International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02098-4

Generate Transferable Adversarial Physical Camouflages via Triplet Attention Suppression

Jiakai Wang[1] · Xianglong Liu[1,2] · Zixin Yin[2] · Yuxuan Wang[2] · Jun Guo[2] · Haotong Qin[2] · Qingtao Wu[3] · Aishan Liu[2]

Received: 6 September 2022 / Accepted: 22 April 2024

Abstract
Deep learning models are vulnerable to adversarial examples. As one of the most threatening types for practical deep learning systems, physical adversarial examples have received extensive attention in recent years. However, due to the insufficient focus on intrinsic characteristics such as model-agnostic features, existing studies generate adversarial perturbations with unsatisfactory transferability on attacking different models. Motivated by the viewpoint that attention reflects the intrinsic characteristics of the recognition process, we propose the Transferable Attention Attack (TA_2) method to generate adversarial camouflages with strong transferable attacking ability by taking advantage of visual attention mechanism, i.e., triplet attention suppression. As for attacking, we generate transferable adversarial camouflages by distracting the model-shared similar attention patterns from the target to non-target regions, therefore promoting the transferable attacking ability. Furthermore, we enhance the attacking ability by converging the model attention of the non-ground-truth class, which exploits the lateral inhibition of visual models and activates the model perception for wrong classes. Besides, considering the visually suspicious appearance, we also introduce human attention to help improve their visual naturalness. We conduct extensive experiments in both the digital and physical worlds for classification tasks and comprehensively investigate the effectiveness of the discovered model attention mechanism, demonstrating that our method outperforms state-of-the-art methods.

Keywords Physical adversarial camouflage · Model attention distraction · Lateral inhibition mechanism · Human attention evasion

1 Introduction

Deep neural networks (DNNs) have achieved remarkable performance across wide areas of applications, e.g., computer vision (Krizhevsky et al., 2012; Jin et al., 2021; Jia et al., 2021; Li et al., 2021), natural language (Sutskever et al., 2014), and acoustics (Mohamed et al., 2012), etc, but they are vulnerable to *adversarial examples* (Szegedy et al., 2013). These elaborately designed perturbations are imperceptible to humans but can easily lead DNNs to wrong predictions, which pose a strong security challenge to deep learning applications in both the digital and physical world (Goodfellow et al., 2014; Eykholt et al., 2018; Liu et al., 2020; Zhang et al., 2021).

Communicated by Oliver Zendel.

✉ Xianglong Liu
xlliu@buaa.edu.cn

Jiakai Wang
wangjk@mail.zgclab.edu.cn

Zixin Yin
yzx835@buaa.edu.cn

Yuxuan Wang
yxwang1231@buaa.edu.cn

Jun Guo
junguo@buaa.edu.cn

Haotong Qin
qinhaotong@buaa.edu.cn

Qingtao Wu
wqt8921@haust.edu.cn

Aishan Liu
liuaishan@buaa.edu.cn

[1] Zhongguancun Laboratory, Beijing, China

[2] Beihang University, Beijing, China

[3] Henan University of Science and Technology, Henan, China

Published online: 01 June 2024

In the past years, several works have been proposed to perform adversarial attacks in different scenarios under diverse settings (Kurakin et al., 2017; Dong et al., 2018; Athalye et al., 2017). Though bringing significant challenges for deep learning, adversarial examples are also valuable for understanding the behaviors of DNNs, which provide insights into the blind spots and help to construct robust models (Ilyas et al., 2019; Tsipras et al., 2019; Li et al., 2021; Zhang et al., 2020). Generally, adversarial attacks can be divided into two categories: *digital attacks*, which attack DNNs by perturbing the input data in the digital space; and *physical attacks*, which attack DNNs by modifying the visual characteristics of the real object in the physical world. In contrast to the attacks in the digital world (Jia et al., 2019; Xie et al., 2019; Inkawhich & Wen, 2019; Zhang et al., 2019), adversarial attacks in the physical world faces great challenges since the complex physical constraints and conditions (e.g., lighting, distance, camera, etc), which impairs the attacking ability of generated adversarial perturbations (Elsayed et al., 2018). In this paper, we mainly focus on the challenging physical world attack task, which is significantly meaningful to deploying deep learning applications in practice.

Though several attempts have been adopted to perform physical attacks (Liu et al., 2020; Huang et al., 2020; Liu et al., 2019), existing works pay insufficient attention to the intrinsic characteristics, such as model-agnostic and human-specific patterns. Thus, there is still a significant distance to satisfying adversarial attacking ability among different models. Especially, the limitations can be summarized as (1) the existing methods ignore the common patterns among models and generate adversarial perturbations using model-specific clues (e.g., gradients and weights of a specific model), which fails to attack across different target models. In other words, the transferability of adversarial perturbations is weak, which impairs their attacking abilities in the physical world; (2) current methods generate adversarial perturbations with a visually suspicious appearance which is poorly aligned with human perception and even attracts human attention. For example, painted on the adversarial camouflage (Huang et al., 2020), the classifier misclassifies the car into a bird. However, as shown in Fig. 1a, the camouflage apparently contains bird-like but not natural features (e.g., bird head, bird eyes), which attracts human attention.

To address the above-mentioned problems, this paper proposes the Transferable Attention Attack (TA_2) by exploiting both the model and human attention for generating transferable and visually-natural adversarial camouflages. Regarding the *attacking ability*, inspired by the biological observation that cerebral activities between different individuals share similar patterns when stimulus features are encountered (Evans et al., 1999) (i.e., selected attention (Tricoche et al., 2020)), we perform transferable adversarial attacks by suppressing the attention patterns shared among different models. Specifically, we distract the model-shared similar attention from target to non-target regions inside the model attention maps via exploiting the idea of connected graphs. Thus, influenced models will be misclassified by not paying attention to the objects in the target regions. Since our generated adversarial camouflage captures model-agnostic structures, it can transfer among different models, which improves the transferability. Based on the above elaboration, we further exploit the model attention with a novel approach to improve the *attacking ability*, i.e., considering the lateral inhibition mechanism, which is the ability of adjacent receptors to inhibit each other (Blakemore et al., 1970). In practice, we achieve this goal by converging the wrong model attention and activating the model predictions to false class (i.e., non-ground-truth or target class), so-called lateral attention inhibition. Since this special design is also based on biological observation (i.e., the lateral inhibition mechanism), it can be easily combined with the model attention distraction strategy and achieve better attacking ability. As for the *visual naturalness*, psychologists have found that the bottom-up attention of human vision will alert people to salient objects (e.g., distortion) (Connor et al., 2004). Thus, we try to evade this human-specific visual attention by generating adversarial camouflage which contains a high semantic correlation to scenario context. As a result, the generated camouflage can get rid of the visually suspicious appearance (e.g., salient to human perception) and be natural in terms of human vision. Fig. 1c is the adversarial camouflage generated by CAMOU (Zhang et al., 2019) which is suspicious to human vision due to the semantic missing. By contrast, our generated adversar-

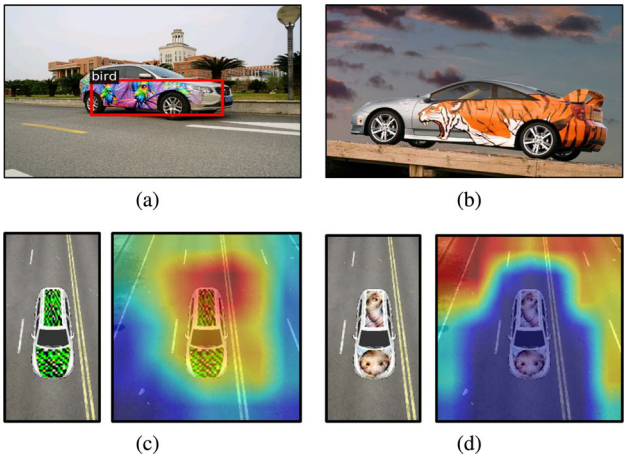

Fig. 1 **a** shows the unnatural appearance of camouflages generated by previous work (i.e., UPC (Huang et al., 2020)). **b** is the painted car that commonly exists in the physical world. **c** shows the adversarial example (classified as `pop bottle`) generated by existing work (i.e., CAMOU (Zhang et al., 2019)) and its corresponding attention map. **d** shows the adversarial example (classified as `Shih-Tzu`) generated by our TA_2 and its confused attention map

ial camouflage is able to yield a more natural appearance with the semantic information "smile face", as shown in Fig. 1d.

Note that we extend our prior conference publication, which mainly concentrated on dual attention suppression, i.e., model-shared attention distraction and human-specific attention evasion. In this paper, we enhance our previous *Dual Attention Attack* (DAS) framework by further investigating the latent attention mechanism (i.e., class lateral inhibition) of the DNN architecture. Inspired by the human lateral inhibition observation and designed a class-wise lateral inhibition strategy, we are surprised to find that the attacking ability of the adversarial camouflages is significantly improved. Based on the improvements, we conducted more extra experiments and provided more in-depth analysis. Also, we introduce several more novel compared methods as baselines in our experiments. The experimental results demonstrate the superiority of our proposed method and reveal some insights for the current DNN architecture. Moreover, we have to highlight the contributions of this paper in (1) the attention mechanism in deep models is investigated further, which can be a solid foundation for the community to understand the model behavior, and (2) Benefited from the model-shared attention attack, we provide a more powerful attacking approach to perform transferable adversarial attacks.

To sum up, our main contribution can be concluded as follows:

- To the best of our knowledge, we are the first to explore and exploit the shared attention characteristics among models for generating adversarial camouflages.
- We comprehensively investigate the shared model attention, once again providing more evidence about the latent correlations between DNN models and human vision systems and revealing the possibility of jointly exploiting various attention characteristics.
- We conduct extensive experiments in both the digital and physical worlds on the basic classification task, demonstrating that our proposed method outperforms other SOTA methods.

The structure of the paper is illustrated as follows: Sect. 2 introduces the related works; Sect. 3 describes the proposed framework and methodology; Sect. 4 demonstrates the effectiveness of the proposed method by thorough experiments; Sect. 5 provides some additional discussions and suggestions; and Sect. 6 summarizes the whole contributions and provides the conclusion.

