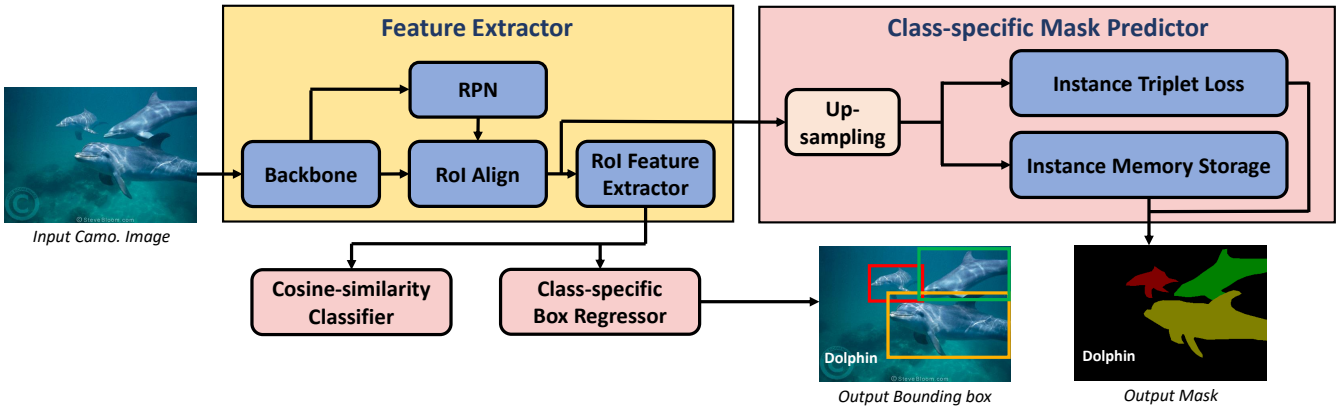


FIGURE 6: Our general FS-CDIS framework for Few-Shot Camouflaged Detection and Instance Segmentation.

shot settings. For example, we get all elements to generate 5-shot training data and discard 2 in 5 objects to make a 3-shot one. In this way, the 5-shot benchmark contains objects of the 3-shot dataset and the 3-shot setting contains the objects of the 2-shot one.

To the best of our knowledge, this is among the first works to address few-shot camouflaged instance segmentation and detection. Given the lack of a large-scale dataset for training and testing purposes on camouflaged animal issues, we build a benchmark for the task of few-shot camouflaged instance segmentation and detection.

B. GENERAL FRAMEWORK

Few-shot instance segmentation formulation. In few-shot learning, we have one set of base classes denoted C_{base} with a large amount of available training data, and one disjoint set of novel classes denoted C_{novel} containing a small amount of training data. This amount is small to a few samples. The ultimate goal is to train a model to predict well on the novel classes $C_{test} = C_{novel}$ [31], [87] or on both base and novel data $C_{test} = C_{base} \cup C_{novel}$ [88]. In few-shot classification, this work [31] introduces the method of episodic training. The method sets up a series of episodes $E_i = (I_q, S_i)$ where S_i is a support set that contains N classes from $C_{train} = C_{novel} \cup C_{base}$ along with K examples per class (so-called N-way K-shot). A network is then trained to classify an input image I_q, termed query image, out of the classes in S_i. The key idea is that solving a different classification task for each episode leads to better generalization and results on C_{novel}. The extended versions of this method are FSOD [42] and FSIS [43], [89]. Those proposals consider all objects in an image as queries and they have a single support set per image instead of per query. However, there exist challenges in FSIS which are not only classification tasks but also how to determine their localization and segmentation. Use an image I_q to query, FSIS returns labels y_i, bounding boxes b_i, and segmentation masks M_i for all objects in I_q that belong to the set of C_{test}.

General framework. Originated from TFA [39] which uses Faster R-CNN [90], MTFA [49] employs a mask prediction branch to return the pixel-wise mask for the segmentation task. In this work, we leverage the architecture of MTFA model [49] based on Mask R-CNN [70] which is a two-stage training and fine-tuning mechanism. We train the first stage of the framework on 80 classes from the COCO dataset. This stage results in the base model weights for the second stage of novel fine-tuning. In the fine-tuning stage, we apply the few-shot technique to learn the novel concepts of camouflaged instances in our proposed CAMO-FS dataset.

Similar to Mask R-CNN, the input images are fed into a feature extractor F consisting of backbone B, RoI Align, RoI feature extractor modules, and a region proposal network. There are three heads specifying three tasks that this scheme supports: a classification head C, a box regression head R, and a new attached mask prediction head M. In the first stage, the network is trained on the base classes C_{base}. Then in the second stage, we froze the backbone network B of the feature extractor F and only perform training on the prediction heads. Thus, only RoI classifier C, box regressor R, and mask predictor M are fine-tuned in the second stage. In Figure 6, there exists a branch called mask predictor M. We apply similarly to Ganea *et al.* [49] by using this two-stage fine-tuning approach. Firstly, the network is trained on base classes with lots of abundant data and then fine-tuning all predictor heads C, R, and M on novel data of K shots for each class.

Not a simple mask predictor M that we use, we enhance the performance of the instance segmentation task by employing the two concepts of instance triplet loss and instance memory storage which are clearly described in the next section. The two improvements are inspired not only by the instance segmentation task in general but also by the camouflaged instance segmentation specifications.

C. FRAMEWORK IMPROVEMENT

One of the characteristics of camouflage instances is the camouflage texture similar to the background. This makes the precise identification of the boundary areas difficult. It is more critical in the context of few-shot learning where the

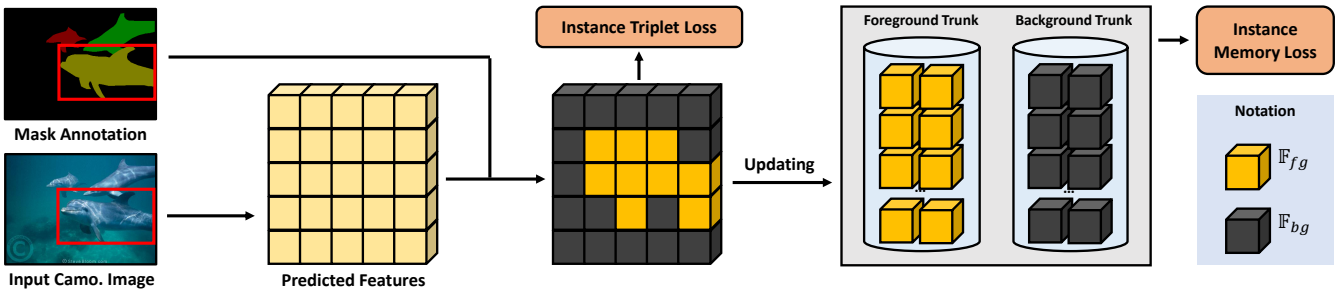


FIGURE 7: Visualization of instance triplet loss and instance memory loss for region proposal.

concepts of a class are represented by only a few samples.

