Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01455-7

ORIGINAL ARTICLE

ATTACK-COSM: attacking the camouflaged object segmentation model through digital world adversarial examples

Qiaoyi Li[1] · Zhengjie Wang[1] · Xiaoning Zhang[1] · Yang Li[1]

Received: 2 November 2023 / Accepted: 17 April 2024

Abstract

The camouflaged object segmentation model (COSM) has recently gained substantial attention due to its remarkable ability to detect camouflaged objects. Nevertheless, deep vision models are widely acknowledged to be susceptible to adversarial examples, which can mislead models, causing them to make incorrect predictions through imperceptible perturbations. The vulnerability to adversarial attacks raises significant concerns when deploying COSM in security-sensitive applications. Consequently, it is crucial to determine whether the foundational vision model COSM is also susceptible to such attacks. To our knowledge, our work represents the first exploration of strategies for targeting COSM with adversarial examples in the digital world. With the primary objective of reversing the predictions for both masked objects and backgrounds, we explore the adversarial robustness of COSM in full white-box and black-box settings. In addition to the primary objective of reversing the predictions for masked objects and backgrounds, our investigation reveals the potential to generate any desired mask through adversarial attacks. The experimental results indicate that COSM demonstrates weak robustness, rendering it vulnerable to adversarial example attacks. In the realm of COS, the projected gradient descent (PGD) attack method exhibits superior attack capabilities compared to the fast gradient sign (FGSM) method in both white-box and black-box settings. These findings reduce the security risks in the application of COSM and pave the way for multiple applications of COSM.

Keywords COSM · Adversarial robustness · White-box setting · Black-box setting

Introduction

The camouflaged object segmentation model (COSM) aims to identify objects that exhibit various forms of camouflage. This field has a wide range of real-world applications, including search-and-rescue operations, the discovery of rare species, healthcare (such as automated diagnosis for colorectal polyps [1] and lung lesions [2], medical image fusion [3]), agriculture (including pest identification [4], fruit ripeness assessment [5] and biological disease diagnosis [6]), and content creation (such as recreational art [7]). Figure 1 depicts different categories of camouflage objects [8], with items (1)–(4) representing natural camouflage, and (5) and (6) showcasing artificial camouflage. More specifically, (1) features a terrestrial camouflage creature, (2) showcases an aquatic camouflage creature, (3) illustrates a flying camouflage creature, and (4) portrays a reptile camouflage creature. On the other hand, (5) displays camouflage soldiers, while (6) exhibits a human body painting camouflage object.

In recent years, this field has seen remarkable advancements, largely attributed to the availability of benchmark datasets such as COD10K [9, 10], and NC4K [11], in tandem with the rapid evolution of deep learning techniques. From SINet [9] in 2020 to POPNet [27] in 2023, the accuracy results for the COD10k test set are shown in Fig. 2, where the E-measure [50] improved from 0.864 to 0.897, the S-measure [49] improved from 0.776 to 0.827, the weighted F-measure [51] improved from 0.631 to 0.789, and the mean absolute error (MAE) decreased from 0.043 to 0.031. Evidently, the accuracy of the models has increased. However, research on

✉ Zhengjie Wang
wangzhengjie@bit.edu.cn

Qiaoyi Li
joe_li@bit.edu.cn

Xiaoning Zhang
xnzhang@bit.edu.cn

Yang Li
xiao-l@bit.edu.cn

[1] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China

Published online: 07 May 2024

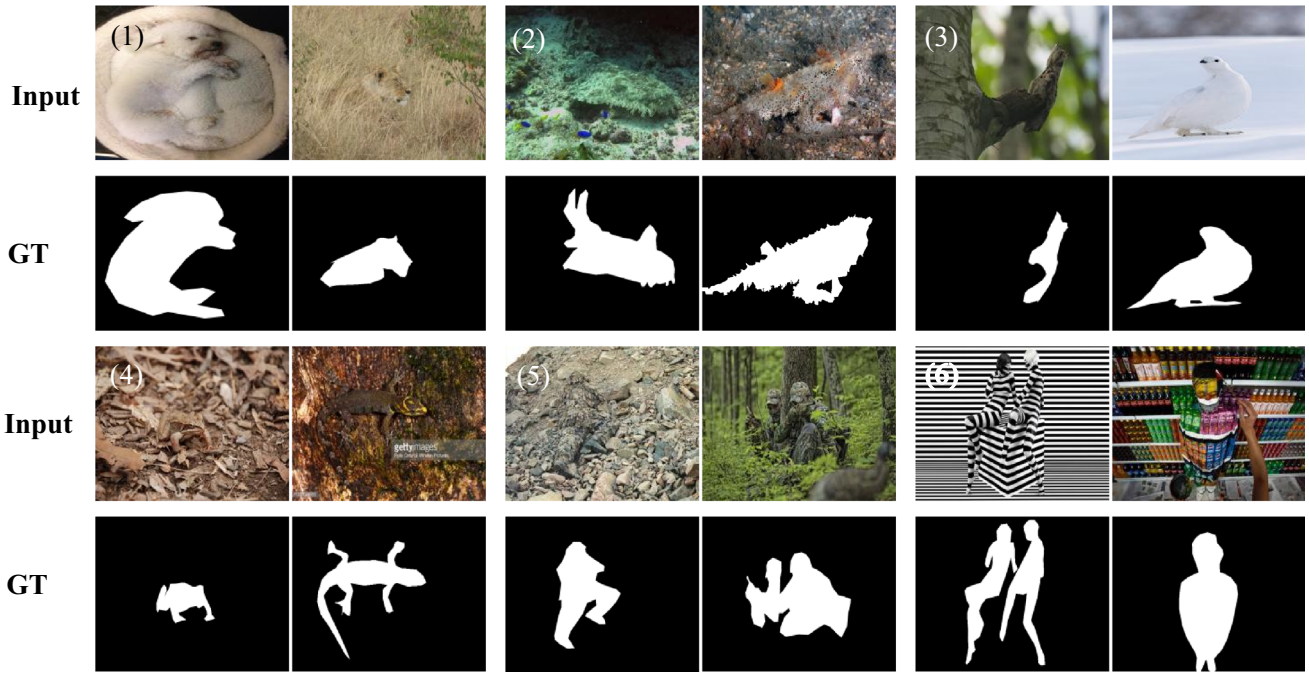


Fig. 1 Camouflaged object segmentation task. The proposed camouflaged object detection task, where the goal is to detect objects that have a similar pattern (e.g., edge, texture, or color) to the natural habitat [8]

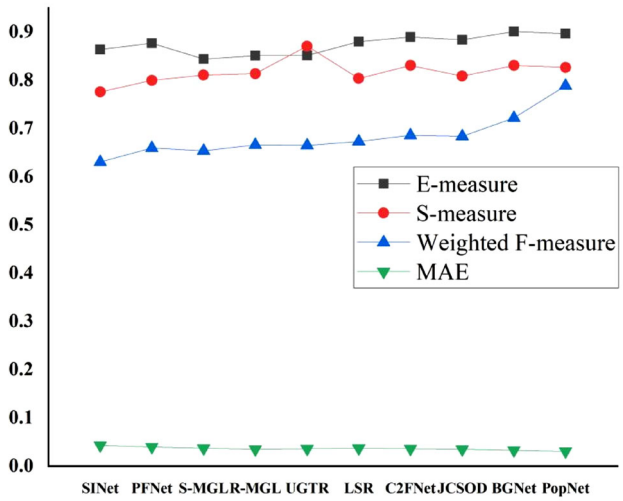


Fig. 2 Accuracy statistics of the camouflaged object segmentation models

the security of COSM against adversarial example attacks is still in its infancy. it remains unclear whether it can withstand adversarial attacks. This raises doubts about COSM being used in safety critical applications (such as, intelligent grid systems [12, 13] and autonomous driving systems [14]) since the networks could inexplicably classify a natural input incorrectly although it is almost identical to examples it has classified correctly before. Therefore, it is crucial to conduct research on the robustness of COSM, reducing the security risks associated with these models, and ultimately promoting their widespread application.