2 Related Work

Adversarial examples are elaborately designed perturbations that are imperceptible to humans may mislead DNNs (Szegedy et al., 2013; Goodfellow et al., 2014). In the past years, a long line of works has been proposed to develop adversarial attack strategies (Kurakin et al., 2018; Eykholt et al., 2018; Liu et al., 2019; Wei et al., 2019; Duan et al., 2020; Liu et al., 2020), (Zhang et al., 2019; Huang et al., 2020; Wang et al., 2022; Duan et al., 2022; Suryanto et al., 2022). In general, there are several ways to categorize adversarial attack methods, e.g., targeted or un-targeted attacks, white-box or black-box attacks, etc. Based on the domain in which the adversarial perturbations are produced and employed, adversarial attacks can be divided into digital attacks and physical attacks.

Digital attacks generate adversarial perturbations for input data in the digital pixel domain. Szegedy et al. (2013) first introduced adversarial examples and used the L-BFGS method to generate them. By leveraging the gradients of target models, Goodfellow et al. proposed the Fast Gradient Sign Method (FGSM) Goodfellow et al. (2014) which can generate adversarial examples quickly. Moreover, Madry et al. (2017) proposed Projected Gradient Decent (PGD), which is currently the strongest first-order attack. Based on the gradient information, a series of attack approaches have been proposed (Kurakin et al., 2018; Dong et al., 2018; Xie et al., 2019; Wu et al., 2020; Dong et al., 2019). For example, Wu et al. exploit the back-propagated gradients to approximate the model attention and attack the focused region (Wu et al., 2020). And the Kazemi introduce the structural-enhancing algorithms to allow for larger distortions size than common $\updownarrow_p$ counterpart compared with PGD (Kazemi et al., 2023). Although these attacks achieve substantial results in the digital world, their attacking abilities degenerate significantly when introduced into the physical world.

On the other hand, physical attacks aim to generate adversarial perturbations by modifying the visual characteristics of the real object in the physical world (Duan et al., 2020; Athalye et al., 2017; Zhang et al., 2019; Eykholt et al., 2018; Huang et al., 2020), (Wang et al., 2022; Duan et al., 2022; Suryanto et al., 2022). To achieve the goal, several works first generate adversarial perturbations in the digital world, then perform physical attacks by painting the adversarial camouflage on the real object or directly create the perturbed objects. By constructing a rendering function, Athalye et al. (2017) generated 3D adversarial objects in the physical world to attack classifiers. However, this exploratory work only aims at the white-box attack, ignoring the transferable attacking requirements in practice. Recently, Huang et al. (2020) proposed the Universal Physical Camouflage Attack (UPC), which crafts camouflage by jointly fooling the region proposal network and the classifier. Aiming at universal attack, UPC shows weak black-box attacking performance and an unnatural appearance. Duan et al. (2020) generate adversarial examples with natural style while showing insufficient tranferability. Besides, this study also adopts huge perturba-

tion strength, i.e., imposing adversarial noises on the whole image. Wang et al. (2022) propose a robust Full-coverage Camouflage Attack (FCA) to fool detectors and Duan et al. (2022) propose the Coated Adversarial Camouflage (CAC) to attack the detectors in arbitrary viewpoints. These two works adopt a relevant big perturbation region (full camouflage coverage on objects) in 3D environments, making them not so practical and feasible in real scenarios. This year, Suryanto et al. (2022) propose the Differentiable Transformation Attack (DTA) for generating a robust physical adversarial pattern on a target object to camouflage with a wide range of transformations. However, this study is full short of the consideration of visual naturalness, leading to suspicious camouflage. And Zhang et al. (2022) propose the Transferable Physical Attack (TPA) to generate physically adversarial textures with separable attention. Another line of work tries to perform physical adversarial attacks by generating adversarial textures in 2D, i.e., patches (Brown et al. 2017), which confine the noise to a small and localized patch without perturbation constraint (Liu et al., 2019, 2020; Feng et al., 2021). For example, Feng et al. (2021) is a pioneer study that introduces the few-shot learning into physical adversarial examples generation and active considerable attacking ability on unseen models and unseen class. While, this attack perturbs the textures in a full image, making it not flexible in the real world. Overall, the 2D-oriented physical adversarial attacks are not well-suitable for real scenarios due to the special characteristics of physical space.

To sum up, despite achieving certain results, existing methods still show shortages in effectively balancing transferable attacking ability and visually-natural appearance. We believe that it is valuable to further investigate the transferable adversarial camouflages for more satisfying physical adversarial attacking performance, not only for more feasible attacking but also for further understanding the model behaviors.

3 Framework

In this section, we first provide the definition of the problem and then elaborate on our proposed framework.

3.1 Problem Definitions

Given a deep neural network $\mathbb{F}_\theta$ and an input clean image $\mathbf{I}$ with the ground truth label y, an adversarial example $\mathbf{I}_{adv}$ in the *digital world* can make the model conduct wrong predictions as follows:

$$\mathbb{F}_\theta(\mathbf{I}_{adv}) \neq y \quad s.t. \quad \|\mathbf{I} - \mathbf{I}_{adv}\| < \epsilon, \tag{1}$$

where $||\cdot||$ is a distance metric to quantify the distance between the two inputs $\mathbf{I}$ and $\mathbf{I}_{adv}$ sufficiently small.

In the *physical world*, let $(\mathbf{M}, \mathbf{T})$ denote a 3D real object with a mesh tensor $\mathbf{M}$, a texture tensor $\mathbf{T}$, and ground truth y. The input image $\mathbf{I}$ for a deep learning system is the rendered result of the real object $(\mathbf{M}, \mathbf{T})$ with environmental condition $e \in \aleph$ (e.g., camera views, distance, illumination, etc.) from a renderer $\mathcal{R}$ by $\mathbf{I} = \mathcal{R}((\mathbf{M}, \mathbf{T}), e)$, where the $\aleph$ is the environmental set. To perform physical attacks, we generate $\mathbf{I}_{adv} = \mathcal{R}((\mathbf{M}, \mathbf{T}_{adv}), e)$ through replacing the original $\mathbf{T}$ with an adversarial texture tensor $\mathbf{T}_{adv}$, which has different physical properties (e.g., color, shape). Thus the definition of our problem can be depicted as:

$$\mathbb{F}_\theta(\mathbf{I}_{adv}) \neq y \quad s.t. \quad \|\mathbf{T} - \mathbf{T}_{adv}\| < \epsilon, \tag{2}$$

where we ensure the naturalness of the generated adversarial camouflage in the physical world by ϵ.

In this paper, we mainly discuss adversarial attacks in the physical world and generate an adversarial camouflage (i.e., texture), which is able to fool the real deep learning systems when it is painted or overlaid on a real object.

3.2 Framework Overview

To generate visually-natural physical adversarial camouflage with strong transferability and strong attacking ability, we propose a novel adversarial camouflage generation framework based on biology vision principles and mechanisms, in which we exploit both the model and human attention and further exploit the lateral inhibition mechanism. The overall framework can be found in Fig. 2.

Regarding the *transferability for attack*, inspired by the biological observation, we suppress the similar attention patterns shared among models. Specifically, we generate adversarial camouflage by distracting the attention of models from target to non-target regions (e.g., background) via connected graphs. Since different deep models yield similar attention patterns towards the same object, our generated adversarial camouflage can capture the model-agnostic structures and transfer them to different models.

Based on the shared model attention, we further investigate this biological attention mechanism and perform a class lateral inhibition attack to active strong *attacking ability*. In detail, we first converge the model attention to the non-ground-truth class. And then we further activate the latent neural representation of the same class with the attention-converged class. Both strategies are intended to fully investigate and take advantage of the latent similarities between biological neural networks and artificial neural networks. This approach has well compatible with the previous model attention distraction strategy and so as to effectively

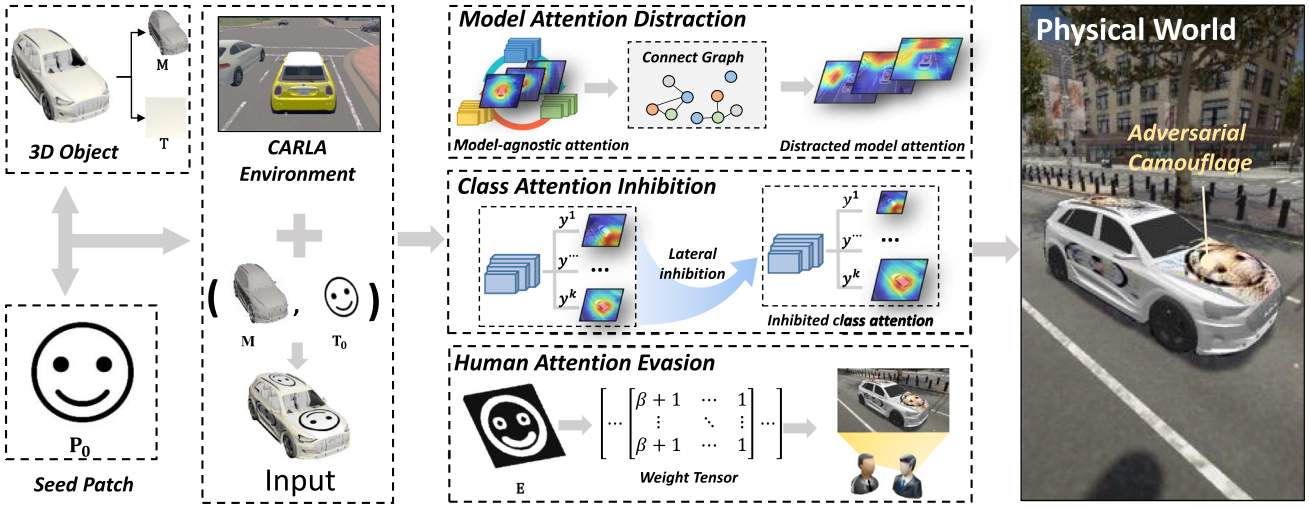

Fig. 2 The framework of our TA_2 method. Based on the intrinsic attention mechanism, we first distract the model attention characteristic by fully exploiting the similar attention patterns of models and forcing the "heat"regions away from the target object with loss function $\mathcal{L}_d$. And then we converge the model attention of the non-ground-truth class to, therefore, further activate the latent neural representation and inhibit that of the ground-truth class for better attacking ability. Besides, we evade the human-specific visual attention mechanism by correlating the appearance of adversaries to the context scenario to generate visually-natural adversarial camouflage

improve the attacking ability of the generated adversarial camouflage.

