Enhancing Camouflaged Object Segmentation in Limited Data Setting

Vinitra Muralikrishnan
UMass Amherst
vmuralikrish@umass.edu

Rahul Saxena
UMass Amherst
rahulsaxena@umass.edu

Mustafa Chasmai
UMass Amherst
mchasmai@umass.edu

Abstract

The essence of image segmentation and detection is to find objects that stand-out or are different from their background, which is also how humans often imagine "objects". But what if an object does not want to be found? Camouflaged objects tend to blend-in with their surroundings by using similar colors and textures. Natural selection has allowed many species to evolve and use sophisticated camouflage mechanisms to avoid detection by predators. Soldiers use camouflage to move covertly and avoid detection by enemies. We address the problem of detecting these objects that expressly wish to avoid detection. Camouflaged objects are abundant in nature and can be extremely challenging to detect, even for humans. Advancements on this problem can help real-world applications like search and rescue operations, tracking aquatic species, ecological surveys, and defect detection. However, unlike generic object detection datasets, benchmark datasets for Camouflaged Object Detection (COD) are relatively smaller and may lack diversity. In this project, we aim to explore various methods to alleviate this limitation and catalyze future research in this area. The source code for this project is available here.

1. Introduction

The difficulty in detecting and identifying camouflaged objects comes from the inherent nature of the human eye attuned to catch conspicuous objects first. Some animals have evolved to take advantage of this and blend themselves into the background with expert command over their body and leveraging the environment. Therefore, they are difficult to identify even for humans. The way humans generally try to find such objects is by employing a search and detect tactic, spending a good amount of time zooming and trying to catch anomalies in the image. Even then it's not perfect. There can be cases where the image has multiple objects, but only one can be detected and identified. Can *you* find the camouflaged objects[1] in images at the top of Fig 1? The difficulty we faced while finding these objects was what made us interested in pursuing this project.

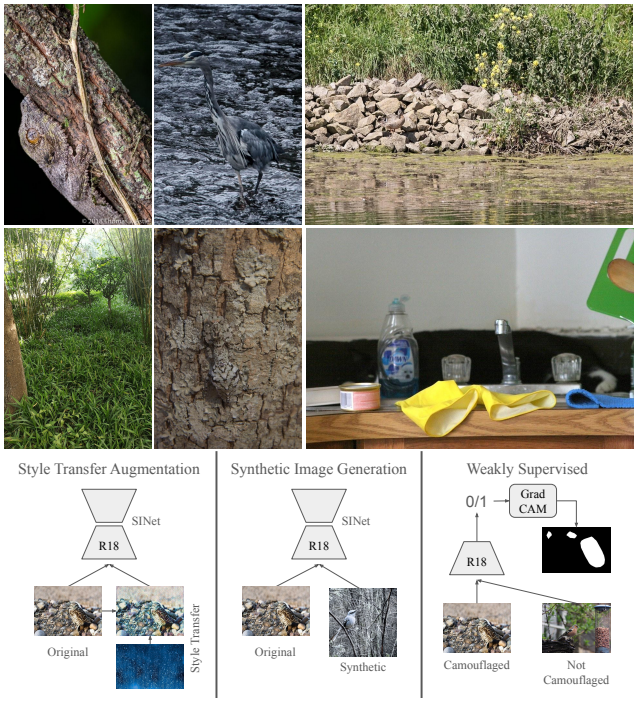

Figure 1. **Overview**. We explore three approaches to either increase training data or reduce annotation effort. Camouflaged images with style transferred from some natural textures can be a good augmentation. Generating synthetic data can also help increase training data. A weakly supervised approach where the annotator only has to provide binary class labels (camouflage or not) can greatly reduce annotation effort.

When a task is difficult for humans, it poses a bigger challenge for a machine-learning solution. To even obtain trainable data for segmenting camouflage images, we would require a human annotator to provide labels for such images. Obtaining good-quality annotations is challenging for camouflaged image segmentation, which is likely a big reason why existing datasets are considerably smaller than other tasks. COD10K [3], is (to the best of our knowledge) the biggest existing benchmark, containing only around

[1]Hints: (top right) There are multiple birds. (bottom left) There is a small black dog in the grass. (bottom right) A cat spans the entire width.