In this work, we thus propose improvements to enhance distinguishable features between background and foreground areas. In particular, we explore two approaches that focus on loss functions. The first one is the triplet loss function which was known as a strong metric to support the network in creating discrimination features between anchor and negative. The second approach is the idea of memory bank, which is used to enhance the distance between foreground and background not only for individual instances but also for each novel class. To this end, our framework is named after FS-CDIS.

To calculate the loss function, we employ the mask annotation for RoI features to collect the $\mathbb{F}_{bg}$ background and $\mathbb{F}_{fg}$ foreground features by location on each RoI. Both $\mathbb{F}_{bg}$ and $\mathbb{F}_{fg}$ for each proposal are used to calculate the respective loss functions which are presented in the following sections.

1) Instance triplet loss

With the idea of enhancing the discrimination between camouflaged instances and their backgrounds, we leverage the power of the triplet loss function [91]. Specifically, we treat the pixels of an object as positive points and the background as negative ones. Accordingly, we force the model to learn the distinguished features among the foreground and background representatives. The more distinguished among features, the better a model can do to detect or segment camouflaged instances. In this way, we highlight the camouflaged instances so that the model is able to recognize them.

For each RoI, we consider the average foreground features $\mathbb{F}_{avg} = \frac{1}{|\mathbb{F}_{fg}|}\sum \mathbb{F}_{fg}$ as anchors with the foreground feature $\mathbb{F}_{fg}$ as positive and the background feature $\mathbb{F}_{bg}$ as negative to apply the triplet loss function [91]. In this way, the model tries to learn to minimize the distance between foreground representatives and maximize the distance between background representatives as shown in Figure 7. We use cosine similarity to calculate the distance instead of Euclidean distance. The loss function is defined as:

$$\begin{aligned} \mathcal{L}_{triplet} &= \max\{d(\mathbb{F}_{avg}, \mathbb{F}_{fg}) - d(\mathbb{F}_{avg}, \mathbb{F}_{bg}) + margin, 0\} \\ d(x,y) &= 1 - \frac{x \cdot y}{\|x\| \cdot \|y\|} \end{aligned} \tag{1}$$

, where *margin* controls the discrimination between foreground and background features. In our experiments, we set *margin* of 0.5.

2) Instance memory storage

The memory bank is designed to store information within a class and the class information is updated during the training. Still, the model can learn information at a global level and has high consistency for each class. On the other hand, storing and updating the features in the memory bank for each iteration during training also creates more variants. By leveraging these advantages, we propose the memory bank for few-shot camouflage instance segmentation. To be specific, we use the memory bank to contain the background and foreground features per each class and make use of features to calculate the discrimination between areas of object and no object in region proposals (shown in Figure 7).

Storing and updating: The memory bank for each class contains 2N features including N of foreground features and N of background features. While the memory bank receives new features, the module concatenates them with existing old features. In case the number of features is greater than the given N features, the memory bank releases the oldest features to maintain the number of features to N. This process updates the features in the memory bank and keeps the quantity of the stored features appropriate to the memory size (also known as the memory capacity).

Sampling: To calculate the loss value, the memory bank has to provide three elements $\mathbb{F}_{fg}$, $\mathbb{F}_{bg}$, and $\mathbb{F}_{general}$. $\mathbb{F}_{fg}$ and $\mathbb{F}_{bg}$ are all foreground and background features that module storing. The $\mathbb{F}_{general}$ is the general foreground feature, and it is created for each class by averaging the $\mathbb{F}_{fg}$.

Let $\mathbb{F}^i_{fg}$ be the i-th foreground feature and τ be a temperature hyper-parameter in [92]. In our experiments, we set τ as 1. The memory loss function for camouflaged instances is introduced as follows:

$$\mathcal{L}_{memory} = -\log \frac{\exp(\mathbb{F}_{general} \cdot \mathbb{F}^i_{fg}/\tau)}{\sum_{j=0}^{|\mathbb{F}_{bg}|} \exp(\mathbb{F}_{general} \cdot \mathbb{F}^j_{bg}/\tau) + \exp(\mathbb{F}_{general} \cdot \mathbb{F}^i_{fg}/\tau)} \tag{2}$$

To this end, the final loss of our training process, which contains an instance triplet loss and memory storage is defined as follows:

TABLE 4: State-of-the-art comparison on CAMO-FS dataset among the baseline model of MTFA [49], Mask RCNN† [70], iFS-RCNN [83], and our proposed methods FS-CDIS with instance triplet loss (ITL) and instance memory storage (IMS). Our performance improves over the utilized baselines.

Model		Novel AP									
Method	Baseline	Instance Segmentation					Object Detection				
		1	2	3	5	Avg.	1	2	3	5	Avg.
MTFA [49]	COCO-80 ResNet-50	2.48	6.67	5.81	6.40	5.34	1.98	6.47	5.82	6.17	5.11
M-RCNN† [70]		4.08	6.79	6.90	8.29	6.52	2.82	5.09	5.46	6.18	4.89
iFS-RCNN [83]		4.17	6.26	5.73	6.38	5.64	3.92	6.06	5.47	6.60	5.51
MTFA [49]	COCO-80 ResNet-101	3.66	6.21	6.16	5.95	5.50	2.93	5.90	5.84	5.84	5.13
M-RCNN† [70]		4.39	7.69	7.94	10.09	7.53	3.03	5.80	6.20	7.79	5.71
iFS-RCNN [83]		4.27	6.55	6.07	7.80	6.17	3.79	6.28	6.01	8.08	6.04
Our performance											
FS-CDIS-ITL	ResNet-101 MTFA	4.46	5.57	6.41	8.48	**6.23**	4.04	7.28	7.49	9.76	**7.14**
FS-CDIS-IMS		5.46	6.95	7.36	9.61	**7.35**	4.50	6.95	7.55	10.36	**7.34**
FS-CDIS-ITL	ResNet-101 M-RCNN	5.73	7.97	8.52	9.92	**8.04**	5.08	7.56	7.85	9.67	**7.34**
FS-CDIS-IMS		5.52	7.84	8.65	9.82	**7.96**	4.92	7.39	7.96	9.52	**7.45**
FS-CDIS-ITL	ResNet-101 iFS-RCNN	5.35	6.01	7.80	9.35	**7.13**	4.71	5.66	7.10	10.36	**6.96**
FS-CDIS-IMS		2.99	6.83	6.14	9.03	**6.25**	2.74	6.39	5.94	8.44	**5.88**

M-RCNN† is Mask R-CNN [70] with sigmoid classifier.

$$\mathcal{L}_{final} = \mathcal{L}_{mrcnn} + \alpha\mathcal{L}_{triplet} + \beta\mathcal{L}_{memory}. \quad (3)$$

Here, the parameter α of $\mathcal{L}_{triplet}$ and β of $\mathcal{L}_{memory}$ are used during the training process to keep the balance between the two loss functions. Details of these functions are mentioned in the following section.