As for the *visual naturalness*, we aim to evade the human-specific bottom-up attention in human vision (Connor et al., 2004) by generating visually-natural camouflage. By utilizing a seed content patch $\mathbf{P}_0$, which has textures with perceptual correlations to the scenario context, the generated adversarial camouflage, in this case, can be more unsuspicious and natural to human perception by preserve the shape information, to wit, evade the human-specific attention correlations, therefore leading to more natural camouflage.

3.3 Model Attention Distraction

Biologists have found that the same stimulus features (i.e., selected attention) yield similar patterns of cerebral activities among different individuals (Evans et al., 1999) (i.e., similar characteristics of the neuron hyper-perception). Since artificial neural networks are implemented from the human central nervous system (Hentrich, 2015), it is also reasonable for us to assume that DNNs may have the same characteristics, i.e., different models have similar attention patterns towards the same objects when making the same predictions. Based on the above observations, we consider improving the transferability of adversarial camouflages by capturing the model-agnostic attention structures.

Visual attention techniques (Zhou et al., 2016) have been long studied to improve the explanation and understanding of deep learning behaviors, such as CAM (Zhou et al., 2016), Grad-CAM (Ramprasaath, 2017), and Grad-CAM++ (Chattopadhay et al., 2018). When making predictions, a model pays most of its attention to the target objects rather than meaningless parts. Intuitively, to successfully attack a model, we directly distract the attention of models from the salient objects. In other words, we distract the model-shared similar attention map on the salient area to other regions and force the attention weights to distribute uniformly through the entire image. Thus, the model may fail to focus on the target object and make the wrong predictions.

Specifically, given an object ($\mathbf{M}$, $\mathbf{T}$), an adversarial texture tensor $\mathbf{T}_{adv}$ to be optimized, and a certain label y, we get $\mathbf{I}_{adv}$ by $\mathcal{R}$ and then compute the attention map $\mathbf{S}^y$ with an attention module $\mathcal{A}$ as

$$\mathbf{S}^y = \mathcal{A}(\mathbf{I}_{adv}, y). \tag{3}$$

More precisely, the attention module $\mathcal{A}$ is

$$\mathcal{A}(\mathbf{I}, y) = relu\left(\sum_k \sum_i \sum_j \alpha_{ij}^{ky} \cdot relu\left(\frac{\partial p^y}{\partial A_{ij}^k}\right) \cdot A_{ij}^k\right), \tag{4}$$

where α_{ij}^{ky} is the gradient weights for a particular class y and activation map k, p^y is the score of the class y, A_{ij}^k is the pixel value in position (i, j) of the k-th feature map, $relu(\cdot)$ denotes the *relu* function, and the $\cdot$ is a dot multiplication sign. Note that the attention module can be an arbitrary deep model rather than the target model.

Given the attention map $\mathbf{S}^y$ calculated by Eq. (3), we aim to distract the attention region and force the model to focus

on non-target regions. Intuitively, the pixel value of the attention map represents to what extent the region contributes to model predictions. To decrease the attention weights of the salient object and disperse these attention regions, we obtain inspiration from the *connected graph*, which contains a path between any pair of nodes within the graph. In an image, a region with attention weights for each pixel higher than a specific threshold can be deemed as a connected region. We utilize the four-way flood-fill algorithm (Smith, 1979) to obtain the connected regions. Specifically, for each pixel that is unmarked but above the threshold in the attention map $\mathbf{S}^y$, we recursively mark its four-direction neighbors above the threshold as in the same connected regions.

To distract the attention using the connected graph, we consider the following two tasks: (1) decrease the overall connectivity by separating connected graphs into multiple sub-graphs; (2) reduce the weight of each node within a connected sub-graph. To achieve these goals, we propose attention distraction loss as

$$\mathcal{L}_d = \frac{1}{K}\sum_k \frac{G_k}{N - N_k}, \quad s.t. \quad G_k \subseteq \mathbf{S}^y, \tag{5}$$

where G_k is the sum of pixel values in the region corresponding to k-th connected graph in $\mathbf{S}^y$, N is the total pixel number of the $\mathbf{S}^y$, and N_k is the total pixel number of G_k.

By minimizing $\mathcal{L}_d$, the salient region in the attention map becomes smaller (i.e., distracted) and the pixel values of the salient regions become lower (i.e., no longer "heated"), leading to the "distracted" attention map.

3.4 Lateral Attention Inhibition

Lateral inhibition, which inspires a considerable number of deep vision techniques (Tao et al., 2021), is one of the basic principles of information processing and plays an important role during information processing in the nervous system. Due to the fact that the shared model attention draws lessons from biological knowledge, it is reasonable for us to further combine it with the lateral inhibition mechanism and further improve the attacking ability of the adversarial camouflages. Specifically, we attempt to improve the attacking ability of the generated adversarial camouflages through the lateral attention inhibition approach, leading joint attacks with our attention distraction attack strategy in the meantime.

Furthermore, for a deep neural network model, its recognition capability is highly correlated to perceptual ability. Therefore, model attention, which represents the model perception to a certain class, can be exploited to perform additional attacks considering the lateral inhibition mechanism. In general, a DNN model can be regarded as a complex vision system, which exhibits human-like traits of attention (Zhou et al., 2016). Thus, we once again employ the model-shared similar attention but perform an opposite operation, to wit, converge the model attention of a non-ground-truth class. The rationale behind this behavior is to force the model to activate the latent neural representation of the false label, thus paying its attention to the wrong class and influencing the attention of true label in turn, i.e., lateral attention inhibition.

Specifically, to accurately converge the model attention of a non-ground-truth class c, we firstly introduce a region mask m, which indicates the region to be covered by the adversarial camouflage (the region is de facto that of the target object). Then we acquire the saliency map of the object region by a simply element-wise multiplication $\odot$ operation. Thus, the lateral inhibition loss can be formulated as

$$\begin{aligned} \mathcal{L}_l &= \frac{(\mathbf{1} - m) \odot \mathbf{S}^c}{m \odot \mathbf{S}^c}, \\ \mathbf{S}^c &= \mathcal{A}(\mathbf{I}, c), \end{aligned} \tag{6}$$

where the $\mathbf{S}^c$ means the attention map of a non-ground-truth c, in practice, we set the c as the class of the seed patch $\mathbf{P}_0$ under a classifier (i.e., the $c = \mathbb{F}_\theta(\mathbf{P}_0)$), $\mathbf{1}$ is a matrix that all elements are set as 1. In this equation during optimization, the sum value of the target region in the saliency map (i.e., denominator) will be maximized and the sum value of the non-target region (i.e., numerator) will be minimized.

By calculating the above equation, we can converge the model attention of the specific class (i.e., non-target class) into the adversarial camouflage's region. However, the above equation is formulated as a division operation, which causes unnecessary computation costs. Thus, we reformulate the lateral inhibition loss as

$$\mathcal{L}_l = \log(\mathbf{1} - m)\mathbf{S}^c - \log m \odot \mathbf{S}^c \tag{7}$$

where the notations are the same as those in Eq. (6). With this reformulation, we can still activate the model attention to the non-ground-truth class, whereas the calculation cost will obviously decrease.

By exploiting the lateral inhibition mechanism, we converge the model attention of the non-ground-truth class and activate the latent representation of neurons, resulting in a better attacking ability of the generated adversarial camouflage.

3.5 Human Attention Evasion

Previous physical attacks might generate adversarial perturbations with a comparatively huge magnitude (Duan et al., 2020; Wang et al., 2022). Since the bottom-up human attention mechanism can alert people to salient objects (e.g., distortion) (Connor et al., 2004), adversarial examples, in this case, can attract human attention due to the salient perturba-

tions, showing suspicious appearance and lower stealthiness in the physical world.

In this paper, we aim to generate more visually-natural camouflage by suppressing the human visual mechanism, which will evade human-specific attention. Intuitively, we expect the generated camouflage to share similar visual semantics with the context to be attacked (e.g., beautiful paintings on vehicles are more perceptually acceptable for humans than meaningless distortions). Thus, the generated adversarial camouflage can be highly correlated to human perception, which is unsuspicious to human perception.

In particular, we first incorporate a seed content patch $\mathbf{P}_0$ which contains a strong semantic association with the scenario context. We then paint the seed content patch on the vehicle ($\mathbf{M}$, $\mathbf{T}$) by $\mathbf{T}_0 = \Psi(\mathbf{P}_0, \mathbf{T})$. Specifically, $\Psi(\cdot)$ is a transformation operator which first transfers the 2D seed content patch into a 3D tensor, and then paints the car through tensor addition.

Since humans pay more attention to shapes when focusing on objects and making predictions (Liu et al., 2020), we aim to further improve the human attention correlation by better preserving the shape of the seed content patch. Specifically, we obtain the edge patch $\mathbf{P}_{edge} = \Phi(\mathbf{P}_0)$ using an edge extractor Φ (Canny, 1986) from the seed content patch. It should be noticed that $\mathbf{P}_{edge}$ has 0–1 value in each pixel. After that, we simply transform the edge patch $\mathbf{P}_{edge}$ to a mask tensor $\mathbf{E}$ which has the same dimension with $\mathbf{T}_0$.

With mask tensor $\mathbf{E}$, we can distinguish the edge and non-edge regions and limit the perturbations added to the edge regions. Thus, the attention evasion loss $\mathcal{L}_e$ can be formulated as

$$\mathcal{L}_e = \|(\beta \cdot \mathbf{E} + \mathbf{1}) \odot (\mathbf{T}_{adv} - \mathbf{T}_0)\|_2^2, \tag{8}$$

where $\beta \cdot \mathbf{E} + \mathbf{1}$ is the weight tensor, $\mathbf{1}$ is a tensor in which each element is 1 and its dimension is the same with $\mathbf{E}$ and $\odot$ denotes element-wise multiplication.