In this paper, Our work is the first investigation into how to launch adversarial attacks against COSM. We employ standard practices found in popular adversarial attack methods, such as the fast gradient sign method (FGSM) attack [26] and the projected gradient descent (PGD) attack [18]. COSM distinguishes itself from traditional image recognition (segmentation) models in two key aspects: (1) it produces masks without making label predictions, and (2) the objects detected by the model closely resemble the background.

For this purpose, we propose a framework named Attack-COSM, which is designed to launch attacks on COSM through the task of mask prediction for camouflaged targets. Specifically, our objective is to mislead COSM, causing the model to reverse its predictions for masked objects and backgrounds by increasing COSM loss. The experimental results indicate that adversarial attacks can effectively reduce the accuracy of COSM, implying that COSM is susceptible to adversarial examples. We also conducted experiments to evaluate the transfer attack performance of adversarial examples and found that adversarial examples generated using one COSM system can be used to effectively attack other COSM systems. In addition to the primary objective of reversing the predictions for masked objects and backgrounds, we consider the question of whether adversarial examples can be used to manipulate COSM and generate any desired mask. To achieve this, we structure the desired task in two settings: (1)

by assigning a manually designed mask to a random position and (2) by generating a mask from another image. In general, we discover that it is feasible to consistently generate the desired mask in most cases, underscoring the susceptibility of COSM to adversarial attacks. Overall, the contributions of this work are summarized as follows.

1. We conduct the first yet investigation on attacking COSM with adversarial examples. We present a framework for attacking COSM with the objective of reversing its predictions for masked objects and backgrounds.
2. We uncover that COSM is susceptible to adversarial attacks in a complete white-box setting. Furthermore, we demonstrate that the adversary can target the model without prior knowledge of its parameters. In other words, COSM can be partially compromised by adversarial examples in a cross-model setup.
3. In addition to the primary objective of reversing its predictions for masked objects and backgrounds, through further investigation, we successfully demonstrate that COSM can be manipulated to generate any desired mask. This further underscores the vulnerability of COSM.

The remainder of this work is structured as follows. "Related work" section provides an overview of related research, summarizing the advancements in COSM and different attack methodologies.

In "Framework for Attack-COSM" section, we present the framework for Attack-COSM, formulating the objective as reversing the predictions of masked objects and backgrounds. In "Main results of Attack-COSM" section, We begin by presenting the results of Attack-COSM in the white-box setting. We then delve into an exploration of transfer-based attacks on COSM. Finally, we investigate techniques for manipulating COSM to generate masks according to specific requirements. "Discussion" section discusses the relationship between attacking label predictions and attacking mask predictions, as well as the limitations of our work.

Related work

In this section, we separately outline the characteristics and classifications of COSM and typical adversarial attack algorithms.

COSM

Numerous projects and papers have explored the above mentioned topic from various perspectives, which can be broadly categorized into two groups: single-task and multitask learning. (a) Single-task learning is the most commonly employed paradigm in COS, focusing solely on the segmentation of concealed targets. Within this paradigm, most recent works, [9, 10, 19] concentrate on the development of attention modules for identifying target regions. Multitask learning introduces an auxiliary task to complement the segmentation task, enhancing the robustness of COS learning. These multitask frameworks can be implemented through various methods, including confidence estimation [20–23], localization/ranking [11, 24], category prediction [25], learning depth methods [26, 27], boundary methods [28–33], and texture [16, 34] cues of camouflaged objects.

Adversarial attacks

Deep neural networks, including CNNs [26, 35, 36] and vision transformers (ViT) [37–40], are well recognized for their vulnerability to adversarial examples. This susceptibility has spurred numerous studies aimed at examining model robustness under various types of adversarial attacks. Adversarial attack methods can be categorized into two settings: the white-box setting [17, 41, 42], which allows full access to the target model, and black-box attacks [43–48], which primarily rely on the transferability of adversarial examples. Another way to classify attacks is as untargeted or targeted. In the context of image recognition (classification), an attack is deemed successful under the untargeted setting if the predicted label differs from the ground-truth label. In the more stringent targeted setting, the attack is considered a failure unless the predicted label matches the predetermined target label. The prior works mentioned have primarily concentrated on manipulating image-level label predictions for image classification tasks. In contrast, our work considers attacking COSM for the task of predicting the masks of camouflaged targets. Attacking COSM also sets itself apart from attacking semantic segmentation models, as the generated masks lack semantic labels. It remains uncertain whether COSM can withstand adversarial attacks.

As shown in Table 1, we summarizes and analyzes the current typical adversarial example generation methods based on attack type, attack target, attack frequency, advantage, limitation etc. Single-step denotes a single iteration, while iteration represents multiple iterations. W stands for white-box attack, B for black-box attack, T for targeted attack, and NT for non-targeted attack.

Framework for Attack-COSM

Inspired by the image classification attack method, we analyze the difference between the image classification task and the COS task, and finally get the attack flow of the COS task.

Table 1 Summary of typical adversarial attack

Adversarial attack	Attack frequency	Attack type	Attack target	Advantage	Limitation
FGSM [17]	Single-step	W	NT	High efficiency in generation and transferability	Low success rate in white-box attacks
PGD [18]	Iteration	W	NT	High success rate in white-box attacks	Poor transferability
MI-FGSM [52]	Iteration	W	NT	Improved transferability of iterative attacks	
D-MI-FGSM [56]	Iteration	W	T, NT	It can be combined with other attacks to increase the success rates in both white-box and black-box scenarios	Low computational efficiency
C&W [41]	Iteration	W	T, NT	1. The resulting disturbance is small 2. It can break a lot of defenses 3. It has portability	The generation time is longer
SBA [57]	Iteration	B	T, NT	1. It can be used to attack other machine learning models 2. It can effectively avoid defense methods that rely on gradient masking	It is rarely used in practice because it is almost impossible for an attacker to get detailed information about the model
DeepFool [53]	Iteration	W	NT	Compared with FGSM, the perturbation is smaller and the calculation speed is improved	The generation time is five times that of FGSM
OPA [58]	Iteration	B	T, NT	It has high classification error rate under various models	It takes a long time to find hackable pixels
AdvGAN [59]	Iteration	W	T, NT	The resulting adversarial example is visually indistinguishable from the clean image	It is relatively limited in adversarial training settings and may not generalize to broader settings
UAP [54]	Iteration	W	NT	1. It proves the existence of a general disturbance 2. It has high portability	There is no guarantee that every updated generic adversarial disturbance will remain adversarial to the data points that occurred before the update
BPDA [55]	Iteration	W	T, NT	Defense methods that rely on confusion gradients can be effectively circumvented	Defense against confusion gradients only