IV. EXPERIMENTS

We first overview the metrics and the experiment settings and the implementation details in Section IV-A and then we evaluate and discuss our improvement on the general framework, as well as ablation study for our core proposed methods in Section IV-B.

A. OVERVIEW

As specified in this work, we utilize the proposed CAMO-FS dataset containing images of camouflaged animals in the wild to establish the evaluation of our baseline and proposed improvement. We follow the concept procedure published in FSOD [39], [42], [43]. We employed the MTFA baseline [49] implemented using Detectron2 framework [93]. The backbone is ResNet-101 [94] with Feature Pyramid Network [95]. The models are trained in two stages: base training and novel fine-tuning stage.

In the first stage of the base phase, we train our model with abundant data from 80 classes with $118K$ images in the `train2017` set of the COCO dataset. The training hyper-parameters of the base phase are set according to Detectron2 settings [93].

In the second stage of the fine-tuning phase, we evaluate the performance of having $K = \{1, 2, 3, 5\}$ shots per each novel class. Specifically, in the 5-shot setting, we train the novel detector on 47 camouflaged classes with 197 images of the CAMO-FS dataset. The training set for other settings is a subset of the 5-shot setting (as presented in Section III-A). The novel phase has a learning rate of 0.00125 inferred from the MTFA configuration. We set the balance parameters $\alpha = 1 \times 10^{-1}$ and $\beta = 1 \times 10^{-2}$ when we train the model with instance triplet loss and instance memory storage, respectively. Other training hyper-parameters of the novel phase are set following TFA [39] settings. Then, the novel models are assessed in a test set including $2,655$ images with $3,107$ instances of 47 camouflaged classes to obtain results. The number of instances for each class is detailed in Table 5. Please visit [39] or [93] for more details on other parameters of both the training and testing phases. Our models are trained and tested on a single GeForce RTX 2080 Ti GPU. The frame per second (FPS) is approximately 15.

TABLE 5: The number of instances for each camouflaged class in the test set of our CAMO-FS.

Class	#ins.	Class	#ins.	Class	#ins.
Shrimp	10	Crab	37	Dolphin	48
Crocodile	10	Snake	54	Turtle	23
Worm	24	Snail	9	Kangaroo	48
Elephant	12	Giraffe	29	Goat	18
Leopard	39	Monkey	35	Deer	21
Sea_Lion	27	Lion	16	Bat	13
Fox	30	Camel	15	Weasel	32
Rabbit	44	Dog	37	Horse	31
Mouse	8	Hedgehog	10	Rhino	35
Tiger	6	Pig	11	Squirrel	19
Sloth	17	Scorpion	7	Bear	19
Octopus	133	Starfish	71	Seahorse	120
Fish	190	Frog	128	Lizard	180
Cat	82	Moth	56	Penguin	67
Spider	125	Insect	228	Bird	363
Body_Paint.	347	Solider	223	**Total**	**3107**

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 [1].

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

TABLE 6: The improvement of our proposed instance triplet loss (ITL) and instance memory storage (IMS) over the baseline MTFA [39]. The best performances are marked in **bold**. # denotes the Number of shots.

#	Method	AP	AP50	AP75	APs	APm	APl	AR1	AR10	ARs	ARm	ARl
Instance Segmentation												
1	**Baseline MTFA**	3.66	5.37	4.09	22.42	4.35	2.01	11.30	13.58	25.97	**12.96**	12.53
	MTFA + ITL	4.46	8.21	4.60	21.33	4.13	**4.01**	12.36	15.04	23.17	9.49	16.67
	MTFA + IMS	**5.46**	**9.20**	**6.17**	**27.79**	**6.20**	**4.01**	17.08	19.99	**29.41**	11.45	20.89
	MTFA + Both	5.02	8.58	5.38	26.29	5.32	3.93	**17.98**	**21.42**	28.13	12.68	**22.64**
2	**Baseline MTFA**	6.21	8.92	7.28	32.64	**7.75**	3.50	18.88	21.12	**35.82**	**15.49**	20.14
	MTFA + ITL	5.57	9.45	6.04	25.83	3.01	5.37	15.67	17.33	26.13	7.37	17.50
	MTFA + IMS	6.95	10.72	7.60	**33.62**	5.73	6.44	20.00	22.15	34.25	13.86	20.92
	MTFA + Both	**7.60**	**11.37**	**8.26**	33.58	5.97	**6.49**	**22.76**	**24.85**	34.02	15.01	**24.50**
3	**Baseline MTFA**	6.16	8.95	6.68	33.74	6.19	5.08	20.25	22.95	36.83	16.31	21.63
	MTFA + ITL	6.41	10.67	6.72	30.39	5.17	5.30	20.69	22.98	31.90	15.69	22.53
	MTFA + IMS	7.36	11.23	8.49	37.03	**6.24**	5.64	**24.40**	**27.69**	38.44	17.02	**26.71**
	MTFA + Both	**7.85**	**11.74**	**9.05**	**37.57**	5.49	**6.69**	24.33	27.42	**39.05**	**17.36**	25.92
5	**Baseline MTFA**	5.95	8.67	6.94	34.71	**6.25**	4.85	21.29	24.42	36.86	**14.51**	24.83
	MTFA + ITL	8.48	13.43	9.80	36.66	5.75	8.04	23.83	26.66	37.03	11.62	25.91
	MTFA + IMS	**9.61**	**14.61**	**11.73**	**38.60**	5.79	**10.40**	**26.65**	**30.37**	39.21	12.26	**30.02**
	MTFA + Both	8.37	13.29	9.45	38.44	5.72	7.60	25.27	29.31	**39.33**	14.19	28.52
Object Detection												
1	**Baseline MTFA**	2.93	5.86	2.20	20.95	4.18	2.03	9.25	10.84	21.74	11.49	8.77
	MTFA + ITL	4.04	8.65	2.98	20.50	4.90	**4.22**	12.89	15.53	20.73	11.45	17.46
	MTFA + IMS	**4.50**	**9.14**	3.45	**22.88**	5.61	3.54	13.14	15.22	**23.14**	8.78	16.33
	MTFA + Both	4.31	8.63	**3.75**	22.15	**6.25**	3.63	**14.67**	**17.47**	22.16	**11.62**	**18.15**
2	**Baseline MTFA**	5.90	8.87	6.83	33.04	9.74	3.10	17.26	19.25	34.04	**15.74**	19.61
	MTFA + ITL	7.28	11.22	8.25	32.31	**10.72**	**6.83**	20.52	22.69	32.34	14.88	23.52
	MTFA + IMS	6.95	10.88	7.75	33.93	7.49	6.81	19.84	22.01	**34.10**	15.04	21.47
	MTFA + Both	**7.60**	**11.58**	**8.78**	**34.13**	8.23	6.70	**23.87**	**25.92**	34.07	15.13	**27.18**
3	**Baseline MTFA**	5.84	8.98	6.29	34.56	7.78	4.31	19.13	21.83	35.80	15.93	21.09
	MTFA + ITL	7.49	11.51	8.23	**38.45**	8.61	6.38	**24.88**	**27.52**	**38.55**	17.66	27.44
	MTFA + IMS	7.55	11.45	8.50	38.07	**9.21**	5.70	24.20	27.29	38.50	**18.10**	**27.56**
	MTFA + Both	**7.94**	**12.07**	**9.11**	37.97	8.96	**6.68**	23.77	26.74	38.32	17.64	26.58
5	**Baseline MTFA**	5.84	9.13	6.04	35.44	8.17	4.22	19.67	22.96	35.94	**14.16**	22.58
	MTFA + ITL	9.76	14.37	11.12	**40.05**	**8.82**	9.89	25.93	29.28	**40.05**	12.53	30.32
	MTFA + IMS	**10.36**	**16.27**	**11.79**	39.32	8.08	**11.36**	**26.34**	**30.30**	39.35	12.37	**30.91**
	MTFA + Both	9.39	14.19	10.42	39.37	8.36	9.28	26.16	30.16	39.50	13.98	30.23