To further improve the naturalness of the camouflage, we introduce the smooth loss $\mathcal{L}_s$ by reducing the difference square between adjacent pixels (Eykholt et al., 2018). Thus, the generated camouflage in this case will be visually correlated to the scenario context, leading to evading human perceptual attention.

3.6 Overall Optimization Process

Overall, we generate the physical adversarial camouflages through 3 kinds of parallel constraints, which respectively aim to enhance the transferability, attacking ability, and stealthiness to perform strong attacks. And in practice, we generate the adversarial camouflage by jointly optimizing all the 4 constraint loss functions mentioned above, i.e., the model attention distraction loss $\mathcal{L}_d$, the lateral inhibition loss $\mathcal{L}_l$, the human attention evasion loss $\mathcal{L}_e$, and smooth loss $\mathcal{L}_s$. Given a seed patch and a 3D object ($\mathbf{M}$, $\mathbf{T}$) and the corresponding environmental conditions, we can obtain the adversarial camouflage based on the above optimization process.

Specifically, we first distract the target model from the salient objects to the meaningless part (e.g., background), jkand then we force the model to pay attention to the non-ground-truth class by covering the corresponding model attention and activating the latent neural representation. Finally, we evade the human-specific attention mechanism by enhancing the strong perceptual correlation to the scenario context. Thus, we can generate transferable and visually-natural adversarial camouflages with strong attacking ability by minimizing the following formulation as

$$\min \mathcal{L}_d + \mathcal{L}_l + \lambda(\mathcal{L}_e + \mathcal{L}_s), \tag{9}$$

where λ controls the balance of the attacking intensity and natural level.

To balance the attacking ability and appearance naturalness, we set λ as 10^{-4}, and set β as 8 according to our experimental results. The overall training algorithm can be described as Algorithm 1.

Algorithm 1 Transferable Attention Attack (TA_2)

Require: environmental parameter set $\aleph = \{e_1, e_2, ...e_n\}$, 3D real object ($\mathbf{M}$, $\mathbf{T}$), seed content patch $\mathbf{P}_0$, region mask m, neural renderer $\mathcal{R}$, attention model $\mathcal{A}$, specific class label c
Ensure: adversarial texture tensor $\mathbf{T}_{adv}$
  $\mathbf{T}_0 \leftarrow \Psi(\mathbf{P}_0, \mathbf{T})$
  $\mathbf{P}_{edge} \leftarrow \Phi(\mathbf{P}_0)$
  transform $\mathbf{P}_{edge}$ to $\mathbf{E}$
  initial $\mathbf{T}_{adv}$ as $\mathbf{T}_0$
  for the number of epochs **do**
    select $minibatch$ environmental conditions from $\aleph$
    for $m = n/minibatch$ steps **do**
      $\mathbf{I}_{adv} \leftarrow \mathcal{R}((\mathbf{M}, \mathbf{T}_{adv}), e_m)$
      $\mathbf{S}^y \leftarrow \mathcal{A}(\mathbf{I}_{adv}, y)$
      $\mathbf{S}^c \leftarrow \mathcal{A}(\mathbf{I}, c)$
      calculate the $\mathcal{L}_d$, $\mathcal{L}_l$, $\mathcal{L}_e$ and $\mathcal{L}_s$
      optimize the $\mathbf{T}_{adv}$ by Eqn (9)
    end for
  end for

4 Experiments

In this section, we first outline the experimental settings and then illustrate the performance of our proposed attacking framework by thorough evaluations in both the digital and physical worlds. Besides, we discuss the detailed effectiveness of the critical factors in our TA_2 framework.

4.1 Experimental Settings

Virtual Environment and Evaluation Metric To perform a physical world attack, we choose CARLA (Dosovitskiy et al., 2017) as our 3D virtual simulated environment, which is the commonly used open-source simulator for autonomous driving research. Based on Unreal Engine 4, CARLA provides many high-resolution open digital assets, e.g., urban layouts, buildings, and vehicles to simulate a digital world that is nearly the same as the real world. To evaluate the performance of our proposed method, we first select the widely used Accuracy as the metric. Facing adversarial attacks, the Accuracy drop reflects the robustness of the models. The bigger Accuracy drop, the stronger the attacking ability. Besides, we also adopt the Attack Success Rate (ASR), which is widely used in various existing works such as (Wang et al., 2022), to comprehensively validate the proposed method. Moreover, we note that the data points used during evaluations are the same as those in our prior reference work.

Compared methods We choose several state-of-the-art works in the 3D attack and physical attack literature, including UPC (Huang et al., 2020), CAMOU (Zhang et al., 2019), MeshAdv (Xiao et al., 2019), AdvCam (Duan et al., 2020), FCA (Wang et al., 2022), CAC (Duan et al., 2022), TPA (Zhang et al., 2022), and our previous conference work [denoted as "Ours-P" (Wang et al., 2021)]. We use ResNet-50 as the base model for the classic classification task. Note that compared with our previous conference work, we extend 4 recent proposed methods for comparison, i.e., AdvCam, FCA, CAC, and TPA. Moreover, since most of the comparisons are designed for detection tasks, we conduct fair adaptation of their released code in our environments.

Target Models In our previous study, we select commonly used model architectures for experiments. Specifically, Inception-V3 (Szegedy et al., 2016), VGG-19 (Zisserman & Simonyan, 2014), ResNet-152 (He et al., 2016), and DenseNet (Huang et al., 2016) are employed for the classification task. In this paper, beyond the aforementioned models, we additionally introduce some other in-fashion models into evaluation, such as ResNext-101 (Xie et al., 2017), and GoogLeNet (Szegedy et al., 2015). Besides, considering the various architectures in practice, we further introduce the ViT family models (Dosovitskiy et al., 2021) (like ViT-T, ViT-S, and Vit-B) and Deit family models (Touvron et al., 2020) (like Deit-T, Deit-S, and Deit-B). Moreover, we also conduct some exploratory evaluations on the large models, such as (Su et al., 2023; Li et al., 2023, 2022, 2023). It should be also noted that in this paper we mainly consider the classification task in the evaluation due to its elementary and analysis-friendly characteristics. For all the models, we use the pre-trained version on ImageNet. Moreover, since there exist several car types in ImageNet, e.g., `taxi`, and `jeep`, some of them are similar to the simulated cars in CARLA, which motivates us to set `taxi`, `jeep`, `sports car`, `race car`, `convertible`, and `limousine` as the correct class, making it more challenging to perform adversarial attacks.

Implementation details We empirically set $\lambda = 10^{-5}$ for classification task, and we set $\beta = 8$. We adopt an Adam optimizer with a learning rate of 0.01, a weight decay of 10^{-4}, and a maximum of 5 epochs. We employ a seed content patch (e.g., a stick smile face image) as the appearance of the 3D object in the training process. For training details of comparisons, we adopt the same configurations, such as environmental settings, same vehicle models, same replacing texture meshes, and same training data points, with ours for most of the baselines to ensure the evaluation fairness. For CAMOU and UPC, we straightly email the authors of the studies and acquire their generated adversarial textures for evaluation. All of our codes are implemented in PyTorch. We conduct the training and testing processes on an NVIDIA Tesla V100-SXM2-16GB GPU cluster. In the physical world attack scenario, adversaries only have limited knowledge and access to the deployed models (i.e., architectures, weights, etc.). Considering this, we focus on attacks in full black-box settings (i.e., the source model employed for generating adversarial camouflages is totally different from the target models to be attacked), therefore leading to more meaningful and applicable results for physical world applications.

4.2 Digital World Attack

In this section, we evaluate the performance of our generated adversarial camouflages on the vehicle classification task in the digital world under black-box settings.

We randomly select 155 points in the simulation environment to place the vehicle and use a virtual camera to capture 100 images at each point using different settings (i.e., angles, and distances). Specifically, we use different distance values (5, 10, 15, and 20), four camera pitch angle values (22.5°, 45°, 67.5°, and 90°), and eight camera yaw angle values (south, north, east, west, and southeast, southwest, northeast, northwest). We then collect 15,500 simulation images with different setting combinations, and we choose 12,500 images as the training set and 3000 images as the test set. To conduct fair comparisons, we use the backbone of ResNet-50 as attention modules in training. Besides, since some compared methods perform attacks by fully coating the vehicles with adversarial camouflages, we thus constrain the changed textures scope of them, making it the same with ours, i.e., top and surroundings. Also, we additionally introduce a "Smile" texture (the smiling face) and "Naive" texture (a common camouflage texture adopted in UPC) to the 3D vehicle surface as a comparison of common noises. Some of the examples with the generated adversarial camouflages are shown in

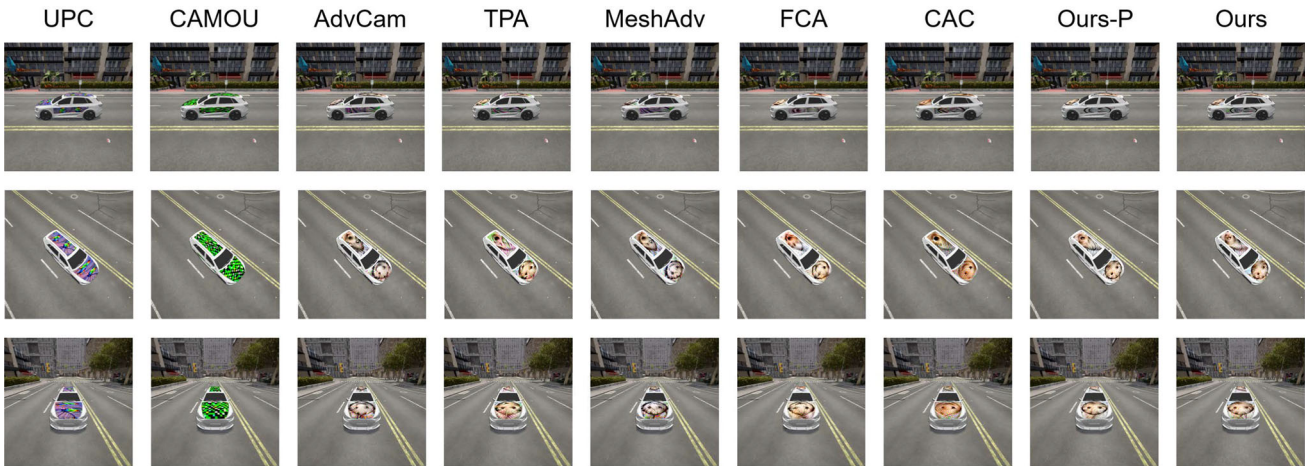

Fig. 3 The generated adversarial example. Note that we only select some sampling angles due to space limitations