3000 images of camouflaged objects. In contrast, COCO [8], a segmentation benchmark for common objects, has over 200K labelled images. Segment Anything Model [6] was trained on over 11M images containing more than 1B annotated objects. While SAM has proved useful for several problems in the segmentation space, it's not perfect, especially in medical image segmentation. In such spaces, getting huge amounts of annotated data is difficult, therefore creating a need for models that can perform segmentation with weak supervision. Our project will explore ways to improve performance in a limited data setting.

1.1. Background

The ability of animals to blend themselves into their surroundings is well studied in biology but research interest in it is quite nascent in the machine learning discipline. To train models to catch objects quite adept at concealment, naturally, we studied the different ways animals camouflage themselves. Animals are able to camouflage themselves using two primary tactics: pigments and physical structures. Some species have natural, microscopic pigments, known as biochromes, which absorb certain wavelengths of light and reflect others. Species with biochromes appear to change colours. Other species have microscopic physical structures that act like prisms, reflecting and scattering light to produce a colour that is different from their skin. The polar bear, for instance, has black skin. Its translucent fur reflects the sunlight and snow of its habitat, making the bear appear white. Our intuition was what if we try to replicate this process of light absorption and generate synthetic data to train the models? A common approach to working with limited data is to use augmentation. While common augmentations like random cropping, jitter and colour transformations are readily available, they may not introduce enough diversity. In this work, 2 of the methods we propose are data augmentation techniques, one of which tries to replicate the camouflage process employed by animals and get synthetic data.

We imagined an augmentation that takes, say an owl camouflaged in a rocky terrain and generates an owl camouflaged in a beach or the rain. To do this, we use style transfer (explained in detail in section 4.1) to transfer the environment "styles" from some naturally occurring textures. An alternate approach is to generate synthetic data directly. Challenges here include the ability to control the generation process and the quality of automatic annotations. Finally, instead of increasing available data, we can try to reduce annotation effort to allow scaling the benchmarks. Weakly supervised methods, relying only on class labels (camouflaged or not) can be a good direction to explore.

A realistic goal for the task would be to beat the Segment Anything Model (SAM) [6]. While our proposed methods improved performance over the benchmark SOTA SINet [3], we observe that there is still a significant gap to beat SAM. We hope these explorations will increase interest in this field and catalyse future research. A good camouflaged object segmentation model can help applications like ecological surveys, search-and-rescue and defect detection. The vision community would also benefit from a challenging benchmark where the typical segmentation challenges like occlusion and small objects are compounded greatly.

2. Related Work

Research into camouflaged object detection has a rich history in biology and the potential impact of tackling this is tremendous. However, within the domain of computer vision, this area remains relatively under-explored compared to other forms of object detection, such as Generic Object Detection [2, 8] and Salient Object Detection [1]. This discrepancy can be partly attributed to the inherent complexity of the task and the scarcity of expert-annotated data. Interest in this space was kickstarted with the contributions of CamouflagedAnimals [10] and CAMO [7], and found momentum with the contribution of the COD10K dataset [3]. CamouflagedAnimals [10] includes videos of camouflaged animals, which are difficult to identify in still frames, but immediately pop out as they start to move. CAMO [7] consists of 1250 images and was the first work dedicated primarily to camouflaged objects. MirrorNet [16] improved over existing approaches on CAMO [7] by fusing predictions from mirrored data. COD10K [3] greatly increased the scale of this problem by constructing a dataset $\sim$ 10 times bigger. They also propose SINet, which mimics the human receptive field and the "search and identification" stages of predators. Many animals are adept at using the environment to blend and conceal themselves to avoid detection. The role of background and textures in this concealment has been studied and explored by Ren el al., 2021 [11]. Xiang el al., 2022 [15] explore the contribution of depth and attempts to use them for detection.

3. SINet Backbone

We used a refined SINet [3] architecture tailored for efficient camouflaged object segmentation under limited data setting. We modified the standard SINet architecture, which originally employed the ResNet-50 backbone, by integrating the lighter ResNet-18 architecture to reduce computational demands while preserving effective performance for object detection tasks. This modification was particularly advantageous for processing large datasets like COD10K [3] and style transfer-enhanced images where computational efficiency was crucial.

