GAN-based Image Purification for Camouflaged Object Detection

Seunghyeon Seo, Jeongin Moon
Seoul National University, Seoul, Republic of Korea
{zzzlssh, moon00}@snu.ac.kr

Abstract

Camouflaged object detection is one of the most challenging tasks in computer visual recognition field in that it is hard to distinguish between camouflaged objects and backgrounds with highly similar texture. Although it is a promising research topic with its usefulness such as in the military and medical purpose, it has not been actively explored yet. It has been focused on segmentation task and intricately designed model to capture camouflaged objects so far. In this paper, we propose much simpler approach for camouflaged object detection and expand spectrum of task to detection with bounding boxes. By experiments, we show that our camouflaged image purification network contributes to improving detector's performance, e.g. mAP of SSD with purified images increases over 14%p than that with original images. It can be utilized in plug-and-play style with other models so that it can be applicable to any other networks.

1. Introduction

Computer visual recognition technology has reached beyond the level of human visual recognition these days. A self-learning algorithm called deep learning allows computers to recognize features in photos or videos more precisely and systematically than human. These achievements are very encouraging from a historical perspective in which computers came to be capable of replacing or supporting the judgement of human through computer algorithm. However, despite the remarkable advances in ordinary computer visual recognition, computer visual recognition for camouflaged object is far from human-level recognition. In particular, images collected in exploration mission, in surgery, or in military operations are representative field that conventional computer visual recognition algorithm has difficulty in object detection and classification. The camouflaged object has properties by which human visual attention could be easily distracted by the surrounded environment. Thus, to prevent the unforeseen risks from camouflaged object in exploration and/or to alleviate the human visual load in high intensive visual recognition task, computer algorithm to help for camouflaged object are needed to be explored.

Figure 1. We aim to process camouflaged object images so that detectors can capture objects better than original ones. We generate corresponding target image by simple per-channel mean subtraction from original image. To make camouflaged objects more detectable, Neural Representation Purifier[12] is reinterpreted and modified for our task.

The term "Camouflage" was originated from natural animal behavior to describe animals changing their appearance in colors and shapes to hide themselves from natural enemies, in modern times the same sense applies to artificial camouflage properties created by human[10]. In general, camouflaged objects means the objects that are difficult to distinguish from background due to its similarity[5]. Contrast to salient objects, computer visual recognition models for camouflaged objects are not easy to outperform due to next three reasons. First, common computer visual recognition algorithms are not easy to capture features of camouflaged objects due to the textural similarity between object and background. Because the boundary between the background and objects is often unclear and parts of objects are even hidden in the background. Second, to obtain high-definition image of camouflage objects is highly difficult because the amount of data for training computer visual algorithms is insufficient compared to salient objects. Third, judgment of the level of camouflage is highly subjective and thus it is confusing to form accurate labeling for image classification. Fan et al[5] stated that the average labelling took 1-hour per single image classification to produce COD10K data. Because of these difficulties, computer visual recognition algorithm for camouflaged object requires additional process to increase model performance.

To increase camouflage objects detection performance

in model, it is necessary to acknowledge that the image features which play an important role in detecting salient objects are no longer helpful in detecting camouflaged objects and the model even need to break away from the elements that salient object attract observers' attention[20]. According to sensory biologists, the principle of natural animal camouflage works in a way that deceiving the visual perception of the observer[18]. In addition, it should also be recognized that most CNN-based image recognition networks operate in a form that mimics the signal system of the human brain, so that CNN-based object recognition networks could also be deceived, just as the human brain is deceived by camouflaged objects[1]. For example, colorblind patients tend to perceive camouflage objects better than ordinary people because they are less dependent on colors and texture, shape to recognize objects whereas they are more trained to focus on overall biological features. Furthermore, the efficacy of camouflage pattern of animal increases as the pattern locate closer to object boundary and have higher contrast[4]. Therefore, to detect camouflage objects well, model developing strategies based on a deep understanding of natural animal camouflage function are needed.