Table 1 The experimental results of un-target attacks in the digital world

Method	VGG-19		ResNet-152		ResNext-101		DenseNet		Inception-V3		GoogLeNet	
	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑
Raw	40.62	–	73.51	–	60.18	–	71.91	–	74.36	–	64.76	–
Smile	58.00	6.76	55.02	33.46	50.98	36.52	75.38	13.68	62.22	24.36	72.76	8.43
Naive	39.69	23.08	53.96	34.89	66.40	20.53	71.51	16.63	59.24	27.37	57.33	25.22
UPC (Huang et al., 2020)	38.00	42.02	48.18	41.46	58.89	27.08	65.87	23.53	42.40	44.65	65.11	24.48
CAMOU (Zhang et al., 2019)	31.46	38.61	48.93	38.91	52.71	41.65	57.56	29.97	47.51	41.38	**44.58**	**40.71**
AdvCam (Duan et al., 2020)	36.53	34.55	39.64	51.20	45.16	46.95	56.22	35.59	42.62	46.50	63.91	21.40
TPA (Zhang et al., 2022)	36.93	**45.26**	33.69	**65.86**	42.76	58.31	62.71	31.02	45.42	50.33	63.20	24.29
MeshAdv(Xiao et al., 2019)	32.44	37.93	35.33	58.12	41.78	49.46	58.04	40.60	42.31	49.71	64.04	20.74
FCA(Wang et al., 2022)	40.00	25.99	49.91	37.08	42.53	52.97	70.27	18.12	53.91	32.36	74.76	8.99
CAC(Duan et al., 2022)	39.56	29.27	38.93	50.60	38.40	57.09	70.22	21.17	46.84	41.69	66.53	14.33
Ours-P (Wang et al., 2021)	30.18	43.02	32.49	61.15	32.22	60.19	55.42	38.05	39.86	52.88	59.11	21.20
Ours	**28.89**	**39.97**	**28.58**	**64.30**	**30.36**	**61.73**	**46.71**	**40.80**	**32.58**	**53.99**	**55.29**	**25.02**

Ours and the best performance are in bold font

Fig. 3. The experimental results are illustrated in Table 1, where we can draw several conclusions as follows:

(1) Our adversarial camouflage ("Ours") achieves significantly better performance than the compared baselines in most cases on different models. Overall, our TA_2 outperforms others by large margins, i.e., the average accuracy drop of Ours is 27.15%, and that of UPC, CAMOU, AdvCam, TPA, MeshAdv, FCA, CAC, and Ours-P are respectively 11.15, 17.10, 16.88, 16.77, 18.57, 8.99, 14.14, and 22.68%. Similarly, the average ASR of Ours is 47.63%, while the others are respectively 33.87, 38.54, 39.36, 45.85, 42.76, 29.25, 35.69, 46.08%. More precisely, the maximum accuracy drop of Ours even achieves 44.93% on ResNet-152. Besides, we find that the attacking performances of TPA and FCA are not as strong as they mentioned in the original papers, we attribute it to their special design, i.e., the former aims is designed for object detection tasks with full consideration of distracting multi-layer attention, and the latter adopts a fully coated adversarial attacking strategy (we only replace part of the vehicle textures).

(2) For each model, we also calculate the detailed attacking ability variance. Specifically, the accuracy drops of the proposed TA_2 on VGG-19, ResNet-152, ResNext-101, DenseNet, Inception-V3, and GoogLeNet are respectively 11.73, 44.93, 29.82, 25.20, 41.78, and 9.47%. What's more, compared with Ours-P, the attacking ability improvement of Ours is **4.48%** on average (with a max improvement **8.71%** on DenseNet). It should be noted that the TA2 shows relevantly weak attacking ability on GoogLeNet, a similar observation can also be witnessed on UPC, AdvCam, TPA, MeshAdv, FCA, and CAC. However, the CAMOU appears higher ASR and lower

ACC (i.e., 40.71 and 44.58%), we attribute it to the potential correlation between CAMOU and GoogLeNet, i.e., the CAMOU learns a multi-view learning strategy, such as jointly receiving the perturbation, the background, and the foreground as inputs during training, which similar to the multi-scale technique in GoogLeNet. While for Inception-V3, it adopts more training techniques like label smoothing, and etc, possibly making it lose the original characteristics in GoogLeNet. Overall, we can conclude that the proposed additional class lateral inhibition is effective in promoting the attacking ability of the generated adversarial camouflage. Also, these experimental results indicate that the different visual attention principles might be correlated to each other, thus motivating us to further study the insights behind these observations.

(3) We found that FCA works comparatively worse than other baselines in many cases. For example, the FCA achieves only 25.99, 18.12, and 8.99% ASR on VGG-19, DenseNet, and GoogLeNet, respectively. We conjecture the reason might be that FCA is primarily designed for fully coated adversarial camouflage, which might adapt not well in these partially coated settings. Moreover, the CAC also shows a similar performance to the FCA. By carefully reviewing the correlated works, we found that the CAC is also designed for fully coated adversarial attacks in the 3D environment. Besides, we find that the UPC performs not so well as it performs in the physical world, which should be attributed to its physical elaboration, i.e., physical simulation, therefore showing worse attacking ability in the digital world. By contrast, our TA_2 attack exploits the intrinsic visual attention mechanism to perform stronger attacking, therefore not only achieving considerable attacking ability in the digital world but also keeping acceptable performance in the physical world consistently. That also means, this intrinsic characteristic among deep models plays a critical role during the decision-making process de facto.

4.3 Physical World Attack

As for the physical world attack, we conduct several experiments to validate the practical effectiveness of our generated adversarial camouflages. Due to the limitation of funds and conditions, we print our adversarial camouflages with an HP Color LaserJet Pro MFP M281fdw printer and stick them on a toy car model with different backgrounds to simulate the real vehicle painting. To conduct fair comparisons, we take 960 pictures of the toy car model on various environmental conditions (i.e., 8 directions {left, right, front, back and their corresponding intersection directions}, 3 angles {0°, 45°, 90°}, 4 scenarios {3 indoor conditions, 1 outdoor condition}, and 10 kinds of textures {Raw, 9 compared methods, and Ours }) for each kind of physical camouflages, using a Huawei P40 phone. It should be noted that we did not deliberately search for scenes with significant differences in lighting, but there are natural differences in lighting intensity in the selected background environments. Here we provide some adversarial examples sampled from the physical world as shown in Fig. 4.

Fig. 4 The adversarial examples in the physical world for attacking toy cars. To show the diversity of the sampling environment, we select 4 physical adversarial camouflages that sampled from 3 indoor environments and 1 outdoor environment, with different distances and angles. These adversarial samples are respectively predicted as `mousetrap`, `waffle`, `Mustela nigripes`, and `car wheel`

The evaluation results can be witnessed in Table 2. Compared with other methods, the TA_2 shows competitive transferable attacking ability, which is significantly better than the baselines, including Ours-P. More precisely, Ours achieves ACC of **33.33%** on VGG-19, **20.83%** on ResNet-152, **21.88%** on ResNext-101, **36.46%** on DenseNet, **38.54%** on Inception-V3, and **46.88%** on GoogLeNet, respectively. Moreover, compared with the SOTA method TPA (Zhang et al., 2022) and Our-P, Ours enjoy stronger physical adversarial attacking ability due to the further exploitation of visual attention mechanism. Besides, it should be noted that both the TPA and Ours-P utilize the model attention distraction strategy. However, our TA_2 considers taking advantage of this kind of attention mechanism in a more comprehensive perspective, i.e., attention distraction and class lateral inhibition, which jointly activate strong attacking performance due to the latent correlation inside the model attention mechanism.

In addition, we also test the generated adversarial camouflages on some on-sale devices that are employed for roadblocks surveillance. Specifically, we purchase a Hikvision DS-2CD7167EWD-IZ detector that can take pictures and then capture the vehicles inside the frame for recording. For realizable tests, we adopt various testing settings. Specifically, test the detecting device by sampling from different visual angles, i.e., horizontal, vertical, and inclined. As shown in Fig. 5, we can find that the physical adversarial camouflages are allowed to mislead the unknown models deployed in the on-sale roadblocks surveillance device successfully, under different angles and environmental settings. Beyond the quantitative physical experiments, we believe this physical qualitative experiment can demonstrate the strong physical attacking ability of the proposed TA_2 method convincingly.