Key components of the SINet backbone:

- **Search Module (SM):** This module is inspired by the

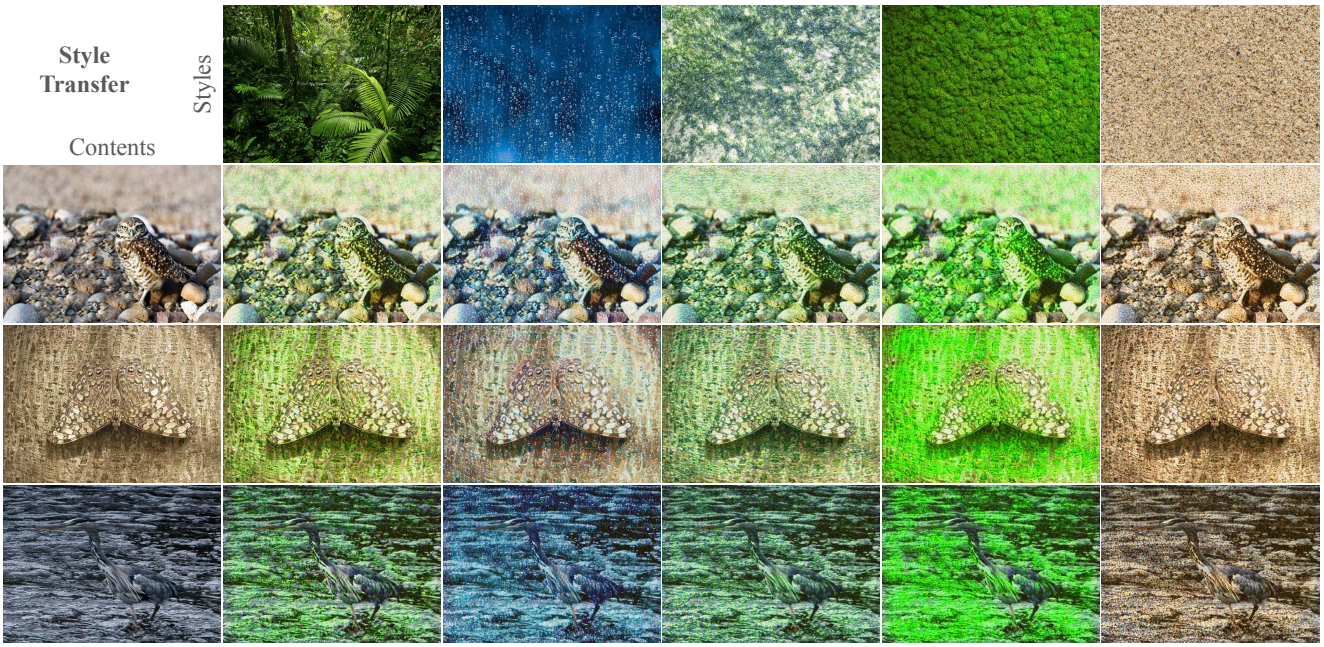

Figure 2. **Style-Transfer as Augmentation.** We consider a few images from the COD10K [3] dataset, and apply style transfer on them with a few natural-like textures. We use a VGG-19 model pre-trained on ImageNet for neural style transfer.

initial phase of hunting in the animal kingdom, where the predator searches for potential prey. It leverages a series of receptive fields (RF) that mimic the human visual system's structure. These receptive fields help to highlight regions in the visual field where camouflaged objects might be located. The RF component in SINet includes several branches, each designed to capture different scales and details, enhancing the detection capabilities of the network.

- **Identification Module (IM):** Following the detection of a potential camouflaged object by the SM, the IM precisely identifies and segments the object. It utilizes a Partial Decoder Component (PDC) that integrates multi-level features processed from the SM. This component refines the detection by focusing on detailed characteristics of the object, ensuring accurate segmentation.

SINet is particularly effective because it processes and integrates multi-scale features, which are crucial for detecting objects that blend seamlessly with their background. It demonstrates superior performance across several metrics when tested on various datasets, indicating its robustness and reliability in detecting camouflaged objects under different conditions.

3.1. ResNet-18

ResNet-18 serves as the backbone for our modified SINet due to its optimal balance between speed and computational efficiency, essential for scenarios with limited computational resources. It is a streamlined variant of the deeper ResNet models, maintaining robust performance with reduced complexity.

The main characteristics of ResNet-18 that make it particularly suited to our application are delineated as follows:

- **Residual Blocks:** Central to ResNet-18 are its residual blocks, which mitigate the vanishing gradient problem—a common challenge in training deep networks. Each block contains two layers connected by skip connections, allowing direct transmission of inputs across layers to preserve signal strength and training stability.