Based on these backgrounds and motive for research, our goal is to develop improved computer visual recognition model for camouflage objects in terms of object detection algorithm rather than segmentation algorithm. Given the fact that camouflaged patterns that distinguish the boundaries between camouflaged objects and backgrounds are highly dependent on the camouflage strategies of living animal, we hypothesized that the structural properties of camouflaged pattern can be a discriminative feature to distinguish objects from background. It is assumed that by learning the specific characteristics of the camouflage pattern, camouflaged objects can be processed with bounding box at instance level. Consequently, through this work, we seek to explore the image processing model which change the style of image towards more detectable fashion. Our main contributions are:

- ***Task generalizability:*** We follow overall architectural style of [12] with appropriate modifications depending on characteristic of our task. Sharing the similar purpose for input image processing, our proposed network can be added before any model.
- ***Separation loss:*** We propose additional loss for better purification quality for camouflaged object images. With minimizing KL-divergence of foregrounds and backgrounds between purified images and target images respectively, we can achieve higher detection accuracy.
- ***Novel try for camouflaged object detection:*** Regarding camouflaged object images, it has been mainly focused on segmentation task so far. We try to expand camouflaged object detection task from segmentation to detection task with bounding boxes. Furthermore, we focus on image itself so that our proposed network can be applicable in wide spectrum of task.

2. Related Work

2.1. Camouflaged object detection

Camouflaged object detection (COD) has been known as a fairly challenging task. Camouflaged objects and its surroundings have very similar textures. For that reason, deep learning algorithms for generic object detection[15, 7, 3, 2, 13, 16, 21] and salient object detection[22, 23] do not perform well.

With the development of computer vision and deep learning, various attempts have been made for camouflaged object detection. Among them, Fan et al[5] collected the first large scale dataset for COD and proposed an algorithm called SINet using a search and identification module to capture camouflaged objects. Ren et al[14] proposed TANet, which increases the texture difference between camouflaged objects and the background by using the affinity function and improves the segmentation quality by boundary-consistency loss.

While both actually dealt with segmentation task, we focus on the existing object detection task that detects objects with bounding boxes. Our study is, to the best of our knowledge, the first attempt to perform detection task with bounding boxes on COD10K dataset.

2.2. Generative adversarial network

We aim to make camouflaged objects in an image more detectable before being sent to model. It can be seen as a image-to-image translation task because we want to change the style of an object from camouflage to normal. With this intuition, we apply GAN[6] framework for our task. GAN[6] refers to adversarial frameworks in which generator and discriminator can be trained in unsupervised ways by competing against each other. GAN[6] is commonly used to generate new image samples due to its capability to produce realistic results. Therefore, it is free for data scarcity or arduous image labeling tasks. Furthermore, with conditional adjustment it could also be used as a domain adaptation technique for mapping training images of source domain to target domain. [12] proposed a Neural Representation Purifier network that obtains normal images from contaminated images by competitively training generator to remove impurities induced by the self-supervised perturbation. Similarly, using GAN framework we tried to train generator to distinguish objects from the background.

3. Methodology

Our main goal is to process an input image so that the image with camouflaged objects used as an input can be