B. RESULTS AND DISCUSSION

State-of-the-art comparison. To prove the effectiveness of our proposed methods, we conducted experiments on our proposed CAMO-FS dataset. We tested with $K = \{1, 2, 3, 5\}$ shots, respectively. Since several recent work have not published their source code [85], [86], we adopted the typical models addressing both detection and instance segmentation tasks to compare with our proposed methods. Table 4 presents the evaluation of the performance of our methods of instance triplet loss and memory storage over our baseline MTFA [49], the model of Mask R-CNN [70] with sigmoid classifier, and the state-of-the-art method iFS-RCNN [83] in the approach of few-shot instance segmentation. We reported experiments on those models and chose the common COCO-80 ResNet-101 as their base model to apply our proposed methods. The details of this decision are declared in the ablation section. In terms of instance segmentation, we improved over MTFA [49], Mask RCNN† [70], and iFS-RCNN [83] by getting average AP values of $6.23\%, 8.04\%, 7.13\%$, respectively thanks to instance triplet loss, and $7.35\%, 7.96\%, 6.25\%$, respectively thanks to instance memory storage. Regarding object detection, our FS-CDIS got average amounts of $7.14\%, 7.34\%, 6.96\%$, respectively with instance triplet loss and $7.34\%, 7.45\%, 5.88\%$, respectively with instance memory storage. The detailed performance of our methods is in Table 4. Despite the limited results, we defeated the very early models on detection and instance segmentation tasks on camouflaged images.

Proposed modules evaluation. In Table 6, we also present the results of the baseline MTFA [49] with its original default configuration along with our proposed improvements. On top of the baseline MTFA [49], we establish fine-tuning configuration on this model by training all heads of classification, box regression, and mask prediction on few-shot novel data. The reported results prove the performance of the proposed instance triplet loss, instance memory storage, and the combination of both loss functions.

In general, our approaches achieve outstanding results in comparison with the baseline. Our improvements surpass MTFA by a remarkable margin. These results manifest the

TABLE 7: Ablation study on the base model with 1-shot results. The best performances are marked in **bold**. "Triplet" stands for Instance Triplet Loss and "Memory" stands for Instance Memory Storage.

Method	Base Model	Segmentation			Detection		
		AP	AP50	AP75	AP	AP50	AP75
Triplet	COCO-80 R-101	**4.46**	**8.21**	**4.60**	**4.04**	**8.65**	**2.98**
	COCO-80 R-50	3.68	6.79	3.81	2.85	6.67	1.65
	COCO-60 R-101	3.87	6.26	3.90	3.37	6.51	2.69
	COCO-60 R-50	2.56	4.25	2.79	2.28	4.13	2.26
Memory	COCO-80 R-101	**5.46**	**9.20**	**6.17**	**4.50**	**9.14**	**3.45**
	COCO-80 R-50	3.87	6.81	3.91	3.40	6.94	2.76
	COCO-60 R-101	2.89	4.50	3.26	2.76	4.66	2.81
	COCO-60 R-50	2.63	4.50	3.02	2.25	4.50	1.65

efficiency of our methods in the context of few-shot camouflaged instance segmentation. Both loss functions enhance the discrimination between foreground and background features which strongly supports the model to segment pixels that belong to the camouflaged animals. Regarding the memory storage and the triplet loss function, the results of the memory loss function are higher than those of the triplet loss function by about 1%. We realize that storing representatives for each class is a crucial element in few-shot learning. This technique not only expands the variants during training but also increases the consistency per class, so thereby model can segment difficult objects better. In these ways, we also improve the corresponding results in camouflage object detection.

In Table 6, our improvements help the model segment animals in various sizes. Specifically, all three metrics including APs, APm, and APl improve in comparison with the baseline model, which demonstrates that our model well segments small, medium, and large animals. This situation also happens in the detection task. When data is very scarce as in a 1-shot or 2-shot setting, the instance triplet loss function has comparative results with the instance memory storage function. However, in the context of 3-shot or 5-shot settings, the instance memory storage demonstrates outstanding efficiency thanks to storing and updating the memory via iterations to create discriminative features on a global level. Actually, our proposed instance triplet loss is designed to differentiate the features among the foreground and background of a single camouflaged instance. Meanwhile, the instance memory storage aims to store features of multiple instances of the same category to enhance the general features. Thus, when combining the two approaches at the same time to train our FS-CDIS, the performance of the model fluctuates but still follows a trend. It can be seen from the reported numbers that with $K = \{1, 2, 3\}$ shots, the results of the combined loss seem to dominate the results of each separate component. However, notably around $K = 5$ shots, the AP, AP50, and AP75 of the instance memory storage yield better performance due to the information increase by more shots.