- **Architecture layout:** The network comprises 18 layers, including convolutional layers, batch normalization, ReLU activations, and culminates in a fully connected layer. This arrangement efficiently captures and processes complex image features.

- **Global Average Pooling:** Positioned before the final dense layer, global average pooling reduces spatial dimensions by averaging out feature maps, decreasing the model's parameter count and helping to avert overfitting while lowering computational demands.

Through these modifications and features, we ensure that our modified SINet framework is well-suited for efficient camouflaged object detection in resource-limited settings.

3.2. Balanced Loss

We observed that the objects in the COD10K datasets tend to take up small portions of the image. Looking at semantic segmentation as simply pixel-level classification, this means there is a class-imbalance in the data. On an average, around 90% of an image is background. To mitigate this imbalance, we borrow a trick commonly used in classification, and apply class balancing to the segmentation loss. In particular, we use a new loss $\mathcal{L}_{bal}$ that assigns a weightage of 1 for background pixels and 10 for foreground pixels in each image. $\mathcal{L}_{bal}$ can be used with the original data as well as with style-transferred images and synthetic data. We included the $\mathcal{L}_{bal}$ into our training, and applied it to both original and augmented datasets.

4. Proposed Methods

Our work would be focused towards exploring different ways to improve performance of a baseline segmentation model with a Resnet-18 [5] backbone. We also plan to conduct a study on leveraging generative models and explore whether it can be used to reduce the gap between the COD datasets and generic image segmentation datasets.

4.1. Style-Transfer as Augmentation

The existing benchmark datasets like COD10K [3] has a great label and object diversity but it is still quite imbalanced at the fine-grained level. To further improve the models become more robust to this task, we built a data augmentation pipeline by generating images using Neural Style Transfer [4]. The idea here is to replicate one of the most common ways animals try to conceal themselves - to blend themselves into the background, either by matching the colours of their body to the background or smartly employing the tactic of blending in the environment using the natural textures of their body. Replicating this technique of camouflage, involved using images from COD10K, transformed into different nature-like backgrounds and textures, generating more camouflage scenarios. The generated images may or may not be realistic. For example, the owl example in 2 is not a realistic example of camouflage that happens in nature with respect to owl. But as the goal is to train model to be more robust towards detecting concealed objects, this example would also be useful. This pipeline can be applied to other datasets of concealed object detection, further improving the robustness towards detection in diverse backgrounds and environments. Fig 2 presents some samples taken from COD10K.

4.2. Weakly Supervised Segmentation

We explored the problem in a weakly supervised setting because of the connection it has to other problems such as anomaly and defect detection. The scarcity of annotated data presents a hurdle towards finding a satisfactory solution in these problem spaces. A solution in the weakly supervised setting is therefore potentially the most impactful if found.

We started off our exploration in this setting by using a basic binary classifier and visualising what it "sees" in these images with camouflaged objects. We train a Resnet-18 model to do binary classification and predict whether an image consists of a camouflaged object or not. This classifier is trained on the 6000 images from the training set (3000 containing camouflaged objects, 3000 containing non-camouflaged objects). It achieved about 90% accuracy. We then proceed to draw up a class-activating map using Grad-CAM [13]. This is an extremely weak supervision which, not surprisingly, isn't able to match up to the performance of the other previous works or the other methods proposed in this work. But it helps give insight into what existing models see in these images. Other extensions in this space are training a multi-class classifier to identify the specific species that are camouflaged, using a vision transformer as the base classifier or using few of the annotated masks from the dataset to strengthen the supervision.

4.3. Synthetic Data

We wish to obtain additional images that have camouflaged objects. Authors of COD10K [3] downloaded images from websites like Flicker and then manually annotated them. This assumes the availability of a filter to identify which images contain camouflaged objects. Recent advances in Generative AI have paved the way to a fine control over generated images. In particular, we use a variant of Stable-Diffusion [12] that generates images with text prompts. Although the quality and benefits of AI-generated images to train new AI is often debatable, this is an interesting direction to explore.

We experimented with a few different prompts and choose to use a fixed prompt template "a photo of a _ colored _ in _ forest with matching texture", where we vary the color of the object and the forest (same color for the two) and we repeat this for all the animals present in the COD10K dataset. The range of animals as well as the different colors are intended to give a good diversity to the generated data. Some cherry-picked generated images can be seen in Fig 3. Considering the relatively smaller size of the datasets here and the difficulty in annotation, generating synthetic data, if helpful, could bridge the gap between camouflaged and generic image segmentation datasets like COCO [8] and Pascal-VOC [2].