Table 2 The experimental results in the physical world

Method	VGG-19		ResNet-152		ResNext-101		DenseNet		Inception-V3		GoogLeNet	
	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑	ACC↓	ASR↑
Raw	60.42	–	50.00	–	58.33	–	61.46	–	57.29	–	61.46	–
UPC (Huang et al., 2020)	46.88	31.03	32.29	45.83	41.67	41.07	42.71	37.29	55.21	18.18	56.25	11.86
CAMOU (Zhang et al., 2019)	45.83	32.76	34.38	45.83	44.79	35.71	50.00	27.12	53.13	20.00	52.08	22.03
AdvCam (Duan et al., 2020)	44.79	31.03	28.13	54.17	38.54	44.64	50.00	25.42	50.00	27.27	54.17	15.25
TPA (Zhang et al., 2022)	40.63	32.76	25.00	54.17	32.29	51.79	43.75	33.90	48.96	27.27	53.13	15.25
MeshAdv (Xiao et al., 2019)	44.79	34.48	27.08	50.00	32.29	50.00	55.21	18.64	50.00	30.91	55.21	13.56
FCA (Wang et al., 2022)	41.67	32.76	27.08	54.17	32.29	58.93	51.04	25.42	47.92	27.27	56.25	10.17
CAC (Duan et al., 2022)	39.58	36.21	25.00	52.08	28.13	57.14	54.17	20.34	53.13	25.45	**46.88**	**25.42**
Ours-P (Wang et al., 2021)	40.63	36.21	25.00	58.33	36.46	50.00	46.88	33.90	39.58	**43.64**	54.17	16.95
Ours	**33.33**	**46.55**	**20.83**	**62.50**	**21.88**	**69.64**	**36.46**	**47.46**	**38.54**	**41.82**	**46.88**	**25.42**

Ours and the best performance are in bold font

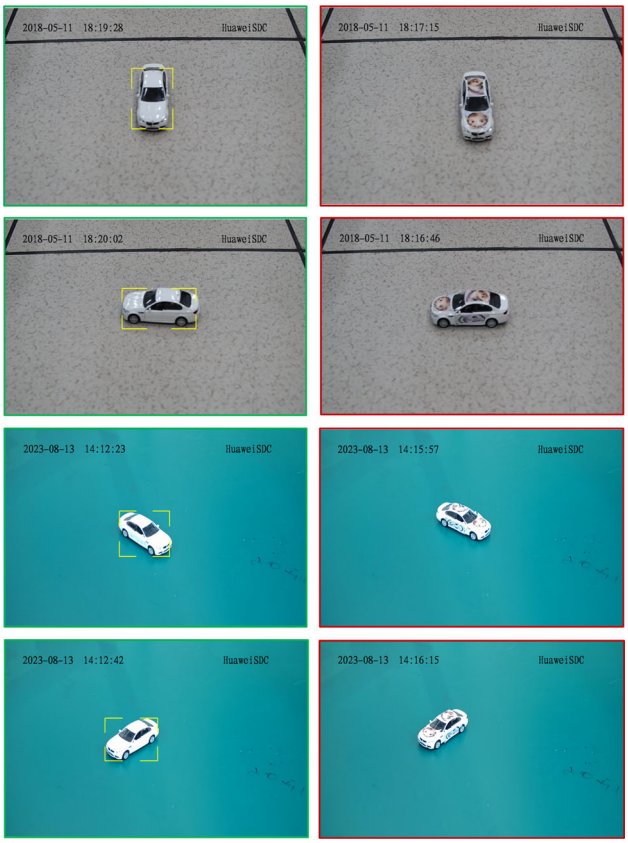

Fig. 5 The experimental example in the physical world by employing the on-sale roadblock surveillance device. We test the transferable attacking ability from different visual angles under indoor conditions

4.4 Human Perception Study

Since one of the critical components in the proposed TA_2, the human attention evasion loss is designed for better visual naturalness, it is necessary for us to evaluate the naturalness of our generated adversarial camouflage. For this purpose, we conduct a human perception study on one of the most commonly used crowdsourcing platforms in mainland, China, i.e., WJX. In detail, we adversarially perturb our 3D car object using different methods (i.e., UPC, CAMOU, AdvCam, TPA, MeshAdv, FCA, CAC, Ours-P, and Ours) and acquire the adversarial textures. Then we replace the texture of the car in specific meshes using these camouflages and get the rendered images for the human perception studies of Naturalness. Specifically, all participants come from the WJX crowdsourcing platforms and independently decide whether to participate in this experiment. Each experimental participant is required to take the experiment seriously and receives a reward of 0.5$ upon completion of the experiment. The participants are asked to score the naturalness of the camouflage from 1 to 10. In particular, we collect the responses from 165 participants.

As shown in Table 3, the average naturalness score is up to **3.982**, which is above the score of the comparison methods due to our human attention evasion mechanism. However, since this improved work mainly focuses on the further mining and utilization of the model's attention to improve the attack capability, the lateral inhibition loss inevitably affects the human attention evasion loss, leading to a decrease in naturalness, which is intuitively manifested as a lower naturalness score than our previous work's adversarial camouflage (Wang et al., 2021), i.e., Ours-P. We believe that there exists a trade-off between aggressiveness and naturalness, which has also been observed by previous works (Duan et al., 2020; Jia et al., 2022). Overall, we think that the proposed TA_2 achieves comparable performance in both naturalness and attack ability.

Table 3 The naturalness results of human perception study

Question	UPC	CAMOU	AdvCam	TPA	MeshAdv	FCA	CAC	Ours-P	Ours
Naturalness	3.909	2.982	3.970	3.897	3.952	3.873	3.691	4.224	3.982

Table 4 The ablation study on attention distraction portion

Models (%)	Setting			
	Raw	$\mathcal{L}_d$	$\mathcal{L}_l$	$\mathcal{L}_d + \mathcal{L}_l$
VGG-19	40.62	30.18	29.38	**28.89**
ResNet-152	73.51	32.49	29.07	**28.58**
ResNext-101	60.18	32.22	32.04	**30.36**
DenseNet	71.91	55.42	49.47	**46.71**
Inception-V3	74.36	39.86	34.71	**32.58**
GoogLeNet	64.76	59.11	56.44	**55.29**

The bold values indicate the best attacking performance (i.e., the lowest accuracy)

4.5 The Effect of Different Loss Terms

Different loss terms play different roles, so we conduct an ablation study to further investigate the effect of loss terms. However, in our previous conference work (Wang et al., 2021), we have drawn some meaningful conclusions about the effectiveness of the model attention distraction loss $\mathcal{L}_d$ and human attention evasion loss $\mathcal{L}_e$, i.e., $\mathcal{L}_d$ mainly provides a transferable attacking ability in our previous DAS method and the human attention evasion provides the natural appearance. Therefore, we only further investigate the effectiveness of the new proposed loss term $\mathcal{L}_l$ and the correlations between attention loss functions, i.e., $\mathcal{L}_d$ and $\mathcal{L}_l$.

Specifically, we optimize the adversarial camouflage using function $\mathcal{L}_d$, $\mathcal{L}_l$, and $\mathcal{L}_d + \lambda\mathcal{L}_l$ respectively (with $\mathcal{L}_e$ and $\mathcal{L}_s$ fixed). As shown in Table 4, the accuracy shows significant drops (i.e., **22.68%** on average under $\mathcal{L}_d$ setting and **25.71%** on average under $\mathcal{L}_l$ setting among the 6 different models), showing the positive influence of the proposed attention mechanism-based loss functions. Further, there is an interesting observation that the $\mathcal{L}_l$ seems has stronger attacking ability compared with $\mathcal{L}_d$, which verifies some visual experience that *attracting one's attention is much easier than distracting*. Besides, we can clearly witness that the attacking ability achieves the peak value when $\mathcal{L}_d$ and $\mathcal{L}_l$ are employed together. That means, the attention mechanism behind deep models plays a critical role during decision-making, and the correlations among different attention principles might be able to be guided into cooperation.

To sum up, this ablation demonstrates that jointly distracting the model attention of the target class and activating the model attention of the non-target class can further improve the attacking ability of adversarial camouflage, revealing that the attention mechanism in deep models is worth investigating.

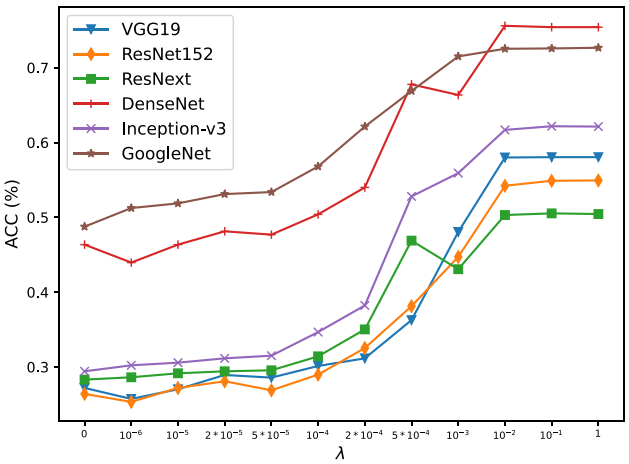

Fig. 6 Ablation on studying the effectiveness of λ. The various lines represent the trend of accuracy change under different λ values

4.6 The Effect of Hyperparameter

Regarding the hyperparameter λ, we believe that it still controls the level of the semantic correlation with the context, i.e., naturalness. Though we have verified this viewpoint in the prior conference work, it is necessary for us to conduct ablations on λ once again due to the updated loss functions that might make some differences. We evaluate the effectiveness of λ on a ResNet-50 model using Accuracy and SSIM. Specifically, we set the λ as $0, 10^{-6}, 10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$, and 1, respectively. As illustrated in Fig. 6, the model accuracy first increases and then keeps a stable value as λ increases. For all different models, although the details are not identical, the trend remains basically consistent. From the results, we can draw the conclusion that λ controls the attacking intensity of the generated adversarial camouflage, i.e., when λ gets bigger, the accuracy gets bigger, which means the lower attacking ability and better appearance according to our prior conclusion. And in this paper, we set the λ as 10^{-4} based on the ablation.

5 Discussion

In this section, we discuss the discovered model attention mechanism and fully investigate the correlations among attention under different labels. Also, we conduct experiments on studying the targeted attacking ability and the transferability of the adversarial camouflages on more different models We believe that this will provide more insights

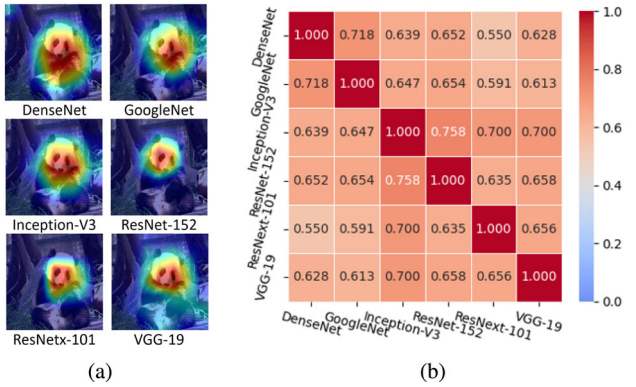

Fig. 7 **a** is the attention map on 6 different models to a particular image. **b** is a heat map drawn according to the SSIM values

for us and promote the understanding of deep neural networks.