Preliminaries

Mask prediction

As illustrated in Fig. 1, we treat COSM as a class-independent, pixel-wise segmentation task. Formally, let $I \in \mathbb{R}^{H\times W\times 3}$ and $C \in \mathbb{R}^{H\times W\times 1}$ denote the input image and output camouflage map, respectively. Given a large collection of such pairs $\{I_i, C_i\}_{i=1}^{N}$, our task is to learn a mapping function $\mathcal{F}_\Theta$ parameterized by weights Θ that can correctly transfer the novel input to its corresponding camouflage map. For each pixel (position) $\rho_o \in [1, H\times W]$, the estimated score $c_{\rho_o} \in [0, 1][0, 1]$ reflects the COD models, prediction, where a score of "1" indicates that it belongs to the camouflaged objects and vice versa. Note that for each pixel (position) $\rho_o \in [1, H \times W]$, the CODM has an intermediate predicted value y_{ρ_o}, which undergoes a sigmoid functions operation to obtain c_{ρ_o}. Namely, when y_{ρ_o} is a positive value, c_{ρ_o} takes the value of 1, and vice versa.

Common attack methods

Before introducing Attack-COSM, we begin by revisiting the commonly used attack methods in traditional classification tasks. We define $f(\cdot, \theta)$ as the target model to be attacked, parameterized by θ. With (X_c, Y_c) as data pairs from the original dataset, the adversarial image X_{adv} is defined as$X_c + \delta^*$, where δ^* is optimized in Eq. 1. More specifically, the

attack algorithm is designed to generate the optimal δ^*. The $\mathbb{S}$ in the formula represents the range of the perturbation limits. In the context of the classification task, Y_c typically signifies the class label, J_c is loss function, and the loss function is often the cross-entropy function.

$$\delta^* = {}_{\delta\in\mathbb{S}}{}^{\max} J_c(f(X_c + \delta;\, \theta),\, Y_c) \tag{1}$$

The typical attack algorithms listed in Table 1 are used to solve the above equation.

Attack-COSM

In typical adversarial attacks targeting image recognition models, the objective is to manipulate the predicted labels at the image level, thereby causing the model to produce inaccurate predictions. From Fig. 3, it can be observed that in adversarial attacks targeting COSM, the objective is to manipulate predicted labels at the pixel level. Additionally, due to the intrinsic similarity between camouflaged objects and background, COSM introduces new detection modules.

Task definition

Since the generated masks from COSM lack semantic labels, a direct approach to successfully attack COSM is to reverse the predictions for masked objects and backgrounds. In this work, we consider the reversal of predictions for masked objects and backgrounds as the fundamental objective of adversarial attacks on COSM. As per "Preliminaries" section, a pixel, denoted as ρ_o, is classified as masked when the intermediate predicted value, denoted as y_{ρ_o}, is positive. Therefore, the task is deemed successful when the predicted values y_{ρ_o} become negative. Conversely, a pixel ρ_o is classified as background when the predicted value y_{ρ_o} is negative, and the task is considered successful when predicted values y_{ρ_o} turn positive.

Loss design

To reverse the predictions of masked objects and background by attacking COSM, the loss design is expected to be adjusted to decrease the predicted values y_{ρ_o} until they become negative in the masked region and increase the predicted values y_{ρ_o} in the background region until they become positive. As shown in Eq. 2, we achieve the aforementioned objective by directly elevating the loss value of COSM to diminish the model's prediction accuracy. The loss function J_s used is BCEWithLogitsLoss. For a dataset related to attacking COS, we seek parameters δ to maximize the loss, i.e.,

$$\delta^* = {}_{\delta\in\mathbb{S}}{}^{\max} \sum_{1}^{N} J_s(\mathcal{F}_\Theta(I_i + \delta_i;\, \Theta),\, C_i) \tag{2}$$

As can be seen from formula 3: Unlike traditional classification tasks, we aggregate the loss values for each pixel to calculate the overall image loss. The loss for each picture is:

$$\begin{aligned} &J_s\left(\mathcal{F}_\Theta\left(I_i + \delta_i;\, \Theta\right),\, C_i\right) \\ &= -\sum_{\rho_o\in\Omega} \left(C_i^{\rho_o} \ln\left(c_{\rho_o}\right) + \left(1 - C_i^{\rho_o}\right) \ln\left(1 - c_{\rho_o}\right)\right) \end{aligned} \tag{3}$$

Here, Ω represents the spatial composition of all pixels for each image, and $C_i^{\rho_o}$ represents the ground truth at pixel position ρ_o for the ith image.

Attack details

The FGSM [17] and PGD [18] are two widely employed methods for assessing model robustness, and they are chosen for their simplicity and effectiveness. FGSM is a single-step attack method based on the model gradient on the input image. PGD is a multi-step attack method, and it is represented as PGD followed by the number of iterations, denoted as PGD-N.

So, we employ the FGSM attack [17] and PGD attack [18], a method commonly used for assessing model robustness in prior studies. Following the established practices of attacking vision models in a white-box scenario, the default maximum perturbation magnitude is set to 8/255. We use step sizes of 8/255 for the FGSM attack and 2/255 for the PGD attack. In cases where no specific attack method is mentioned, we default to using the PGD-40 attack, where the "40" indicates that the attack involves 40 iterations.

Attack process

Using the PGD attack algorithm as an example, the algorithmic process for attacking a camouflage object detection model is as follows (Table 2):

Main results of Attack-COSM

We first introduce the experimental setting in detail, and then test the attack effect of the algorithm in white box and black box setting with reverse mask target and background as attack targets. Finally, we use our attack framework to realize the expansion of the mask and the generation of the specified shape mask.