Figure 3. **Synthetic data**. Few images Generated by Stable Diffusion. We use the same set of animals (objects) present in COD10K and vary the color and texture of the object as well as the surroundings via text prompts.

5. Experiments, Results and Benchmark

5.1. Dataset Description

For this project we will be using 3 benchmark datasets to test our hypotheses: CAMO [7] and COD10K [3] and CHAMELEON [14]. CAMO [7] consists of 1250 images with naturally camouflaged objects like fish, chameleons and insects as well as artificially camouflaged objects like soldiers and body painting. Each image has at least one camouflaged object and corresponding manually-annotated semantic segmentation masks. COD10K [3] consists of 10,000 images containing terrestrial, aquatic and aerial animals as well as amphibians and body art. Most images were downloaded from Flicker by filtering with appropriate key-words. Each image is accompanied with rich hierarchical annotations including bounding-box, object-level, and instance-level labels. We will be working on a semantic segmentation task and will use only the corresponding labels in COD10K [3].

5.2. Segment Anything

Segment Anything Model (SAM) [6] was recently proposed by Meta AI and has demonstrated great generalization capacity to a very wide range of naturally occurring objects. The zero-shot variant of this model (also uses [9]) expects a text prompt and an image and can segment out objects indicated in the prompt. An interesting observation was that just the prompt "animal, insect, fish or human" was sufficient for it to detect camouflaged objects. SAM is trained on many images, and we expect at least part of it to contain reasonably camouflaged objects. The much higher performance of SAM than any of our baselines hints that perhaps a good understanding of object semantics is sufficient to detect camouflaged objects, without explicit training with camouflaged objects. This establishes a good benchmark for us to match and evaluate the performance of our methods in the limited data settings. All our proposed methods have been compared to the results we obtain by using SAM on the dataset.

5.3. Evaluation Metrics

For an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, let the predicted segmentation mask be $\mathcal{P} \in \mathbb{B}^{H\times W}$ and true segmentation mask be $\mathcal{G} \in \mathbb{B}^{H\times W}$, where $\mathbb{B} = \{0, 1\}$. We use $\mathbb{1}$ as the indicator function.

Intersection Over Union (IOU). This metric measures the overlap between the predicted and true segmentation masks. It is a crucial measure for evaluating the precision of the segmentation models, especially in the cases where the object of interest occupies a small portion of the image space. The IOU metric is given by

$$IOU = \frac{\sum_{i=0,j=0}^{H,W} \mathbb{1}[\mathcal{P}_{i,j} = 1 \wedge \mathcal{G}_{i,j} = 1]}{\sum_{i=0,j=0}^{H,W} \mathbb{1}[\mathcal{P}_{i,j} = 1 \vee \mathcal{G}_{i,j} = 1]} \tag{1}$$

Pixel Accuracy (P-ACC) Here, we compute the class-averaged accuracy of the foreground and background classes. Pixel accuracy provides a simple and direct measurement of the proportion of correctly classified pixels.

$$ACC_l = \frac{\sum_{i=0,j=0}^{H,W} \mathbb{1}[\mathcal{P}_{i,j} = l \wedge \mathcal{G}_{i,j} = l]}{\sum_{i=0,j=0}^{H,W} \mathbb{1}[\mathcal{G}_{i,j} = l]} \quad (2)$$

$$P\text{-}ACC = \frac{1}{2}(ACC_0 + ACC_1) \quad (3)$$

Dice Score (Dice). The Dice score, also known as Dice coefficient, is used to gauge the similarity between the predicted mask and the ground truth. It is specially useful for datasets with class imbalances, as it considers both precision and recall of the predictions.

$$Dice = \frac{2\sum_{i=0,j=0}^{H,W} \mathbb{1}[\mathcal{P}_{i,j} = 1 \wedge \mathcal{G}_{i,j} = 1]}{\sum_{i=0,j=0}^{H,W} \mathbb{1}[\mathcal{P}_{i,j} = 1] + \mathbb{1}[\mathcal{G}_{i,j} = 1]} \quad (4)$$

Each of these metrics provided a unique perspective, and insight into the model performance. They allowed us to comprehensively assess our segmentation approach.