In terms of quantitative comparison, Figure 8 illustrates the qualitative comparison among the results of 5-shot settings of the baseline MTFA [49] and our proposed methods of Instance Triplet Loss and Instance Memory Storage. We chose to visualize the images with a confidence threshold of 0.5, which released a huge number of predictions with low confidence from the models. The four final rows indicate exemplary cases that either FS-CDIS-ITL or FS-CDIS-IMS can figure out camouflaged instances compared to the baseline. In these cases, our methods seem to be better the baseline although there are some imperfect cases where the prediction mask in the sixth row overlays irrelevant parts of the object (over-segmentation) or the results in the fourth, fifth, and last rows only capture some main parts of objects (under-segmentation). We conjecture the reasons for that is the image contains background clutter and the extremely vague boundary between foreground and background. To alleviate the phenomenon, it is feasible to further apply the post-processing methods on the output segmentation masks.

Base model ablation study. We also conduct ablation experiments on different backbone base models of the COCO settings including general and few-shot concepts. To be detailed, we report the performance of our proposed method of instance triplet loss and instance memory storage over four different backbones. The considered backbones are ResNet-50 and ResNet-101 [94]. The two base datasets are MS-COCO with 80 classes and 60 classes, respectively. Thus, it led to the combination of four different base models (i.e. COCO-80/60 R-101, COCO-80/60 R-50). As can be seen from Table 7, the performance of applying COCO-80 R-101 base weight yields better results among others evaluated on AP, AP@50, and AP@75 in both segmentation and detection tasks. In both cases of our two proposed improvements, the ablation results demonstrate our selection of COCO-80 R-101 is the best among the tested backbones of the base phase. For the segmentation task, we achieve 4.46 and 5.46 of AP reported for triplet loss and memory storage, respectively. For the detection task, we reach 4.04 and 4.50 also of AP, respectively. In summary, the chosen backbone of the base weight presents a higher performance of around 1% to 2% of evaluated on common metrics as in the table. To be explained, the base from COCO-80 contains more semantic concepts in comparison with the COCO-60 base, which leads to the higher performance reported. Note that all the results in this ablation section are reported for the 1-shot setting.

Ablation on instance triplet loss component. In terms of the instance triplet loss described in Eq. 1, we establish ablation experiments to evaluate the performance of the model with different configurations of margin and α value when the instance memory storage component is disabled (i.e. $\beta = 0$). To this end, we set up the margin varying from 0 to 1, with a step of 0.25. For the α ratio of the loss function (Eq. 3), we check out $\alpha = \{1, 1 \times 10^{-1}, 1 \times 10^{-2}\}$. To be enhanced, the margin value indicates how distinguished foreground and background features are. Meanwhile, the α controls the effect of the instance triplet loss on the total loss of the framework. Table 8 presents the evaluation of both detection and segmentation issues in 1-shot manner. As can be inferred from the table, the effect of α decides which margin should be selected for the triplet loss. With $\alpha = 1$ meaning we keep the original ratio of the loss, the segmentation result in 1-shot

TABLE 8: Ablation study on the margin and the α ratio of the instance triplet loss in 1-shot settings. The best performances are marked in **bold**.

AP	Segmentation					Detection				
α \ Margin	**0**	**0.25**	**0.50**	**0.75**	**1.00**	**0**	**0.25**	**0.50**	**0.75**	**1.00**
1	3.89	4.50	3.92	**5.16**	4.43	3.34	3.65	3.22	**4.22**	3.68
1×10^{-1}	**4.82**	4.74	4.46	4.58	**4.57**	**4.36**	**4.27**	4.04	4.16	**3.79**
1×10^{-2}	4.29	**4.74**	**4.69**	4.46	4.39	4.02	3.97	**4.24**	4.06	3.71

TABLE 9: Ablation study on the capacity of the instance memory storage in 1-shot settings. The best performances are marked in **bold**.

Capacity	Segmentation			Detection		
	AP	**AP50**	**AP75**	**AP**	**AP50**	**AP75**
32	4.56	7.30	5.02	3.85	7.72	2.91
64	4.51	7.67	4.49	3.94	8.37	2.84
128	4.53	7.55	4.87	4.13	7.98	3.62
256	4.56	7.50	5.02	4.01	8.22	3.39
512	**4.76**	7.57	**5.37**	**4.48**	8.25	**4.44**
1024	4.72	**8.06**	5.20	4.14	**8.45**	3.79

TABLE 10: Ablation study on the β ratio of the instance memory loss (Eq. 3) in 1-shot settings. The best performances are marked in **bold**.

β	Segmentation			Detection		
	AP	**AP50**	**AP75**	**AP**	**AP50**	**AP75**
1×10^{-1}	3.36	6.58	2.91	3.69	8.02	2.90
1×10^{-2}	4.57	8.02	4.74	3.73	7.78	2.76
1×10^{-3}	4.51	7.15	4.67	3.87	7.16	3.51
1×10^{-4}	**5.12**	**8.71**	**5.54**	**4.58**	**9.23**	**3.69**
1×10^{-5}	4.44	7.58	3.89	4.06	7.99	3.63

setting yields the highest performance of 5.16% mAP with a 0.75 margin value. Meanwhile, the detection result gets the highest performance of 4.36% with $\alpha = 1 \times 10^{-1}$ and zero margin. This table offers a better understanding of the impact of α and the margin over the total performance.

Ablation on instance memory storage component. As for the instance memory storage, as introduced in Eq. 2, and Eq. 3, there are several parameters that need analyzing, listed as the amount of capacity in the memory storage and the β ratio controlling the effect of the memory storage loss in the total loss. Table 9, and Table 10 present the ablation experimental results of those issues when the instance triplet loss function is disabled (i.e. $\alpha = 0$), respectively. In terms of the capacity of the memory storage, we establish experiments on a range of memory capacity of 2^i where $i = \{5, 6, 7, 8, 9, 10\}$. The reported results figure out that the performance on both segmentation and detection tasks increases with a larger capacity of memory storage. To be detailed, with a capacity of 512, the mAP metric achieves the highest value among configurations, i.e. 4.76% and 4.48% for segmentation and detection, respectively. Empirically, we select 512 to be the suitable capacity of the memory storage, not the largest. To this end, the larger capacity can confuse the model in the process of learning when retrieving information in such a large memory bank. Besides, Table 10 expresses the effectiveness of the memory loss to the total loss function. As can be inferred, $\beta = 1 \times 10^{-4}$ gives the best performance evaluated on mAP, AP50, and AP75 among all configurations.