5.1 The Shared Attention among Models

In this part, we conduct a detailed analysis of attention through both qualitative and quantitative studies to validate the effectiveness of the model attention distraction.

Firstly, we conduct a qualitative study by visualizing the attention regions of 6 different models toward the same image on several layers. As shown in Fig. 7a, different DNNs show similar attention patterns towards the same image. And for different layers, this observation is consistent. That means different models pay their attention to similar regions, inspiring us that the shared attention can be deemed as a model-agnostic characteristic.

We then conduct a quantitative study by calculating the structural similarity index measure (SSIM) (Wang et al., 2004), which is a well-known quality metric used to measure the similarity between two images (Horé & Ziou, 2010). Specifically, we generate the attention maps of a specific image (i.e., Panda) on all 6 different models and calculate the SSIM values between each pair of the attention maps on different models. As shown in Fig. 7b, different models demonstrate comparatively high similarities of the attention maps.

In addition, we further visualize the attention maps by changing the model predictions (i.e., class). As shown in Fig. 8, when changing the class label, the attention map is distracted from the salient objects and becomes more sparse over the entire image. Finally, we visualize the attention differences before and after attacks as shown in Fig. 9. It can be observed that the attention of the model is distracted away from the salient regions.

In summary, we can draw several conclusions as follows: (1) different DNNs show similar attention patterns towards the same class in a specified image; (2) we can attack a DNN

Fig. 8 The visualization of attention maps on the same image `panda` with different target labels using ResNet-152. The attention maps differ significantly when different target labels are provided to the model

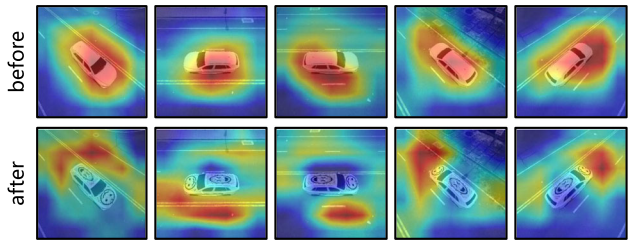

Fig. 9 The attention maps before and after attacking, which indicated that the model attention is distracted significantly. Note that we only employ the proposed distraction loss, evasion loss, and smooth loss for getting rid of the influence of lateral inhibition loss

adversarially to wrong predictions by distracting its attention under a specific class label, i.e., the ground-truth label.

5.2 The Correlation among Classes

In this part, we further investigate the latent correlation between different classes by activating the model attention under different class labels in turn, therefore excavating the insights behind the model attention mechanism.

Specifically, we try to study the latent correlation of model attention with different classes from 2 aspects, i.e., (1) activate the attention to non-ground-truth class then observe the model attention differences of ground-truth classes, and (2) activate the model attention to different non-ground-truth classes then observe the detailed variance (i.e., the confidence variance) of ground-truth ones. In detail, we first observe the attention map with the ground-truth label before and after activating the model attention of the randomly selected class labels, i.e., $c =$ `Washing machine`, `Hinder`, `Handkerchief`, `Puma`, and `Kangaroo`. By calculating the saliency map, we provide the model attention variance under the target class label as Fig. 10, where we can conclude that the model attention of different classes will influence each other. More precisely, from the attention maps we find that the salient region of Panda will be confused when the attention of other class is activated on a certain model, i.e., lateral inhibition.

Further, we then investigate the effect of activating the attention of different non-target classes. Here, we select 4 coarse-grained categories, including *vehicle*, *animal*, *plant*, and *human*. For each category, we select 20 classes, the

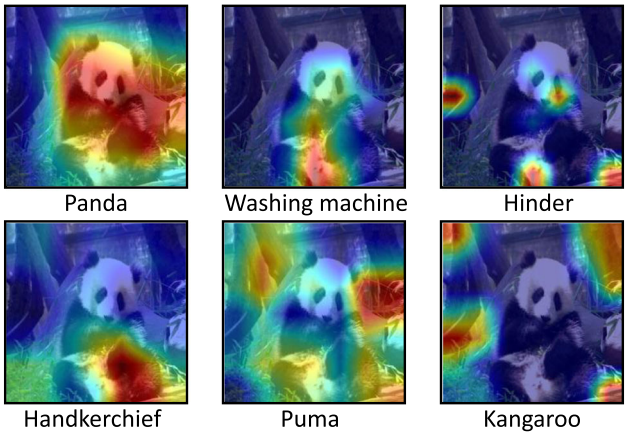

Fig. 10 The variance before and after activating the attention of the non-ground-truth class. The original model attention under the label "Panda" shows significant differences when activating other class labels, e.g., "Washing machine", "Hinder", and "Handkerchief", etc. This observation strongly supports that human visual lateral inhibition is de facto existing in deep models also. Note that we calculate these attention maps on ResNet-152

detailed classes can be found in Table 5. In this study, we quantitatively evaluate the model average confidence of the target class after activating the model attention with non-ground-truth labels. Specifically, we exploit the lateral inhibition loss to activate a specific class with certain iterations and then record the model confidence of the ground-truth class. In detail, we select the panda" as object class and sample 50 images for evaluation. We report the average confidence (50 images) of the class "panda" as shown in Fig. 11a. It can be concluded that for different coarse categories, the lateral inhibition influence is quite different, i.e., the different box plots. More precisely, we conjecture that the similar class (but not the same) might achieve higher lateral inhibition effectiveness, i.e., the "animal" achieves the lower average confidence, due to the possible latent correlations behind the model behavior such as shared attention. Besides, we further report the overall average confidence on all 1000 classes (i.e., 999 non-ground-truth classes) as shown in Fig. 11b. It can be also found that the different classes indeed make differences in the performance of model attention lateral inhibition.

5.3 The Target Attacking Ability

In our main experiments as shown in Sect. 4.2, we only evaluate the untargeted attacking ability of the proposed TA_2 method. Beyond this, the targeted attacking ability of adversarial examples is another focus aspect. For example, compared with misleading a car for a pedestrian, misclassifying it as a straight-ahead sign might cause more serious consequences. Thus, we provide additional results about the

Table 5 The selected non-ground-truth class labels

Category	Classes			
Animal				
Tench	Cock	Ostrich	Goldfinch	Robin
Bald eagle	Tree frog	American lobster	Golden retriever	Doberman
Hyena	Red fox	Tiger cat	Marmot	Zebra
Macaque	Coho	Llama	Ox	Hog
plant				
Guacamole	Head cabbage	Broccoli	Cauliflower	Zucchini
Spaghetti squash	Acorn squash	Butternut squash	Cucumber	Artichoke
Bell pepper	Cardoon	Mushroom	Strawberry	Orange
Lemon	Pineapple	Banana	Custard apple	Pomegranate
Vehicle				
Airliner	Ambulance	Beach wagon	Bicycle-built-for-two	Cab
Electric locomotive	Fireboat	Fire engine	Forklift	Garbage truck
Go-kart	Golfcart	Lifeboat	Minivan	Motor scooter
Mountain bike	Police van	Racer	Recreational vehicle	School bus
tool				
Backpack	Ballpoint	Barrel	Baseball	Basketball
Bathing cap	Beer bottle	Binoculars	Broom	Coffee mug
Goblet	Golf ball	Hammer	Hatchet	Mitten
Mouse	Pencil sharpener	Ping-pong ball	Screwdriver	Soccer ball

All classes can be basically divided into 4 categories

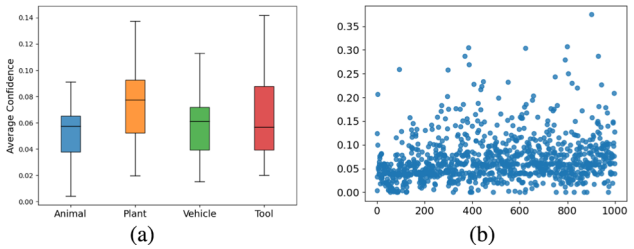

Fig. 11 The investigation of the lateral inhibition mechanism. **a** is the average confidence of class "panda" after activating some selected classes. **b** is the overall average confidence of "panda" after activating non-ground-truth classes

targeted attacking ability of the proposed adversarial attack in this section by misleading the models to predict the input as the designated label.

However, our proposed TA_2 is not designed for performing targeted attacks, since there exists no hard label in our training scheme to mislead models into specific labels (the lateral attention inhibition loss is a soft direction that can not be employed for targeted attack directly). Therefore, we introduce the cross-entropy loss function into the optimization process of TA_2 to mislead the models to predict the target class. Besides, considering the proposed lateral attention inhibition is also correlated with the class labels, we set the attacked class to be the same as the non-ground-truth c. As a result, our TA_2 method obtains a targeted attack success rate of 3.42% in VGG-19, 14.22% in ResNet-152, 15.07% in ResNext-101, 1.51% in DenseNet, 0.09% in Inception-V3, and 2.04% in GoogLeNet. According to the results, we can summarize: (1) the targeted attacking ability of the proposed TA_2 is not very satisfactory. Since the exploited model perception is not fully studied, attacking with this mechanism might not be powerful for targeted adversarial attacks, while also indicating that there is a space to exploit the model perception for targeted attacks. (2) Though showing limited attacking performance, the TA_2 achieves a comparable targeted attacking ability with the AdvCam method, which is the only one that evaluates targeted attacking performance in the baselines. Specifically, under targeted attack, the AdvCam achieves about $2\% \sim 8\%$ ASR under similar settings to ours.

5.4 Attack on Diverse Architectures

Beyond the DNN architectures, there also exist several different but popular model architectures for classification tasks, such as vision transformers (ViTs) and large vision language models (LVLMs). For comprehensively evaluating the transferable attacking ability of the proposed TA_2 framework, it is reasonable for us to conduct the assessment on the mentioned architectures and in turn draw more realizable conclusions.