5.4. Style transfer - Styles and Hyper-parameters

We used a pre-trained VGG-19 model as the backbone to perform neural style transfer. After fixing the content weight for all images, we individually fine-tuned style weights for every texture. 15 different natural scene textures were picked from internet or iNaturalist dataset consisting of trees, shrubs, tropical forests, rain, wind, snow, etc. In Fig 2 we present the ones that brought about the most realistic transformations over the images without completely disrupting the image properties. Since the content images were picked from the dataset itself, there was no need to make new annotated masks. For a modified butterfly image, we could pair it with the annotated mask of the original butterfly and make it part of the dataset, to be fed into SINet.

This pipeline presents a way to get new data for natural image settings without needing more expert annotations and utilizing the existing ones. This potentially could be useful in other settings such as concealed object detection (identifying hidden people in the environment) or defect detection (generating different samples of defects by replicating defect creation).

5.5. SINet training with Style Transfer

We integrated style transfer augmentation into the SINet architecture. We conducted two main experiments for the style transfer augmentation approach. For the first experiment, we expanded the training set with 15,190 images generated by applying style transfer to 3038 camouflaged images from COD10K, utilizing five distinct textures. In the second experiment, we adjusted our approach by randomly sampling 3000 images from the complete set of 15,190 style-transferred images. The subset of 3000 images is resampled from the complete pool every epoch. This modification aimed to prevent potential over-fitting on the styled images which out-number the original images. While This is better than generating just 3000 images in the first place, because over the course of multiple epochs, we expect it to cover all 15K images, thereby introducing higher diversity while preserving the importance of original images.

5.6. Analysis of Empirical Results

The table 1 lists the empirical results obtained using the proposed methods in this work. As shown in this table, the inclusion of data modified using style transfer brings a significant improvement over the SINet backbone and reduces the performance gap between the SINet and SAM model. The synthetic data generated using the stable diffusion pipeline brings the most boost to the performance although the numbers are comparable to those obtained with style transfer methodology. This is an indication of the prowess of the synthetic data and its versatile applications in other similar problems such as anomaly or defect detection that struggle to find a suitable solution without expert annotation.

Our experiments demonstrate that incorporating $\mathcal{L}_{bal}$ significantly enhanced model performance across all metrics; improving IOU from 33.16 to 43.10 and P-ACC from 69.19 to 79.22 on the COD10K dataset. Using style-transfer augmented images, we observe an improvements of 4.1%, 1.8% and 4.8% for IOU, P-ACC and Dice scores. Similar improvements are reflected in the CAMO and CHAMELEON datasets as well. Using synthetic data in addition to the original images improves the IOU by 7.8%, 10.7% and 5.6% for the three datasets. $\mathcal{L}_{bal}$ can be used directly with the other two methods as well, leading to further improvements. Even with our best method, there is a gap of 27.6% and 29.6% IOU in COD10K and CAMO respectively to beat SAM. The gap for the much smaller CHAMELEON dataset is a more achievable 10.2% IOU.

We ablate the synthetic data and style transfer experiments in Table 2 and Table 3 respectively. For the synthetic data, we vary the ratio of synthetic to original images in Fig 4. The general trend for both these settings was that using a roughly equal amount of original and new images is a better idea than either extreme.

In the style transfer experiments, we modified the quantity of style transfer images incorporated into the training dataset. The original COD10K dataset comprised 3038 camouflaged images. Given that style transfer was applied to five textures for all 3038 images, the total number of images augmented with style transfer amounted to approximately 15,190. We conducted two experiments using two different subsets of these images: the complete set of 15,190 images with the original training dataset and a randomly se-