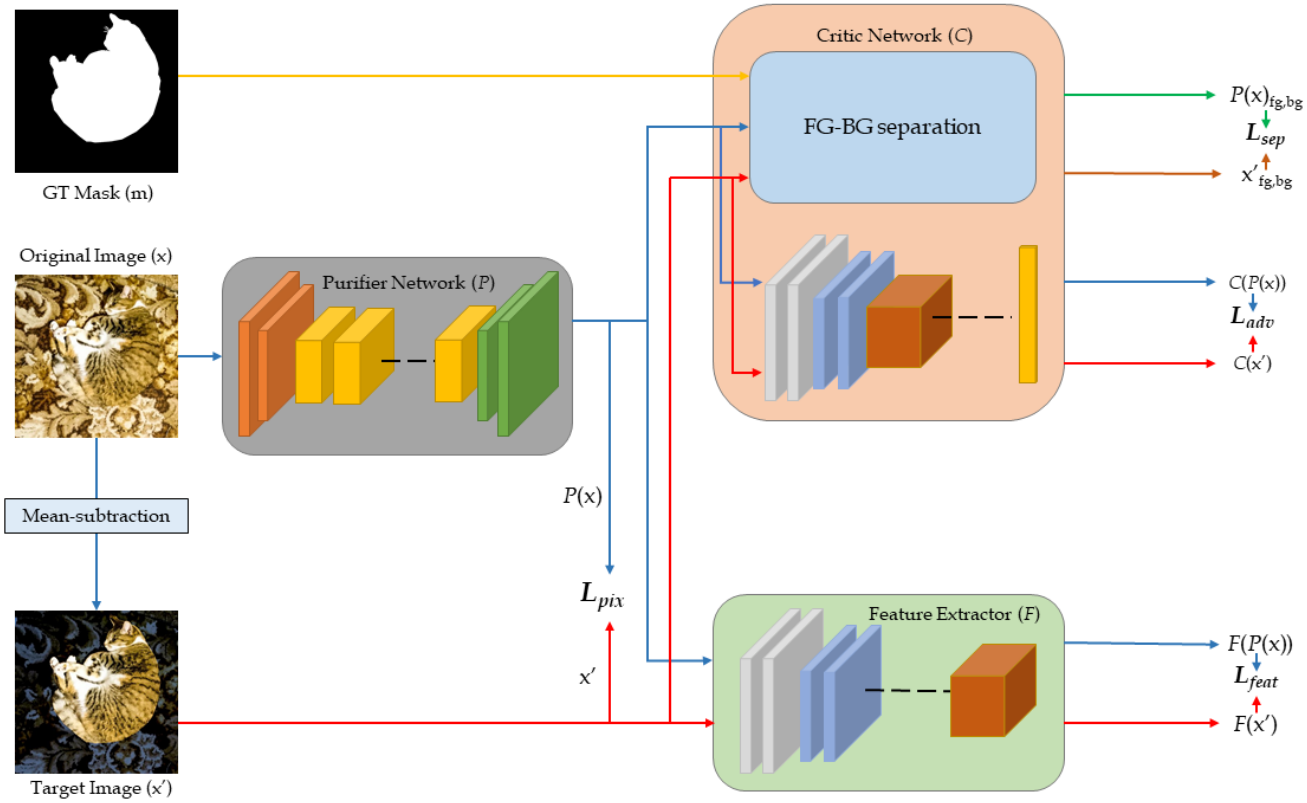

Figure 2. *Schematic diagram for CAM-image purifier network training*. Based on Neural Representation Purifier Network[12], we slightly modify the setting and loss functions according to our camouflaged object image purification task. For enhancing purification quality, we add separation loss term for better discrimination foreground and background.

captured well in tasks such as detection and classification. In our proposed algorithm, using the GAN framework, an original image is purified by referring to target image that has been processed to make foreground objects more prominent. Overall, we follow the architecture of the Neural Representation Purifier proposed by [12] (see Fig. 2).

Our basic intuition behind this processing is that if the area corresponding to the object in camouflaged image is well distinguished from the background, the detector will be able to localize well. So, using ground-truth masks of original images, we created target images by subtracting the mean value of each RGB channel for the background area. As a result, we were able to obtain a target image in which the camouflaged object and the texture of the background are clearly distinguished, and we used those target images as references for original images (see Fig. 3). Since the purification network is an operation that transforms the input image, it can be used as a input processor regardless of the following model. We modified the setting according to the goal of our task. While [12] purifies adversarial examples generated from its original ones, we put the original image in the position of adversarial example of the framework and the mean-subtracted target image in the original image. In addition, we modified the loss as well and also changed the structure of the purifier network due to resource problems and task differences. Its ablation study can be found in section 4.

3.1. Architecture

Overall, we constructed the network according to the NRP architecture proposed by [12], and it is largely composed of a purifier, a critic, and a feature extractor.