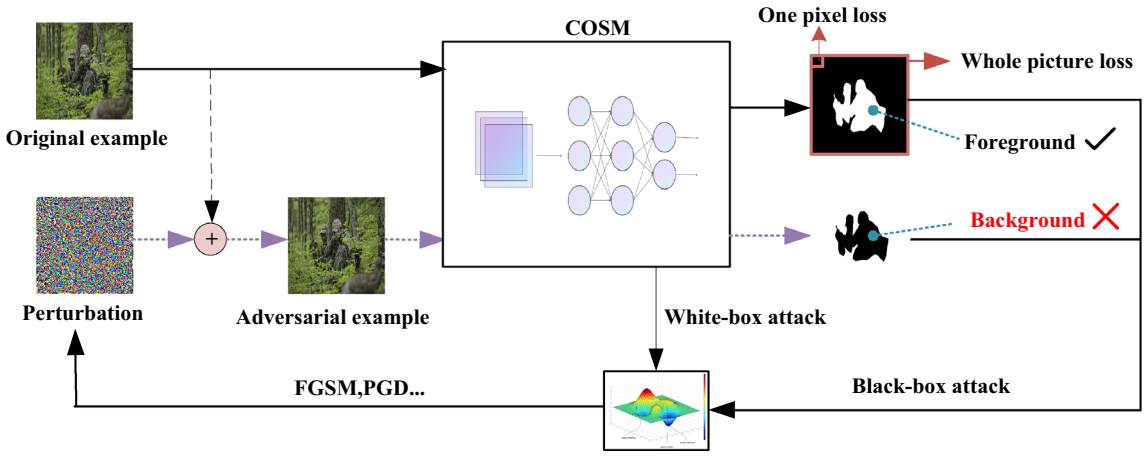


Fig. 3 COSM adversarial example generation process

Table 2 Flow table of ATTACK-COSM algorithm

Algorithm 1

Input: A COSM $\mathcal{F}_\Theta$ with loss function J_s; a real example I_i and ground-truth label C_i;

Input: Maximum allowable perturbation size ϵ; iterations T; step sizes α

Output: An adversarial example I_i^* with $\|I_i^* - I_i\|_\infty \leq \epsilon$

1: $I_i^* = I_i$

2: for $t = 0$ to $T - 1$ do

3: Input I_i^* to $\mathcal{F}_\Theta$ and obtain the gradient $\nabla_{I_i} J_s(\mathcal{F}_\Theta(I_i^*; \Theta), C_i)$

4: Update $I_i^*(t+1)$ by applying the sign gradient as

$$I_i^*(t+1) = I_i^*(t) + \alpha \cdot sign(\nabla_{I_i} J_s(\mathcal{F}_\Theta(I_i^*(t); \Theta), C_i)) \quad (4)$$

5: end for

6: return $I_i^* = I_i^*(T)$

Experimental setup

The experiments were conducted on Intel(R) Xeon(R) Platinum 8260 CPU@2.40 GHz × 96 and RTX 5000 platforms.

COSM

In the white-box setting, we used the FGSM [17] and PGD [18] attack methods to carry out adversarial attacks on the SINet model by generating adversarial examples. Subsequently, we performed transferability tests (black-box testing) on six representative COSM algorithms. These six representative camouflaged object detection algorithms are listed in Table 3.

Table 3 Introduction of the six representative camouflaged object detection models

COSM	Pub./year	Backbone	Representativeness
Single-task learning			
C2FNet [28]	IJCAI'21	Res2Net-50	Attention guidance
PFNet [15]	CVPR'21	ResNet-50	Biomimetic framework
Multitask learning			
JCSOD [29]	CVPR'21	ResNet-50	Conducting confidence estimation
LSRNet [11]	CVPR'21	ResNet-50	Conducting localization/ranking
BGNet [33]	IJCAI'22	Res2Net-50	Conducting boundary
POPNet [27]	arXiv'23	Res2Net-50	Conducting learning depth

Dataset

We conducted evaluations on two benchmark datasets: CAMO [25] and COD10K [9]. CAMO comprises 1250 camouflaged images spanning various categories, with 1000 images designated for training and 250 for testing. On the other hand, COD10K is presently the largest benchmark dataset, featuring 5066 camouflaged images sourced from multiple photography websites. It includes 3040 images for training and 2026 for testing, covering 5 superclasses and 69 subclasses. In line with prior research [9], we utilized the combined training sets of CAMO and COD10K, totaling 4040 images, and the testing set of COD10K for our evaluations.

Evaluation metrics

In the experiment, we employed four well-established evaluation metrics.

Structure measure (S_α) [49] is used to measure the structural similarity between a non-binary prediction map Y and a ground-truth mask C:

$$S_\alpha = (1-\alpha)S_o(\mathrm{Y},\,\mathrm{C}) + \alpha S_r(\mathrm{Y},\,\mathrm{C}) \tag{5}$$

where α balances the object-aware similarity S_o and region-aware similarity S_r. we set $\alpha = 0.5$.

MAE (**mean absolute error,** M) is a conventional pixel-wise measure, which is defined as:

$$M = \frac{1}{W \times H}\sum_{x}^{W}\sum_{y}^{H}|Y(x,\,y) - C(x,\,y)| \tag{6}$$

where (x, y) are pixel coordinates in C.

Enhanced-alignment measure (E_ϕ) [50] is a recently proposed binary foreground evaluation metric, which considers both local and global similarity between two binary maps. Its formulation is defined as follows:

$$E_\phi = \frac{1}{W \times H}\sum_{x}^{W}\sum_{y}^{H}\varphi[Y(x,\,y),\,C(x,\,y)] \tag{7}$$

where φ is the enhanced-alignment matrix.

Weighted F-measure (F_β^ω) [51] can be defined as:

$$F_\beta^\omega = (1+\beta^2)\frac{P^\omega \cdot R^\omega}{\beta^2 \cdot P^\omega + R^\omega} \tag{8}$$

P^ω represents weighted Precision, which measures exactness, while R^ω denotes weighted Recall, measuring completeness. β indicates the effectiveness of detection concerning a user who assigns β times as much importance to R^ω as to P^ω.

Main results under white-box settings

As part of the basic setup, we initially attacked COSM with the objective of reversing the predictions for masked objects and the background, as discussed in "Attack-COSM" section. The attack is considered successful if the precision of $\mathrm{Mask_{adv}}$ is significantly smaller than the precision of $\mathrm{Mask_{clean}}$.

Qualitative results under white-box settings

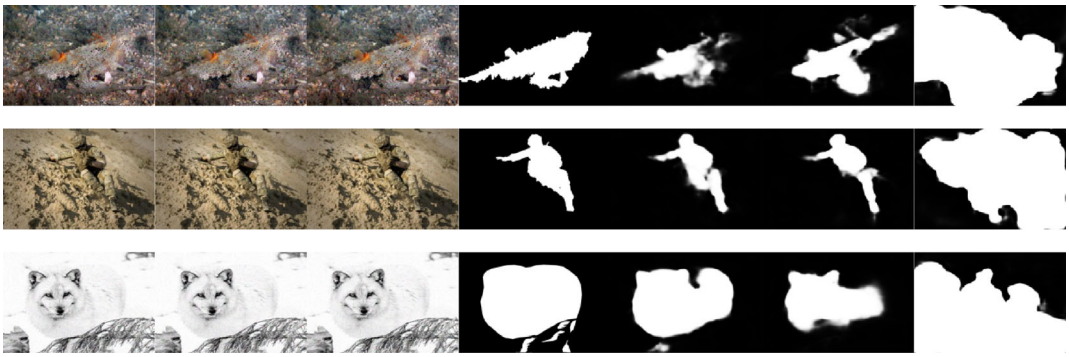
For our white-box attack testing, we selected the first deep learning-based COSM algorithm, the SINet model. We present partial visualization results of adversarial images and predicted masks in Fig. 4. The model is capable of producing adversarial images with imperceptible perturbations following the FGSM and PGD attacks (refer to Fig. 4b, c). While COSM is capable of generating a high-quality $\mathrm{Mask_{clean}}$ in Fig. 4e, both the FGSM and PGD attacks are effective in reversing the predictions for masked objects and the background, particularly the extensive white area of $\mathrm{Mask_{pgd}}$ in Fig. 4g. Figure 4 demonstrates that COSM is susceptible to adversarial attacks, with the PGD attack outperforming the FGSM attack in the context of the COS task. In the experiment, we observed a phenomenon, as seen in Fig. 4g, where the model effectively reverses the prediction of the background into foreground, but the model had poor results when reversing the prediction of the foreground into the background. By examining the values of $Mask_{clean}$, we found that the output value $\mathrm{y_{\rho_o}}$ for the foreground is approximately 20, while the output value $\mathrm{y_{\rho_o}}$ for the background is approximately -5. Therefore, through iterative attacks, the background can be predicted as foreground more quickly.