Experiment	Back bone	COD-10K IOU	COD-10K P-ACC	COD-10K Dice	CAMO IOU	CAMO P-ACC	CAMO Dice	CHAMELEON IOU	CHAMELEON P-ACC	CHAMELEON Dice
SINet (released model)	R-50	43.83	76.27	54.26	43.26	73.54	52.36	50.94	78.23	60.43
SINet (reproduced)	R-50	33.82	70.24	44.91	26.23	64.04	34.73	55.66	80.06	65.90
UNet	R-50	17.09	70.16	26.47	27.81	70.75	40.97	24.72	72.18	37.07
SINet (reproduced)	R-18	33.16	69.19	43.07	22.83	62.11	29.69	56.21	79.95	64.60
SINet + balanced loss ($\mathcal{L}_{bal}$)	R-18	43.10	79.22	55.15	34.41	70.13	43.68	62.55	88.76	**73.20**
SINet + Style Transfer ($\mathcal{ST}$)	R-18	37.29	71.01	48.22	25.81	63.49	33.37	60.37	81.91	69.29
SINet + Synthetic Data ($\mathcal{SD}$)	R-18	40.95	73.40	51.93	33.52	67.80	42.26	61.89	83.03	70.72
SINet + $\mathcal{L}_{bal}$ + $\mathcal{ST}$	R-18	43.40	78.26	55.27	32.73	68.41	41.21	**62.88**	87.94	72.59
SINet + $\mathcal{L}_{bal}$ + $\mathcal{SD}$	R-18	**43.87**	**80.49**	**55.66**	**37.11**	**71.75**	**47.03**	61.37	**89.30**	72.27
Weak sup (GradCAM etc)	R-18	3.99	88.34	6.95	7.59	78.83	12.55	3.27	69.37	14.20
Segment Anything Model	-	71.48	88.42	79.07	66.68	83.91	74.86	73.10	88.96	81.00

Table 1. **Main Experiments.** Performance of the different methods on the three evaluation benchmarks used by COD10K.

Original Data	Synthetic Data	COD-10K IOU	COD-10K P-ACC	COD-10K Dice
✓		33.16	69.19	43.07
	✓	29.15	67.38	38.07
✓	✓	**40.95**	**73.40**	**51.93**

Table 2. **Synthetic data experiments.** Using synthetic data alone gives worse performance than using both synthetic and original.

Experiment	COD-10K IOU	COD-10K P-ACC	COD-10K Dice
Original only	33.16	69.19	43.07
Original + all ST	34.43	69.10	44.53
Original + sample ST	**37.29**	**71.01**	**48.22**

Table 3. **Style Transfer experiments.** Randomly sampling a roughly equal set of style transfer images performs better than keeping all. This may be because the model over-fits on style transferred images dominating the original ones 5 to 1 in the latter.

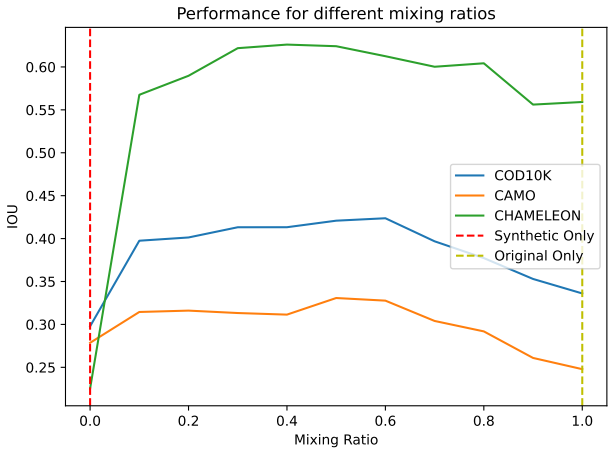

Figure 4. **Synthetic Mix**. We vary the ratio of the original images and synthetic images to be used for training. We observe that having a roughly equal split is better than either extreme.

lected sample of 3000 images (approximately one style per image) with the original dataset. The experimental results are summarized in Table 3.

The results revealed that adding all style-transferred images to the original dataset yielded a modest improvement in IOU and Dice metrics, but a slight decrease in Pixel Accuracy. Conversely, the use of a selectively sampled subset of the style-transferred images resulted in significant enhancements across all metrics: IOU increased to 37.29, Pixel Accuracy improved to 71.01, and the Dice coefficient rose to 48.22. We noticed that by focusing on a diverse yet concise subset of augmented images, we can achieve more pronounced gains in model performance.

We visualise some interesting predictions in Fig 5. Masks generated by SAM are usually sharper and better than other methods. While methods tend to miss parts of a limb, the general shape and location of the masks tend to be quite good. Masks obtained from the weakly supervised setting are able to localise the object, but do not have shapes comparable to the rest.