V. CONCLUSION

In this work, we investigated the interesting yet challenging problem of few-shot learning for camouflaged animal detection and segmentation. We first collect a new dataset, CAMO-FS, for benchmarking purposes. We then propose a novel method to efficiently detect and segment the camouflaged animals in the images. In particular, we introduce the instance triplet loss and the instance memory storage. The extensive experiments demonstrated that our proposed method achieves state-of-the-art performance on the newly constructed dataset. We expect our work will encourage more research work in this field. In the future, we would like to extend our work with more shots for new classes. In addition, we aim to improve the computational model by taking the context into consideration.

ACKNOWLEDGMENT

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

REFERENCES

[1] S. Singh, C. Dhawale, and S. Misra, "Survey of object detection methods in camouflaged image," *IERI Procedia*, vol. 4, pp. 351 – 357, 2013.
[2] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *CVIU*, vol. 184, pp. 45–56, 2019.
[3] Z. Zhou, B. Zhang, and X. Yu, "Immune coordination deep network for hand heat trace extraction," *Infrared Physics & Technology*, vol. 127, p. 104400, 2022.
[4] 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.
[5] 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.
[6] X. Wang, R. Ma, X. Xu, P. Niu, and H. Yang, "Non-linear statistical image watermark detector," *Applied Intelligence*, vol. 53, no. 23, pp. 29 242–29 266, 2023.
[7] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," in *ICCV*, 2021.
[8] E. L.-M. Pia Bideau, "It's moving! a probabilistic model for causal motion segmentation in moving camera videos," in *ECCV*, 2016.
[9] H. Lamdouar, C. Yang, W. Xie, and A. Zisserman, "Betrayed by motion: Camouflaged object discovery via motion segmentation," in *ACCV*, November 2020.
[10] P. Skurowski, H. Abdulameer, J. Baszczyk, T. Depta, A. Kornacki, and P. Kozie, "Animal camouflage analysis: Chameleon database," *Unpublished Manuscript*, 2018.

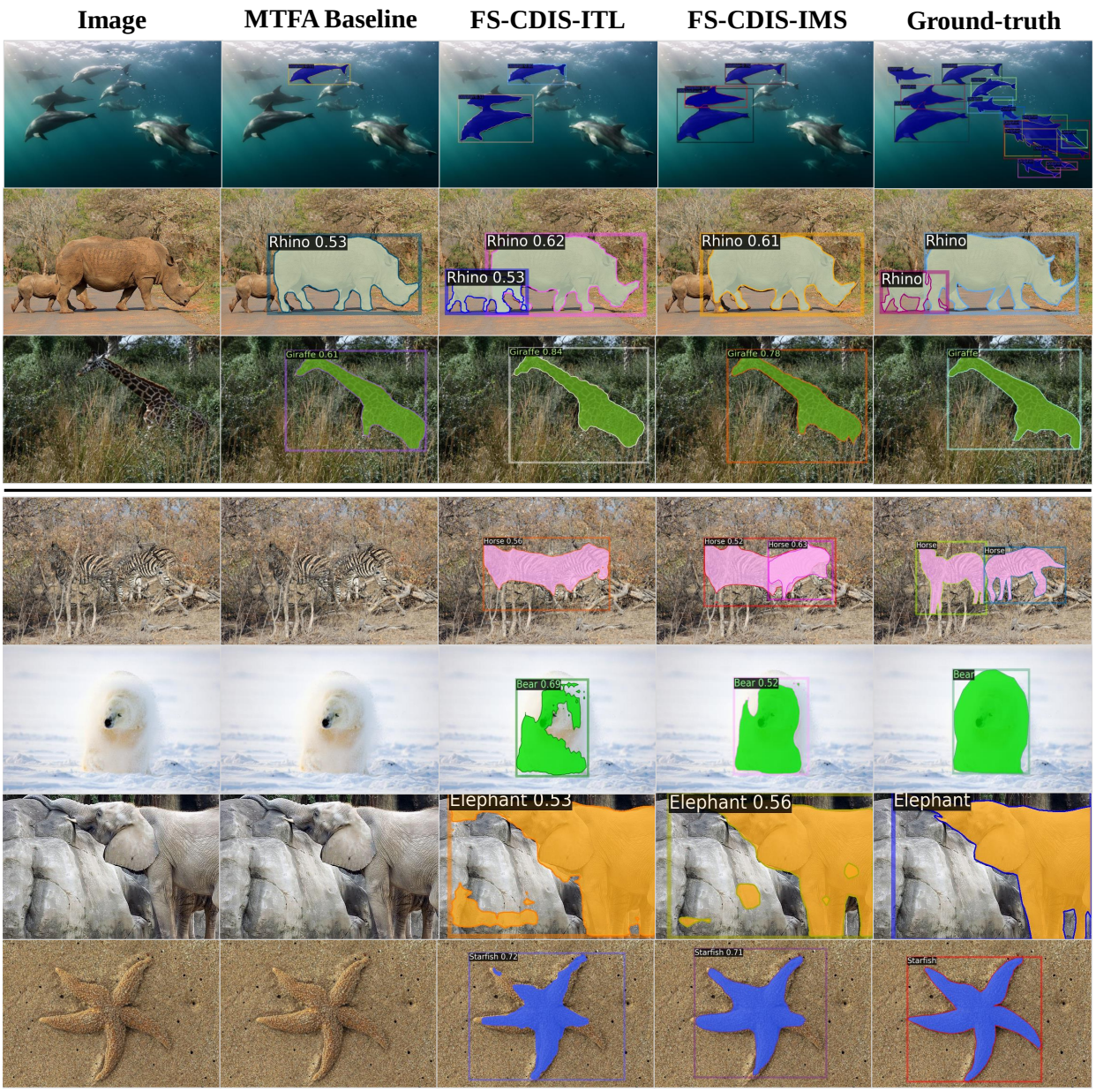


FIGURE 8: Qualitative comparison among the selected baseline MTFA [49] and our proposed methods. The results are from 5-shot settings. Predicted images are visualized with a confidence threshold of 0.5, which released a huge number of predictions with low confidence from the models. The four final rows indicate exemplary cases that either FS-CDIS-ITL or FS-CDIS-IMS can figure out camouflaged instances compared to the baseline.