Quantitative results under white-box settings

We present the evaluation metric results following various attacks in Table 4. With the proposed loss function, the detection accuracy of the SINet model has shown a significant decrease after the PGD attack (e.g.,$\mathrm{E_\varphi}$ drops from 0.817 to 0.279). Although the FGSM attack also results in a noticeable decrease in detection accuracy compared to the original model, the outcome under the FGSM attack is worse than that under the PGD attack due to its weaker attack strength. This indicates that it is difficult for the FGSM attack to cause a significant change in the predicted $\mathrm{y_{\rho_o}}$ and the label value within a single attack step. This result is consistent with the visualization in Fig. 4.

Main results under black-box settings

"Main results under white-box settings" section demonstrates that COSM is vulnerable to adversarial attacks in a full white-box setting. This naturally raises the question: Is COSM resilient to transfer-based attacks? In the black-box setting, the attacker does not have access to all the necessary information when targeting a specific model. In this section, we use the adversarial examples generated by attacking the SINet model in "Main results under white-box settings" section to attack the other six COSM algorithms and assess the performance of transfer-based attacks.



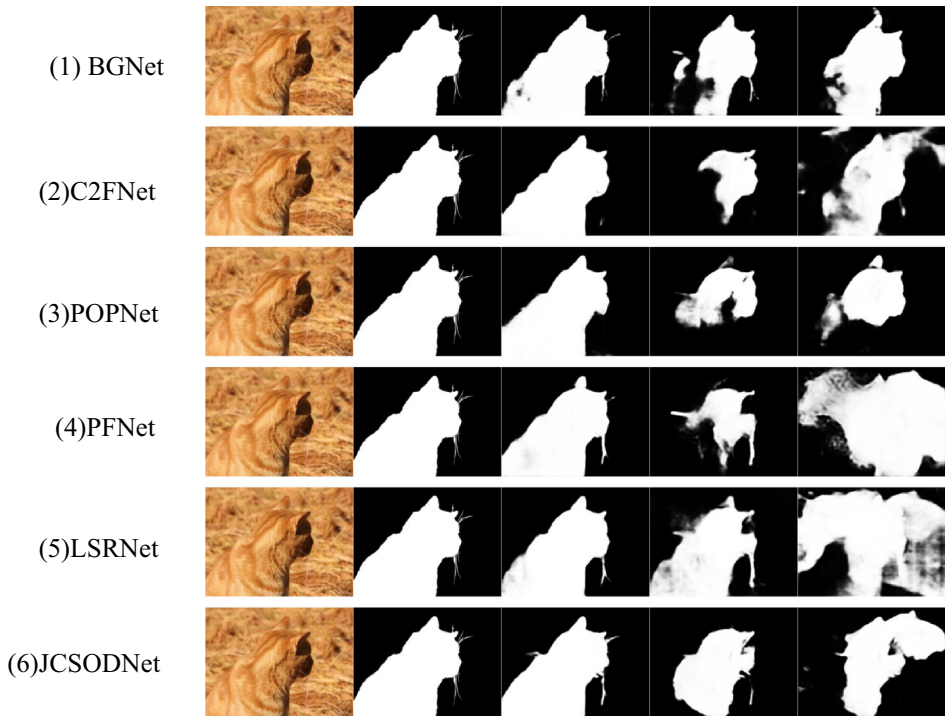
(a) x_{clean} (b) x_{fgsm} (c) x_{pgd} (d) GT (e) $Mask_{clean}$ (f) $Mask_{fgsm}$ (g) $Mask_{pgd}$

Fig. 4 Attacking the SINet model to reverse the predictions of masked objects and background, **a** represents the clean image, **b**, **c** show adversarial images generated by FGSM and PGD attacks, respectively, **e**–**g** represent masks predicted by COSM based on the images shown in **a**–**c**, respectively

Table 4 Results of the change in detection accuracy after attacking the SINet model. Both the FGSM and PGD-40 attacks result in significantly lower detection accuracy compared to the setting with no attack, and PGD-40 results in the lowest detection accuracy

Attack method	Detection accuracy			
	S_α	E_φ	F_β^W	M
SINet				
No attack	0.778	0.817	0.617	0.046
FGSM	0.709	0.771	0.522	0.067
PGD	0.282	0.279	0.145	0.565

Fig. 5 Masks predicted in the cross-model transfer task



(a) x_{clean} (b) GT (c) $Mask_{clean}$ (d) $Mask_{fgsm}$ (e) $Mask_{pgd}$

Qualitative results under black-box settings

As shown in Fig. 5, even though x_{fgsm} and x_{pgd} are generated by attacking the SINet model, they can still successfully attack other models. This can be observed by comparing $Mask_{fgsm}$ and $Mask_{pgd}$ in Fig. 5d, e to $Mask_{clean}$ in Fig. 5c. Comparing $Mask_{pgd}$ in Fig. 5e to $Mask_{fgsm}$ in Fig. 5d, we found that in the COS task, PGD's transfer attack performance is better than that of FGSM.

Quantitative results under black-box settings

We present the evaluation metric results under black-box settings in Table 5. After using the two categories of adversarial examples, x_{fgsm} and x_{pgd}, generated in "Qualitative results under white-box settings" section, to target six representative camouflaged target models, the accuracy of the COSM algorithms decreased. This indicates that adversarial examples have transfer attack capabilities in the COS task. Once more, in the black-box setting for COS, the PGD attack is more effective than the FGSM attack (e.g., the last row, E_φ drops from 0.910 to 0.746 and 0.698 when the input is x_{fgsm} and x_{pgd}, respectively). This result is consistent with the visualization in Fig. 5.

Beyond reversing the predictions of masked objects and background

In the above sections, our primary focus was on reversing the predictions of masked objects and background. Here, we consider a more intriguing scenario, which involves using adversarial examples to generate any desired masks. Conceptually, the goal is to create entirely new masks rather than simply reversing the predictions of masked objects and background, as discussed above.

Mask enlargement

After investigating the mask reverse attack, it is natural to ask whether it is possible to add new masks to a segmentation map. In our preliminary investigation, we begin by attempting to enlarge the mask area without considering the shape or size of the original mask. To enlarge masks through an attack on COSM, the loss design should be aimed at increasing the predicted values y_{ρ_o} until they become positive. To mitigate the randomness effect, the goal is to ensure that the predicted values y_{ρ_o} are significantly higher than zero, rather than just slightly higher. To achieve this, the mean squared error (MSE) loss with a positive threshold is a suitable choice.