6. Conclusions

Both the style transfer and synthetic data approaches proved to be potent strategies for augmenting the training datasets, and enhancing the segmentation models' performance. They were effective in introducing variability and complexity into the training process, thereby enabling the

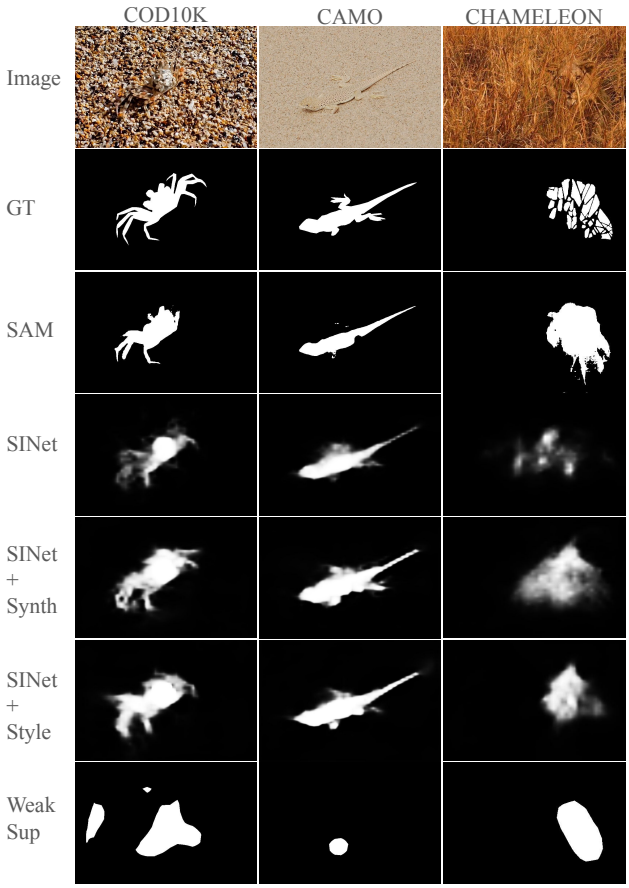

Figure 5. **Predictions**. Visualisations of masks predicted by the different methods for a few images. For the first and second, most methods are able to do good, missing partial limbs but capturing the overall shape of the object. The third one is a good case of very difficult occlusion caused by thin grass. SAM is able to capture the structure of the object here, but is not able to remove occluded regions. The difference between SINet and SINet + synthetic is also more apparent here.

model to perform better than the baseline SINet.

Our experiments not only underscored individual improvements but also showcased the advantages of combining $\mathcal{L}_{bal}$ with style transfer and synthetic augmented datasets to yield the most substantial performance gains. This suggests that a multifaceted strategy is critical for addressing the intricate challenges associated with segmentation tasks in limited data settings.

These advances have the potential to be used in realistic settings for ecological surveillance, military applications, and autonomous navigation where the need for object detection is required to be very precise and robust. Future work could extend the scope of our research by training the model on a broader dataset where camouflaged objects are not predominantly centred. This would potentially improve the model's ability to detect camouflaged objects in more complex and varied scenarios, further enhancing its applicability and performance in real-world settings.

7. Feedback from Poster Session

We received positive feedback and interest, particularly for the style transfer method. One suggestion was to include additional human-based camouflaged images which may be easier to obtain from techniques in film-making like the green screen and track-able suits. For the search and rescue application, a suggestion was to include additional modalities like infrared images, which can make the task much easier. An interesting idea was to use style transfer at test time where an object augmented to a different style may make it pop-out, reducing the difficulty of segmentation. Another idea was to explore recent advances in generative AI and make generators that allow better control over generated images. This can help get harder camouflaged images and further improve segmentation robustness. It was recommended to dig deeper into how the Segment-Anything-Model outperforms our approaches even without explicitly seeing camouflaged objects. Perhaps we do not need camouflaged images for training at all, and a model good enough on natural images can detect camouflaged objects too. Even if this is the case, camouflaged objects can serve as a robust evaluation benchmark that covers challenging corner cases.

References

[1] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media*, 5:117–150, 2019. 2

[2] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 2, 4

[3] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020. 1, 2, 3, 4, 5

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 5

[7] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019. 2, 5

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 4

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[10] Erik Learned-Miller Pia Bideau. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[11] Jingjing Ren, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Yangyang Xu, Weiming Wang, Zijun Deng, and Pheng-Ann Heng. Deep texture-aware features for camouflaged object detection, 2021. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019. 4

[14] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 5

[15] Mochu Xiang, Jing Zhang, Yunqiu Lv, Aixuan Li, Yiran Zhong, and Yuchao Dai. Exploring depth contribution for camouflaged object detection, 2022. 2

[16] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021. 2