[11] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2777–2787.

[12] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 591–11 601.

[13] T.-N. Le, Y. Cao, T.-C. Nguyen, M.-Q. Le, K.-D. Nguyen, T.-T. Do, M.-T. Tran, and T. V. Nguyen, "Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite," *IEEE Transactions on Image Processing*, vol. 31, pp. 287–300, 2022.

[14] C. Kervrann and F. Heitz, "A markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics," *IEEE TIP*, vol. 4, no. 6, pp. 856–862, 1995.

[15] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *IJCV*, vol. 70, no. 2, pp. 109–131, 2006.

[16] X. Li and H. Sahbi, "Superpixel-based object class segmentation using conditional random fields," in *ICASSP*, 2011, pp. 1101–1104.

[17] L. Sulimowicz, I. Ahmad, and A. Aved, "Superpixel-enhanced pairwise conditional random field for semantic segmentation," in *ICIP*, 2018, pp. 271–275.

[18] M. Galun, E. Sharon, R. Basri, and A. Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements," in *ICCV*, Oct 2003, pp. 716–723.

[19] L. Song and W. Geng, "A new camouflage texture evaluation method based on wssim and nature image features," in *International Conference on Multimedia Technology*, Oct 2010, pp. 1–4.

[20] F. Xue, C. Yong, S. Xu, H. Dong, Y. Luo, and W. Jia, "Camouflage

performance analysis and evaluation framework based on features fusion," *Multimedia Tools and Applications*, vol. 75, pp. 4065–4082, 2016.
[21] Y. Pan, Y. Chen, Q. Fu, P. Zhang, and X. Xu, "Study on the camouflaged target detection method based on 3d convexity," *Modern Applied Science*, vol. 5, no. 4, p. 152, 2011.
[22] Z. Liu, K. Huang, and T. Tan, "Foreground object detection using top-down information based on em framework," *IEEE TIP*, vol. 21, no. 9, pp. 4204–4217, Sept 2012.
[23] A. W. P. Sengottuvelan and A. Shanmugam, "Performance of decamouflaging through exploratory image analysis," in *ICETET*, 2008, pp. 6–10.
[24] J. Yin, Y. Han, W. Hou, and J. Li, "Detection of the mobile object with camouflage color under dynamic background based on optical flow," *Procedia Engineering*, vol. 15, pp. 2201 – 2205, 2011.
[25] J. Gallego and P. Bertolino, "Foreground object segmentation for moving camera sequences based on foreground-background probabilistic models and prior probability maps," in *ICIP*, Oct 2014, pp. 3312–3316.
[26] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 997–13 007.
[27] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 071–10 081.
[28] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8772–8781.
[29] 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.
[30] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflage objects by texture-aware interactive guidance network," in *AAAI*, 2021.
[31] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.
[32] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
[33] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4080–4088.
[34] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, and C. Gagné, "Matching feature sets for few-shot image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9014–9024.
[35] J. Rajasegaran, S. Khan, M. Hayat, F. S. Khan, and M. Shah, "Self-supervised knowledge distillation for few-shot learning," *arXiv preprint arXiv:2006.09785*, 2020.
[36] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 573–10 582.
[37] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 836–10 846.
[38] L. Xing, Y. Ma, W. Cao, S. Shao, W. Liu, and B. Liu, "Rethinking few-shot remote sensing scene classification: A good embedding is all you need?" *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[39] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *ICML*, 2020.
[40] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8681–8690.
[41] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, "Semantic relation reasoning for shot-stable few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[42] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *ICCV*, 2019.
[43] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *ICCV*, 2019.
[44] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[45] G. Zhang, Z. Luo, K. Cui, and S. Lu, "Meta-detr: Image-level few-shot object detection with inter-class correlation exploitation," *arXiv preprint arXiv:2103.11731*, 2021.
[46] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5321–5330.
[47] A. Bulat, R. Guerrero, B. Martinez, and G. Tzimiropoulos, "Fs-detr: Few-shot detection transformer with prompting and without re-training," in *ICCV*, 2023, pp. 11 793–11 802.
[48] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proceedings of (CVPR)*, 2021.
[49] D. A. Ganea, B. Boom, and R. Poppe, "Incremental few-shot instance segmentation," in *Proceedings of (CVPR)*, June 2021, pp. 1185–1194.
[50] A.-K. N. Vu, T.-T. Do, N.-D. Nguyen, V.-T. Nguyen, T. D. Ngo, and T. V. Nguyen, "Instance-level few-shot learning with class hierarchy mining," *TIP*, 2023.
[51] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[52] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4650–4666, 2022.
[53] C. Lang, G. Cheng, B. Tu, and J. Han, "Few-shot segmentation via divide-and-conquer proxies," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 261–283, 2024.
[54] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 581–23 591.
[55] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7352–7362.
[56] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[57] A. Li and Z. Li, "Transformation invariant few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[58] J. Li, Y. Zhang, W. Qiang, L. Si, C. Jiao, X. Hu, C. Zheng, and F. Sun, "Disentangle and remerge: Interventional knowledge distillation for few-shot object detection from a conditional causal perspective," in *AAAI*, vol. 37, no. 1, 2023, pp. 1323–1333.
[59] J. Xu, H. Le, and D. Samaras, "Generating features with increased crop-related diversity for few-shot object detection," in *CVPR*, 2023, pp. 19 713–19 722.
[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[61] L. Song and W. Geng, "A new camouflage texture evaluation method based on wssim and nature image features," in *2010 International Conference on Multimedia Technology*. IEEE, 2010, pp. 1–4.
[62] P. Siricharoen, S. Aramvith, T. Chalidabhongse, and S. Siddhichai, "Robust outdoor human segmentation based on color-based statistical approach and edge combination," in *The 2010 International Conference on Green Circuits and Systems*. IEEE, 2010, pp. 463–468.
[63] M. Galun, E. Sharon, R. Basri, and A. Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements." in *ICCV*, vol. 3, 2003, p. 716.
[64] C. Kavitha, B. P. Rao, and A. Govardhan, "An efficient content based image retrieval using color and texture of image sub-blocks," *International Journal of Engineering Science and Technology (IJEST)*, vol. 3, no. 2, pp. 1060–1068, 2011.
[65] F. Xue, C. Yong, S. Xu, H. Dong, Y. Luo, and W. Jia, "Camouflage performance analysis and evaluation framework based on features fusion," *Multimedia Tools and Applications*, vol. 75, no. 7, pp. 4065–4082, 2016.