Purifier: A purifier network that makes the object of the original input image well-marked can be said to play the role of a generator of the GAN framework[6]. We followed the style of the generator proposed by [12], and filter size, number of blocks, etc. were adjusted due to shortage of GPU memory. In addition to such a DenseNet[9]-based structure, we explored a lighter ResNet[8]-based structure and a SINet[5] structure designed for a camouflaged object segmentation task, but DenseNet[9]-based purifier showed the best performance. A related ablation study can be found in section 4.3.

Critic: We also reproduced a critic structure based on the network of [12]. The pretrained VGG16 network[17] with batch normalization layers is used. We attached a fully con-

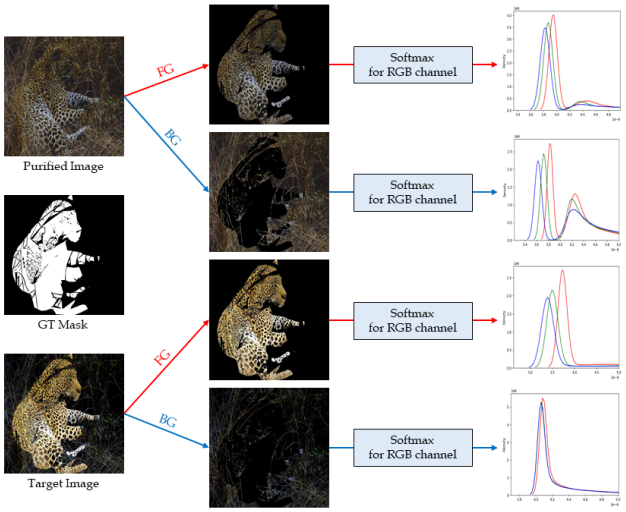

Figure 3. Foreground-only and background-only image are generated using GT mask. Pixel values are converted to probability distribution by channel through softmax function.

nected layer as a classifier.

Feature Extractor: Similar to the Critic network, a pretrained VGG16 network but without batch normalization layers is used. During training, We fixed the feature extractor. [12] showed that minimizing the distance between features from this feature extractor has positive effect on purification.

3.2. Loss Functions

We design the hybrid loss function proposed by [12] to be more suitable for the camouflaged object image purification task by adding **separation loss**. As a result, our loss function consists of four loss terms at total, and each term is as follows.

Adversarial loss: As suggested by [12], we used relativistic average GAN loss for better loss convergence. The adversarial loss is given as:

$$\mathcal{L}_{adv} = -\log\big(\sigma\big(\mathcal{C}(\mathcal{P}(x)) - \mathcal{C}(x')\big)\big). \quad (1)$$

Pixel loss: A loss term is added to compare the target and purified image in pixel space so that purification occurs in a direction similar to the target image style. An L2 loss was used in [12], but we compared L1, L2, and SSIM loss[19]. The L2 loss is found empirically to be the most appropriate through experiments. A detailed comparison can be found in Table 1. The pixel loss term is as follows:

$$\mathcal{L}_{pix} = \|\mathcal{P}(x) - x'\|_2. \quad (2)$$

Feature loss: The network is trained by narrowing the distance between features from the fixed pretrained VGG16[17]. Mean Absolute Error is used as a metric. The distance is as follows:

$$\mathcal{L}_{feat} = \Delta\big(\mathcal{F}(x'), \mathcal{F}(\mathcal{P}(x))\big), \quad (3)$$

where Δ is distance metric, MAE in our case.

Separation loss: A separation loss is added to further enhance purification toward the target image style. Foreground-only and background-only images are generated from purified and target images respectively by using GT mask. And we minimize Kullback-Leibler divergence of foreground and background RGB channel values between purified and target images respectively. As shown in Fig. 3, RGB channel values are represented as a probability distribution through softmax function so that we can use KL-divergence loss. We show that separation loss enhances image processing in the direction of improving detection task performance. The corresponding loss term is as follows:

$$\mathcal{L}_{sep} = \mathcal{D}_{KL}(\mathcal{P}(x)_{fg} \| x'_{fg}) + \mathcal{D}_{KL}(\mathcal{P}(x)_{bg} \| x'_{bg}), \quad (4)$$

and total loss terms are:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{adv} + \beta \cdot \mathcal{L}_{pix} + \gamma \cdot \mathcal{L}_{feat} + \lambda \cdot \mathcal{L}_{sep}. \quad (5)$$