We define $\mathcal{H}_\Theta(I_i; \Theta) = y$, where y represents the output value y_{ρ_o} for all pixels of each image and $\text{Sigmoid}(\mathcal{H}_\Theta(I_i; \Theta)) = \mathcal{F}_\Theta(I_i; \Theta)$. As shown in Eq. 9, the predicted value $\mathcal{H}_\Theta(I_i + \delta; \Theta)$ is optimized to be close to a positive threshold P_t after the attack. In the extreme case where $\mathcal{H}_\Theta(I_i + \delta; \Theta) = P_t$ for all predicted values y, the MSE loss reaches its minimum: zero.

$$\delta^* = \min_{\delta \in \mathbb{S}} \|\mathcal{H}_\Theta(I_i + \delta; \Theta) - P_t\|^2 \tag{9}$$

The parameter P_t is set to 20 in this experiment. We visualize the result of mask enlargement in Fig. 6. The experimental results in Fig. 6 show that the mask of adversarial images $Mask_{pgd-160}$ is much larger than $Mask_{clean}$, as seen in Fig. 6g, c. This indicates that the adversarial attack is capable of not only reversing the predictions of the mask and background but also enlarging them. This motivates us to explore attacking COSM to generate any desired mask.

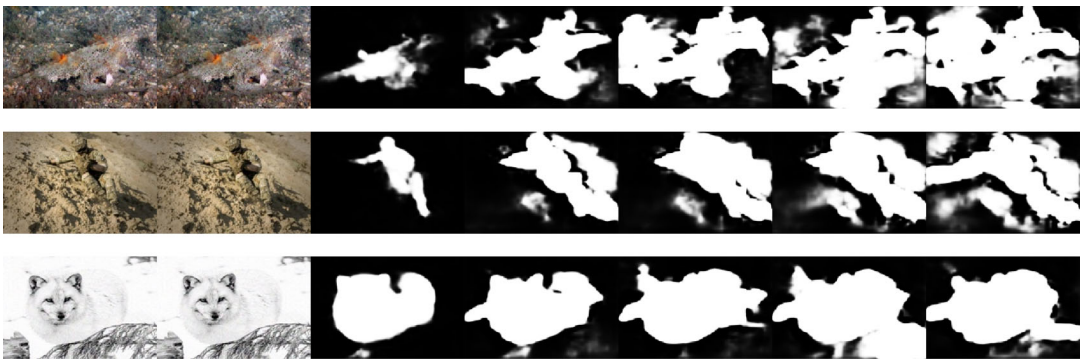
Generating any desired mask

Setting 1: Manually designed mask at a random position In this setting, we explore whether an adversarial attack can generate manually desired masks at random positions. To maintain generality, we design masks in the form of geometric shapes, including circles and squares. Figure 7 illustrates that this goal can be achieved by setting the mask target as a circle and square at a random position when generating x_{pgd}^{circle} and x_{pgd}^{square} in Fig. 7a, b, respectively. Although the input image expects a mask of a fish, as in $Mask_{clean}$ of Fig. 7c, the desired circle or square masks can be obtained in $Mask_{adv}^{circle}$ and $Mask_{adv}^{square}$. Manually designing more complex masks than circles or squares can be challenging. Therefore, we further explore using the real object masks generated by COSM as the target masks to attack COSM (see Setting 2).

Setting 2: A mask generated on a different image We investigate this setting with two example images in Fig. 8. Taking the first row of Fig. 8 as an example, a tuna mask in Fig. 8c is predicted based on the clean images x_{clean} in Fig. 8a. We take the batfish mask in Fig. 8d from the second row as the mask target and attack the x_{clean} image of the tuna in Fig. 8a, and the resulting adversarial image x_{adv} of the tuna is shown in Fig. 8b. Interestingly, a batfish mask, $Mask_{adv}$, is predicted in Fig. 8e based on x_{adv} from the tuna image in Fig. 8b. A similar observation can also be made in the second row of Fig. 8, that is, predicting a tuna mask in $Mask_{adv}$ in Fig. 8e based on x_{adv} from the batfish image in Fig. 8b.

Table 5 Results of the change in detection accuracy under black-box settings. Both the FGSM and PGD-40 attacks result in significantly lower detection accuracy compared to the setting with no attack, and PGD-40 results in the lowest detection accuracy

Model	No attack				x_{fgsm}				x_{pgd}			
	S_α	E_φ	F_β^W	M	S_α	E_φ	F_β^W	M	S_α	E_φ	F_β^W	M
C^2FNet	0.808	0.882	0.697	0.036	0.735	0.829	0.604	0.054	0.667	0.760	0.495	0.087
PFNet	0.801	0.877	0.685	0.038	0.728	0.813	0.577	0.058	0.440	0.500	0.216	0.326
JCSOD	0.817	0.893	0.716	0.033	0.741	0.824	0.604	0.057	0.420	0.480	0.221	0.372
LSRNet	0.811	0.871	0.690	0.036	0.718	0.770	0.534	0.068	0.360	0.493	0.198	0.459
BGNet	0.831	0.902	0.739	0.032	0.752	0.847	0.629	0.049	0.691	0.785	0.526	0.077
POPNet	0.851	0.910	0.771	0.027	0.704	0.746	0.539	0.077	0.656	0.698	0.467	0.099



（a）x_{clean} （b）x_{pgd} （c）$Mask_{clean}$（d）$Mask_{pgd-40}$（e）$Mask_{pgd-80}$（f）$Mask_{pgd-120}$

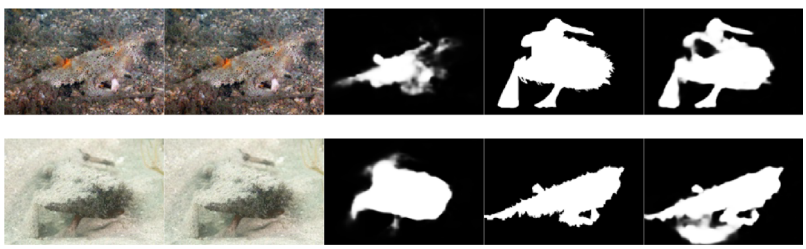
（g）$Mask_{pgd-160}$

Fig. 6 Results of the mask enlargement attack. $Mask_{clean}$ in **c** and $Mask_{adv}$ in **d**–**g** are generated on x_{clean} and x_{pgd} in **a**, **b**, respectively. The results demonstrate that the mask predicted by COSM can be enlarged through the adversarial attack. With the number of iterations increasing from 40 to 160, the effectiveness of the attack improves

(a) x_{pgd}^{circle} (b) x_{pgd}^{square} (c) $Mask_{clean}$ (d) $Mask_{circle}$ (e) $Mask_{adv}^{circle}$ (f) $Mask_{square}$ (g) $Mask_{adv}^{square}$

Fig. 7 Generating any desired masks (Setting 1). $Mask_{adv}^{circle}$ and $Mask_{square}$ in **e**, **g** are generated from x_{pgd}^{circle} and x_{pgd}^{square} in **a**, **b**, respectively. With the adversarial attack, manually designed masks in **d**, **f** can be generated at random positions

Fig. 8 Generating any desired masks (Setting 2)



(a) x_{clean} (b) x_{adv} (c) $Mask_{clean}$ (d) $Mask_{target}$ (e) $Mask_{adv}$

Discussion

Attack goals: label prediction versus mask prediction

In contrast to existing works that mainly focus on attacking the model to change label predictions, our work investigates how to attack COSM to alter the mask predictions of camouflaged targets. Conceptually, our investigation to reverse the predictions of masked and background and generate any desired mask is conceptually similar to a targeted attack setting.