[66] J. Y. Y. H. W. Hou and J. Li, "Detection of the mobile object with camouflage color under dynamic background based on optical flow," *Procedia Engineering*, vol. 15, pp. 2201–2205, 2011.
[67] N. Price, S. Green, J. Troscianko, T. Tregenza, and M. Stevens, "Background matching and disruptive coloration as habitat-specific strategies for camouflage," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.
[68] T.-N. Le, V. Nguyen, C. Le, T.-C. Nguyen, M.-T. Tran, and T. V. Nguyen, "Camoufinder: Finding camouflaged instances in images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 16 071–16 074.
[69] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.
[70] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988.
[71] Y. Xiao and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," in *European conference on computer vision*. Springer, 2020, pp. 192–210.
[72] G. Han, Y. He, S. Huang, J. Ma, and S.-F. Chang, "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3263–3272.
[73] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 780–789.
[74] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *ECCV*, 2020.
[75] S. Zhang, L. Wang, N. Murray, and P. Koniusz, "Kernelized few-shot object detection with efficient integral aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 207–19 216.
[76] W. Zhang and Y.-X. Wang, "Hallucination improves few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 008–13 017.
[77] S. Khandelwal, R. Goyal, and L. Sigal, "Unit: Unified knowledge transfer for any-shot object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[78] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[79] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning." in *BMVC*, vol. 3, no. 4, 2018.
[80] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
[81] O. Saha, Z. Cheng, and S. Maji, "Ganorcon: Are generative models useful for few-shot segmentation?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 9991–10 000.
[82] Z. Tian, X. Lai, L. Jiang, S. Liu, M. Shu, H. Zhao, and J. Jia, "Generalized few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 563–11 572.
[83] K. Nguyen and S. Todorovic, "ifs-rcnn: An incremental few-shot instance segmenter," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7010–7019.
[84] B.-B. Gao, X. Chen, Z. Huang, C. Nie, J. Liu, J. Lai, G. Jiang, X. Wang, and C. Wang, "Decoupling classifier for boosting few-shot object detection and instance segmentation," in *NeurIPS 2022*, 2022.
[85] 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.
[86] H. Wang, J. Liu, Y. Liu, S. Maji, J.-J. Sonke, and E. Gavves, "Dynamic transformer for few-shot instance segmentation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2969–2977.
[87] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.
[88] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4367–4375.
[89] Z. Fan, J.-G. Yu, Z. Liang, J. Ou, C. Gao, G.-S. Xia, and Y. Li, "Fgn: Fully guided network for few-shot instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9172–9181.
[90] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.
[91] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks." in *Bmvc*, vol. 1, no. 2, 2016, p. 3.
[92] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
[93] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.
[94] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[95] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

THANH-DANH NGUYEN is currently a researcher at the Laboratory of Multimedia Communications, University of Information Technology, VNU-HCM. He received his M.Sc. and B.Sc. degree (Hons.) in computer science from University of Information Technology, VNU-HCM, in 2024 and 2021, respectively. His research interests are computer vision and deep learning.

ANH-KHOA NGUYEN VU received M.Sc. and B.Sc. degrees in computer science from University of Information Technology, VNU-HCM in 2023 and 2021, respectively. His current research interests include few-shot learning, computer vision, and deep learning.

NHAT-DUY NGUYEN is a former researcher at the Laboratory of Multimedia Communications, University of Information Technology, VNU-HCM. He obtained his M.Sc. degree and B.Sc degree (Hons.) from University of Information Technology, VNU-HCM. His research interests are computer vision and machine learning, especially in object detection and segmentation. He also worked as a lecturer at the faculty of computer science. He is currently a software engineer at the FPT corporation.

VINH-TIEP NGUYEN is currently a lecturer at University of Information Technology, Head of the Laboratory of Multimedia Communications (MM-Lab), VNU-HCM. He obtained his Ph.D. degree from University of Information Technology, VNU-HCM, and M.Sc. degree from University of Science, VNU-HCM in a co-program with John von Neumann Institute, VNU-HCM. His research interests are computer vision and machine learning. He has a great passion for transferring knowledge and research skills to his students.

THANH DUC NGO is a lecturer in the Faculty of Computer Science, University of Information Technology, VNU-HCM where he has been since 2014. He completed his Ph.D. at The Graduate University for Advanced Studies (SOKENDAI) in 2013. His research interests lie in the areas of Computer Vision and Multimedia Content Analysis.

THANH-TOAN DO is currently a Senior Lecturer at the Department of Data Science and AI, Faculty of Information Technology, Monash University. In 2012, he obtained his Ph.D. in computer science at the French National Institute for Research in Computer Science and Control (INRIA), under the supervision of Professors Laurent Amsaleg, Ewa Kijak, and Teddy Furon. From 2013 to 2016, he was a Research Fellow with Professor Ngai-Man Cheung at the Singapore University of Technology and Design. From 2016 to 2018, he was a Research Fellow with Professor Ian Reid at the Australian Centre for Robotic Vision (ACRV) and the University of Adelaide. From 2018 to 2020, he was a Lecturer at the Department of Computer Science, University of Liverpool. His research interests include Computer Vision and Machine Learning. Some particular research problems are Visual Search, Metric Learning, and Compact Deep Learning.

MINH-TRIET TRAN obtained his B.Sc., M.Sc., and Ph.D. degrees in computer science from University of Science, VNU-HCM, in 2001, 2005, and 2009. He joined the University of Science, VNU-HCM, in 2001. His research interests include cryptography and security, computer vision and machine learning, and human-computer interaction. He was a visiting scholar at National Institutes of Informatics (NII, Japan) in 2008, 2009, and 2010, and at University of Illinois at Urbana-Champaign (UIUC) in 2015-2016. He is currently the Vice Rector of University of Science, VNU-HCM. He is also the Head of Software Engineering Laboratory and the Deputy Head of Artificial Intelligence Laboratory, University of Science, VNU-HCM. He is a member of the Management Board of Vietnam Information Security Association (South Branch) and also a member of the Executive Committee of ICT Program for Smart Cities (2018-2020) of Ho Chi Minh City.

TAM V. NGUYEN is an Associate Professor at Department of Computer Science, University of Dayton. He received his Ph.D. degree from the National University of Singapore in 2013. His research topics include artificial intelligence, computer vision, deep learning, multimedia content analysis, and mixed reality.He is one of the pioneers of camouflage analysis research. He has authored and co-authored more than 120 papers with more than 3,000 citations. His H-index is 29. He is an IEEE Senior Member.