4. Experiments

4.1. Training Settings

Training for purifier network is done on COD10K dataset[5]. We follow overall training protocols of [12]. The images are resized to $480 \times 480 \times 3$, and the corresponding target images are generated by subtracting background pixels' mean value of each RGB channel. Batch size is set to 8 and two TITAN RTX GPUs are used for training. We use Adam optimizer with learning rate 10^{-4} and hyper-parameters for loss, $\alpha = 5 \times 10^{-3}$, $\beta = 1 \times 10^{-2}$, $\gamma = 1$ and $\lambda = 0.5$. We study two representative detection models, SSD[11] from one-stage models and Faster R-CNN[15] from two-stage models. Default versions of two detectors without our network were used as our baselines. We followed default training settings for both of them.[1]

4.2. Results

First of all, we trained vanilla SSD[11] and Faster R-CNN[15] with default training protocol as mentioned above. Original images were used as input, but resized to $480 \times 480 \times 3$ for fair comparison with purified images. Because of memory issue, our purifier network should return images with fixed size as output, $480 \times 480 \times 3$.

[1] We followed the training setting of github and webpage referenced.
SSD: https://github.com/amdegroot/ssd.pytorch
Faster R-CNN: https://pytorch.org/tutorials/intermediate/torchvision_tutorial.html

Detector	Input	1-class	5-class
SSD[11]	Original	**0.037**	**0.012**
	Target	**0.077**	**0.043**
	Purified (ResNet) with L1 Loss[8]	0.086	0.007
	Purified (ResNet) with L2 Loss[8]	0.209	0.003
	Purified (ResNet) with SSIM Loss[8, 19]	0.182	0.002
	Purified (DenseNet) with L1 Loss[9]	0.019	0.008
	Purified (DenseNet) with L2 Loss[9]	**0.280**	0.001
	Purified (DenseNet) with SSIM Loss[9, 19]	0.051	0.007
	Purified (SINet) with L1 Loss[5]	0.068	0.034
	Purified (SINet) with L2 Loss[5]	0.053	0.027
	Purified (SINet) with SSIM Loss[5, 19]	0.127	**0.042**
Faster R-CNN[15]	Original	**0.044**	**0.015**
	Target	**0.480**	**0.096**
	Purified (ResNet) with L1 Loss[8]	0.042	**0.016**
	Purified (ResNet) with L2 Loss[8]	0.022	0.001
	Purified (ResNet) with SSIM Loss[8, 19]	0.025	0.014
	Purified (DenseNet) with L1 Loss[9]	0.026	0.005
	Purified (DenseNet) with L2 Loss[9]	0.028	0.003
	Purified (DenseNet) with SSIM Loss[9, 19]	0.049	0.009
	Purified (SINet) with L1 Loss[5]	0.028	0.011
	Purified (SINet) with L2 Loss[5]	**0.066**	0.009
	Purified (SINet) with SSIM Loss[5, 19]	0.025	0.001

Table 1. Quantitative results (mAP) on different number of classes, purifier architectures and pixel losses. n-class means the number of classes to be classified by the detector.

And we suppose that the performance with mean-subtracted target images could be higher bound of our purifier network performance because we use mean-subtracted image as a reference for processing original image to be well detected in models. With purified images, we explore three architectures[9, 8, 5] for purifier network and three loss functions for pixel loss.

Table 1. shows the quantitative results of combinations. In case of 1-class detection task, Purified images enhance increasing accuracy of detector in case of SSD[11], showing even much higher mAP than mean-subtracted target images. However, Faster R-CNN model[15] shows much greater difference in mAP between original images and target images compared to SSD[11]. Furthermore, there are

Figure 4. Visualization of 1-class detection results of three different purifiers with L2 pixel loss. While camouflaged objects become more salient with SINet-based purifier from human's perspective, DenseNet-based purifier captures objects most accurately.

not salient enhancing in detection performance with purified images. We conjecture that it comes from the architectural difference between two models in localizing and classifying objects.