Limitations

We propose an adversarial exmple attack framework tailored for the COS task. Leveraging our framework, we conducted numerous experimental setups. The results of these experiments confirmed the susceptibility of COSM to adversarial example attacks, thereby highlighting its weak robustness. Some of the experimental results also point the way for future research. For example, in the field of COS, PGD has better attack capability than FGSM in both white box and black box Settings. Subsequent research should aim to uncover its internal mechanisms, laying the groundwork for the development of more potent attack methods. In some challenging scenarios, such as the mask enlargement task, success is only partial when the number of iterations is less than 160. However, increasing the number of iterations requires more time. To address this issue, future research could explore ways to enhance attack performance by designing a more effective loss function. Furthermore, we have only explored the robustness of COSM in the digital realm and have not investigated its robustness in the physical world, accounting for factors such as lighting, weather, and sensor influences. In our next steps, we will conduct further research in the domain of physical adversarial attacks.

Conclusion

Our work represents the first investigation into attacking COSM with adversarial examples. In the full white-box setting, we discovered that COSM is vulnerable, as we successfully reversed the predictions of masked objects and background. We also experimented with cross-model transferability and found that the adversarial examples generated by attacking the SINet model can successfully be used to attack other models. In addition to the fundamental goal of reversing the predictions of masked objects and background, we aim to generate any desired mask, achieving an overall satisfactory level of success. Our primary aim is not to discover the most potent method for attacking COSM. Instead, we concentrate on the adaptation of common attack methods, transitioning from attacking label prediction to targeting mask prediction, to assess the robustness of COSM against adversarial examples. The discovery that COSM is susceptible to adversarial examples underscores the importance of investigating the security implications of deploying COSM in safety–critical applications. In the future, we will continue to explore from the following aspects: (1) The attack method in this paper does not consider the attack migration, and the black box attack capability will be studied; (2) Deeply explore the impact of attack parameters on attack performance; (3) Research on defense technology to improve the robustness of the COSM; (4) Study the robustness of COSM in the field of physical attack.

Acknowledgements We would like to thank Dr. Hongbao Du who provided insightful feedback throughout the research process. We express our gratitude to Dr. Zhide Zhang for his advice on the experimental scheme.

Funding This work was supported by the national level Frontier Artificial Intelligence Technology Research Project (Approval No. 672020109).

Data availability Our code can be accessed publicly on the following website: http://github.com/justin-gif/Aaatck-cosm.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Ji GP, Xiao G, Chou YC, Fan DP, Zhao K, Chen G, Van Gool L (2022) Video polyp segmentation: a deep learning perspective. Mach Intell Res 19(6):531–549. https://doi.org/10.1007/s11633-022-1371-y
2. Fan DP, Zhou T, Ji GP, Zhou Y, Chen G, Fu H, Shen J, Shao L (2020) Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans Med Imaging 39(8):2626–2637. https://doi.org/10.1109/tmi.2020.2996645
3. Srivastava A, Singhal V, Aggarawal AK (2017) Comparative analysis of multimodal medical image fusion using PCA and wavelet

transforms. Int J Latest Technol Eng Manag Appl Sci (IJLTEMAS) 6:1
4. Liu L, Wang R, Xie C, Yang P, Wang F, Sudirman S, Liu W (2019) PestNet: an end-to-end deep learning approach for large-scale multi-class pest detection and classification. IEEE Access 7:45301–45312. https://doi.org/10.1109/ACCESS.2019.2909522
5. Rizzo M, Marcuzzo M, Zangari A, Gasparetto A, Albarelli A (2023) Fruit ripeness classification: a survey. Artif Intell Agric 7:44–57. https://doi.org/10.1016/j.aiia.2023.02.004
6. Aggarwal AK (2022) Biological Tomato Leaf disease classification using deep learning framework. Int J Biol Biomed Eng 16(1):241–244
7. Chu HK, Hsu WH, Mitra NJ, Cohen-Or D, Wong TT, Lee TY (2010) Camouflage images. ACM Trans Graph 29(4):51:51-51:58. https://doi.org/10.1145/1778765.1778788
8. Fan DP, Ji GP, Xu P et al (2023) Advances in deep concealed scene understanding. Vis Intell 1(1):16. https://doi.org/10.48550/arXiv.2304.11234
9. Fan DP, Ji GP, Sun G et al (2020) Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2777–2787. https://doi.org/10.1109/CVPR42600.2020.00285
10. Fan DP, Ji GP, Cheng MM, Shao L (2021) Concealed object detection. IEEE Trans Pattern Anal Mach Intell 44(10):6024–6042. https://doi.org/10.1109/TPAMI.2021.3085766
11. Lv Y, Zhang J, Dai Y et al (2021) Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11591–11601. https://doi.org/10.1109/CVPR46437.2021.01142
12. Ghiasi M, Niknam T, Wang Z et al (2023) A comprehensive review of cyber-attacks and defense mechanisms for improving security in smart grid energy systems: past, present and future. Electric Power Syst Res 215:108975. https://doi.org/10.1016/j.epsr.2022.108975
13. Ghiasi M, Ghadimi N, Ahmadinia E (2019) An analytical methodology for reliability assessment and failure analysis in distributed power system. SN Appl Sci 1(1):44. https://doi.org/10.1007/s42452-018-0049-0
14. Zhang R, Du Y, Shi P et al (2023) ST-MAE: robust lane detection in continuous multi-frame driving scenes based on a deep hybrid network. Complex Intell Syst 9(5):4837–4855. https://doi.org/10.1007/s40747-022-00909-0
15. Mei H, Ji GP, Wei Z et al (2021) Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8772–8781. https://doi.org/10.1109/CVPR46437.2021.00866
16. Ji GP, Fan DP, Chou YC, Dai D, Liniger A, Van Gool L (2023) Deep gradient learning for efficient camouflaged object detection. Mach Intell Res 20(1):92–108. https://doi.org/10.1007/s11633-022-1365-9
17. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. Preprint arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
18. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. Preprint arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083
19. Chen G, Liu SJ, Sun YJ, Ji GP, Wu YF, Zhou T (2022) Camouflaged object detection via context-aware cross-level fusion. IEEE Trans Circuits Syst Video Technol 32(10):6981–6993. https://doi.org/10.1109/TCSVT.2022.3178173
20. Li A, Zhang J, Lv Y, Liu B, Zhang T, Dai Y (2021) Uncertainty-aware joint salient object and camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10071–10081. https://doi.org/10.1109/CVPR46437.2021.00994
21. Yang F, Zhai Q, Li X, Huang R, Luo A, Cheng H, Fan DP (2021) Uncertainty-guided transformer reasoning for camouflaged object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4146–4155. https://doi.org/10.1109/ICCV48922.2021.00411
22. Liu J, Zhang J, Barnes N (2022) Modeling aleatoric uncertainty for camouflaged object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1445–1454. https://doi.org/10.1109/WACV51458.2022.00267
23. Zhang M, Xu S, Piao Y, Shi D, Lin S, Lu H (2022) Preynet: preying on camouflaged objects. In: Proceedings of the 30th ACM international conference on multimedia, pp 5323–5332. https://doi.org/10.1145/3503161.3548178
24. Lv Y, Zhang J, Dai Y, Li A, Barnes N, Fan DP (2023) Towards deeper understanding of camouflaged object detection. IEEE Trans Circuits Syst Video Technol 333:3462–3476. https://doi.org/10.1109/TCSVT.2023.3234578
25. Le TN, Nguyen TV, Nie Z, Tran MT, Sugimoto A (2019) Anabranch network for camouflaged object segmentation. Comput Vis Image Underst 184:45–56. https://doi.org/10.1016/j.cviu.2019.04.006
26. Xiang M, Zhang J, Lv Y, Li A, Zhong Y, Dai Y (2021) Exploring depth contribution for camouflaged object detection. Preprint arXiv:2106.13217. https://doi.org/10.48550/arXiv.2106.13217
27. Wu Z, Paudel DP, Fan DP, Wang J, Wang S, Demonceaux C, Timofte R, Van Gool L (2023) Source-free depth for object pop-out. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1032–1042. https://doi.org/10.48550/arXiv.2212.05370
28. Zhai Q, Li X, Yang F, Chen C, Cheng H, Fan DP (2021) Mutual graph learning for camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12997–13007. https://doi.org/10.1109/CVPR46437.2021.01280
29. Zhuge M, Lu X, Guo Y, Cai Z, Chen S (2022) CubeNet: X-shape connection for camouflaged object detection. Pattern Recogn 127:108644. https://doi.org/10.1016/j.patcog.2022.108644
30. Ji GP, Zhu L, Zhuge M, Fu K (2022) Fast camouflaged object detection via edge-based reversible re-calibration network. Pattern Recogn 123:108414. https://doi.org/10.1016/j.patcog.2021.108414
31. Zhu H, Li P, Xie H, Yan X, Liang D, Chen D, Wei M, Qin J (2022) I can find you! boundary-guided separated attention network for camouflaged object detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 3608–3616. https://doi.org/10.1609/aaai.v36i3.20273
32. Zhou T, Zhou Y, Gong C, Yang J, Zhang Y (2022) Feature aggregation and propagation network for camouflaged object detection. IEEE Trans Image Process 31:7036–7047. https://doi.org/10.1109/TIP.2022.3217695
33. Sun Y, Wang S, Chen C, Xiang TZ (2022) Boundary-guided camouflaged object detection. In: Proceedings of the 31st international joint conference on artificial intelligence, pp 1335–1341. https://doi.org/10.24963/ijcai.2022/186
34. Zhu J, Zhang X, Zhang S, Liu J (2021) Inferring camouflaged objects by texture-aware interactive guidance network. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 3599–3607. https://doi.org/10.1609/aaai.v35i4.16475
35. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. Preprint arXiv:1312.6199. https://doi.org/10.48550/arXiv.1312.6199
36. Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Li C, Yang J, Su H, Zhu J (2023) Grounding dino: marrying dino with grounded pre-training for open-set object detection. Preprint arXiv:2303.05499. https://doi.org/10.48550/arXiv.2303.054-99

37. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2020) An image is worth 16 × 16 words: transformers for image recognition at scale. Preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
38. Benz P, Ham S, Zhang C, Karjauv A, Kweon IS (2021) Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs. Preprint arXiv:2110.02797. https://doi.org/10.48550/arXiv.2110.02797
39. Bhojanapalli S, Chakrabarti A, Glasner D, Li D, Unterthiner T, Veit A (2021) Understanding robustness of transformers for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10231–10241. https://doi.org/10.48550/arXiv.2103.14586
40. Mahmood K, Mahmood R, Van Dijk M (2021) On the robustness of vision transformers to adversarial examples. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7838–7847. https://doi.org/10.1109/ICCV48922.2021.00774
41. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (SP), pp 39–57. https://doi.org/10.1109/SP.2017.49
42. Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, Li J (2018) Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9185–9193. https://doi.org/10.1109/CVPR.2018.00957
43. Xie C, Zhang Z, Zhou Y, Bai S, Wang J, Ren Z, Yuille AL (2019) Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2730–2739. https://doi.org/10.1109/CVPR.2019.00284
44. Liu Y, Chen X, Liu C, Song D (2016) Delving into transferable adversarial examples and black-box attacks. Preprint arXiv:1611.02770. https://doi.org/10.48550/arXiv.1611.02770
45. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: attacks and defenses. Preprint arXiv:1705.07204. https://doi.org/10.48550/arXiv.1705.07204
46. Wu D, Wang Y, Xia ST, Bailey J, Ma X (2020) Skip connections matter: on the transferability of adversarial examples generated with resnets. Preprint arXiv:2002.05990. https://doi.org/10.48550/arXiv.2002.05990
47. Guo Y, Li Q, Chen H (2020) Backpropagating linearly improves transferability of adversarial examples. Adv Neural Inf Process Syst 33:85–95
48. Zhang C, Benz P, Karjauv A, Cho JW, Zhang K, Kweon IS (2022) Investigating top-k white-box and transferable black-box attack. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15085–15094. https://doi.org/10.1109/CVPR52688.2022.01466
49. Fan DP, Cheng MM, Liu Y, Li T, Borji A (2017) Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp 4548–4557. https://doi.org/10.1109/ICCV.2017.487
50. Fan DP, Ji GP, Qin X, Cheng MM (2021) Cognitive vision inspired object segmentation metric and loss function. Sci Sin Inform 51(6):1475. https://doi.org/10.1360/SSI-2020-0370
51. Margolin R, Zelnik-Manor L, Tal A (2014) How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2014.39
52. Dong Y, Liao F, Pang T et al (2018) Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9185–9193. https://doi.org/10.1109/CVPR.2018.00957
53. Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2574–2582. https://doi.org/10.1109/CVPR.2016.282
54. Moosavi-Dezfooli SM, Fawzi A, Fawzi O et al (2017) Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1765–1773. https://doi.org/10.1109/CVPR.2017.17
55. Athalye A, Carlini N, Wagner D (2018) Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: International conference on machine learning. PMLR, pp 274–283
56. Xie C, Zhang Z, Zhou Y et al (2019) Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2730–2739. https://doi.org/10.1109/CVPR.2019.00284
57. Guo C, Gardner J, You Y et al (2019) Simple black-box adversarial attacks. In: Proceedings of the 36th international conference on machine learning, PMLR, pp 2484–2493
58. Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput 23(5):828–841. https://doi.org/10.1109/TEVC.2019.2890858
59. Xiao C, Li B, Zhu JY et al (2018) Generating adversarial examples with adversarial networks. Preprint arXiv:1801.02610. https://doi.org/10.48550/arXiv.1801.02610