For 5-class detection task, we can see performance improvement of SSD[11] as well when trained with purified images from SINet-based[5] purifier. We conjecture that poor increasing of mAP in 5-class detection compared to 1-class results from difficulty of classification for camouflaged objects.

Qualitative results are shown in Fig. 4. We compare 1-class detection results of SSD[11] with three different purified images from ResNet-based[8], DenseNet-based[9], SINet-based[5] purifier with L2 pixel loss. As the most performance-boosting purifier, the results with purified images from DenseNet-based[9] purifier show the most accurate and tightest bounding box. Interestingly, purified images from SINet-based[5] purifier seem well-distinguished between objects and background from human perspective, but result in the lowest mAP among three sets of images from different purifiers.

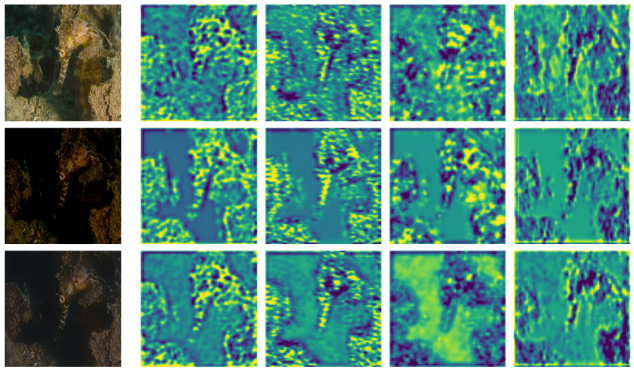

Figure 5. Visualization of feature maps from three different images, original image, purified image without separation loss and purified image with separation loss (from top to bottom).

4.3. Ablation Study

We could produce purified images which result in higher mAP in detection with two representative models when trained with separation loss. Therefore, we add separation loss into our loss terms for training purifier. Here we explore why those images trained with additional separation loss result in better detection performance.

We visualize feature maps of original images, purified images from network trained without separation loss and those with separation loss respectively, as shown in Fig. 5. Compared to feature maps from original images, purified ones capture more salient shape of objects and the feature map with separation loss can make more clear contour of objects than the one without separation loss. Interestingly, we can see that the feature map with separation loss at the third column seems to have information related to backgrounds. We conjecture that the separation loss gives detector an extra signal to learn foregrounds and backgrounds discriminatively.

5. Conclusion

We propose a camouflage purification network based on Neural Representation Purifier[12]. Compared to other studies, our proposed network focuses on input processing stage which is more generalizable approach than existing networks with highly complicated design. It exhibits an applicability as a performance booster with better performance when used for processing input images prior to detection models. Exploration of camouflaged object images' own characteristic from perspectives of features and image itself can be an interesting future research.

References

[1] Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4974–4983, 2019. 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 834–848, 2017. 2

[4] I. Cuthill, M. Stevens, J. Sheppard, T. Maddocks, C. Párraga, and T. Troscianko. Disruptive coloration and background pattern matching. *Nature*, 434:72–74, 2005. 2

[5] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 4, 5

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2, 3

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3, 5

[9] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. 3, 5

[10] 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. 1

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016. 4, 5

[12] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 4, 6

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2

[14] 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

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015. 2, 4, 5

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015. 2

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 3, 4

[18] Martin Stevens and Sami Merilaita. Animal camouflage: Current issues and new perspectives. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364:423–7, 12 2008. 2

[19] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4, 5

[20] Jinnan Yan, Trung-Nghia Le, Khanh Duy, Minh-Triet Tran, Thanh-Toan Do, and Tam Nguyena. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, PP:1–1, 03 2021. 2

[21] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2881–2890, 2017. 2

[22] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Int. Conf. Comput. Vis.*, pages 8779–8788, 2019. 2

[23] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3085–3094, 2019. 2