Table 6 The experimental results in the attacking on vision transformers

Models	ASR↑					
	ViT-T	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B
TA_2	36.56	33.12	32.99	24.20	24.24	25.97

Table 7 The experimental results in the attacking on large vision language models

Models	ASR↑				
	PandaGPT-13b	Otter	Blip2-6.7b	Blip2-2.7b	Blip
TA_2	7.53	36.49	0.18	11.55	29.75

Specifically, considering the time limitations and computation constraints, we select several in-fashion models that have different architectures and can be employed for the classification task. Specifically, for ViTs, we select the ViT-T, ViT-S, ViT-B, Deit-T, Deit-S, and Deit-B. For LVLMs, we select the PandaGPT-13b, Otter, Blip, Blip2-2.7b, and Blip-6.7b. Note that for both ViTs and LVLMs, the selected models include different architectures and parameter numbers. The experimental results are shown in Tables 6 and 7, respectively. To sum up, we draw some conclusions from the aspect of the transferable attacking ability on diverse model architectures, including (1) the proposed TA_2 shows certain transferable attacking capabilities across different model architectures. For example, the ASR of the proposed method achieves 29.51% on average for ViTs and 17.1% on LVLMs. However, compared with the ASR in CNN-based models, the attacking transferability appears significant drop (-18.13% on ViTs and -30.54% on LVLMs), which means that the model attention might exist a bigger difference among CNNs and ViTs/LVLMs. (2) For LVLMs, the larger models show better robustness to the proposed adversarial camouflages. For example, in ViTs, the ASR on Blip2$-$6.7 is much lower than that on Blip2$-$2.7. And both of them show better defense than Blip. We think that this observation indicates that the models with large scales might result in a less easily perturbed model perception, therefore promoting defense against attention-driven attacks.

6 Conclusion

In this paper, we propose the Transferable Attention Attack (TA_2) to generate adversarial camouflage in the physical world inspired by the human attention mechanism. To improve the transferable attacking ability of adversarial camouflages, we first distract the model-shared similar attention from target to non-target regions. For further promoting attacks, we then converge the model attention of

the non-target class to laterally inhibit the model perception to the ground-truth category. To generate more visually-natural camouflage, we suppress human attention by evading human-specific bottom-up attention. We conduct extensive experiments on classification tasks in both the digital and physical world (including evaluations on on-sale surveillance devices) under totally black-box settings. Besides, we also provide several discussions and analyses to help fully understand the proposed method, such as introducing diverse ViTs and LVLMs as attacking models. The experimental results demonstrate that our TA_2 attack shows considerable attacking performance compared with baselines.

In the future, we are interested in investigating the attack abilities of our adversarial camouflage using a real vehicle in a real-world scenario. Using projection or 3D printing, we might simply paint our camouflage on a real-world vehicle. Further, we would also like to investigate the effectiveness of our generated camouflage to improve model robustness against different noises. We believe that the model attention mechanism is worth investigating in other vision scenarios and tasks and promoting a better understanding of explainable artificial intelligence.

Acknowledgements This work was supported by The National Key Research and Development Plan of China (2020AAA0103502), and the National Natural Science Foundation of China (62022009, 61872021).

Data Availability Statement In this paper, we employ the 3D environment to generate training and testing datasets. All the experiments and ablations are based on them. The datasets generated and analyzed during the current study are available at https://drive.google.com/drive/folders/1vspvRxnZ3shOV4kM5ELcO9-xztapBThS. Beyond the data availability, our code will also be released freely after acceptance.

References

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2017). Synthesizing robust adversarial examples. arXiv e-prints arXiv:1707.07397.

Blakemore, C., Carpenter, R. H., & Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature, 228*(2), 37–39.

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665.

Canny, J. (1986). A computational approach to edge detection. In *PAMI, PAMI-8*.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV* (2018).

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology, 14*(19), R850–R852.

Dong, Y., Liao, F., Pang, T., & Su, H. (2018). Boosting adversarial attacks with momentum. In *CVPR*.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *CVPR*.

Dong, Y., Pang, T., Su, H., & and Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *ICLR 2021*.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *CoRL*.

Duan, Y., Chen, J., Zhou, X., Zou, J., He, Z., Zhang, J., Zhang, W., & Pan, Z. (2022). Learning coated adversarial camouflages for object detectors. In L. De Raedt (Ed.), *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 2022* (pp. 891–897). ijcai.org.

Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A. K., & Yang, Y. (2020). Adversarial camouflage: Hiding physical-world attacks with natural styles. In *CVPR*.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *NeurIPS*.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *CVPR*.

Feng, W., Wu, B., Zhang, T., Zhang, Y., & Zhang, Y. (2021). Meta-attack: Class-agnostic and model-agnostic physical adversarial attack. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7787–7796).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hentrich, M. (2015). Methodology and coronary artery disease cure. SSRN 2645417.

Horé, A., & Ziou, D. (2010). *Image quality metrics: PSNR vs*. In *ICPR SSIM*.

Huang, L., Gao, C., Zhou, Y., Xie, C., Yuille, A. L., Zou, C., & Liu, N. (2020). Universal physical camouflage attacks on object detectors. In *CVPR*.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2016). Densely connected convolutional networks. https://doi.org/10.48550/arXiv.1608.06993

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In *NeurIPS*.

Inkawhich, N., Wen, W., Li, H. H., & Chen, Y. (2019). Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7066–7074).

Jia, Y., Lu, Y., Velipasalar, S., Zhong, Z., & Wei, T. (2019). Enhancing cross-task transferability of adversarial examples with dispersion reduction. arXiv preprint arXiv:1905.03333.

Jia, W., Li, L., Li, Z., & Liu, S. (2021). Deep learning geometry compression artifacts removal for video-based point cloud compression. *International Journal of Computer Vision, 129*(11), 2947–2964.

Jia, S., Yin, B., Yao, T., Ding, S., Shen, C., Yang, X., & Ma, C. (2022). Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *Advances in Neural Information Processing Systems, 35*, 34136–34147.

Jin, H., Liao, S., & Shao, L. (2021). Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision, 129*(12), 3174–3194.

Kazemi, E., Kerdreux, T., & Wang, L. (2023). Minimally distorted structured adversarial attacks. *International Journal of Computer Vision, 131*(1), 160–176.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99–112). Chapman and Hall/CRC.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2017). Adversarial examples in the physical world. In *ICLR workshop*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900). PMLR.

Li, T., Liu, A., Liu, X., Xu, Y., Zhang, C., & Xie, X. (2021). Understanding adversarial robustness via critical attacking route. *Information Sciences, 547*, 568–578.

Li, H., Tao, R., Li, J., Qin, H., Ding, Y., Wang, S., & Liu, X. (2021). Multi-pretext attention network for few-shot learning with self-supervision. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). IEEE.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C. and Liu, Z. (2023). Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726.

Liu, A., Huang, T., Liu, X., Xu, Y., Ma, Y., Chen, X., Maybank, S. J., & Tao, D. (2020). Spatiotemporal attacks for embodied agents. In *ECCV*.

Liu, A., Liu, X., Fan, J., Ma, Y., Zhang, A., Xie, H., & Tao, D. Perceptual-sensitive GAN for generating adversarial patches. In *AAAI*.

Liu, A., Wang, J., Liu, X., Zhang, C., Cao, B., & Yu, H. (2020). Patch attack for automatic check-out. In *ECCV*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Mohamed, A., Dahl, G. E., & Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(1), 14–22.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Smith, A. Ray. (1979). Tint fill. *SIGGRAPH. Computer Graphics, 13*(2), 276–283.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., & Cai, D. (2023). Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355.

Suryanto, N., Kim, Y., Kang, H., Larasati, H. T., Yun, Y., Le, T. T. H., Yang, H., Oh, S. Y., & Kim, H. (2022). Dta: Physical camouflage attacks using differentiable transformation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 15305–15314).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NeurIPS*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tao, R., Wei, Y., Li, H., Liu, A., Ding, Y., Qin, H., & Liu, X. (2021). Over-sampling de-occlusion attention network for prohibited items detection in noisy x-ray images. arXiv preprint arXiv:2103.00809.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. arXiv:2012.12877.

Tricoche, L., Ferrand-Verdejo, J., Pélisson, D., & Meunier, M. (2020). Peer presence effects on eye movements and attentional performance. In *Front Behav Neurosci*.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. In *ICLR*.

Wang, D., Jiang, T., Sun, J., Zhou, W., Gong, Z., Zhang, X., Yao, W., & Chen, X. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2414–2422).

Wang, J., Liu, A., Yin, Z., Liu, S., Tang, S., & Liu, X. (2021). Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 8565–8574).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Wang, D., Jiang, T., Sun, J., Zhou, W., Gong, Z., Zhang, X., Yao, W., & Chen, X. (2022). Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(2), 2414–2422.

Wei, X. S., Cui, Q., Yang, L., Wang, P., & Liu, L. (2019). RPC: A large-scale retail product checkout dataset. arXiv preprint arXiv:1901.07249.

Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M. R., & Tai, Y. W. (2020). Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1161–1170).

Xiao, C., Yang, D., Li, B., Deng, J., & Liu, M. (2019). Meshadv: Adversarial meshes for visual recognition. In *CVPR*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks, 2017.

Xie,Cihang, Zhang,Zhishuai, Zhou,Yuyin, Bai,Song, Wang,Jianyu, Ren,Zhou, Yuille,Alan L. (2019). Improving transferability of adversarial examples with input diversity. In *CVPR*.

Zatorre, R. J., Mondor, T. A., & Evans, A. C. (1999). Auditory attention to space and frequency activates similar cerebral systems. In *Neuroimage*.

Zhang, Y., Foroosh, H., David, P., & Gong, B. (2019). CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*.

Zhang, Y., Gong, Z., Zhang, Y., Li, Y., Bin, K., Qi, J., Xue, W., & Zhong, P. (2022). Transferable physical attack against object detection with separable attention. CoRR arXiv:2205.09592.

Zhang, C., Liu, A., Liu, X., Xu, Y., Yu, H., Ma, Y., & Li, T. (2020). Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing, 30*, 1291–1304.

Zhang, X., Qin, H., Ding, Y., Gong, R., Yan, Q., Tao, R., Li, Y., Yu, F., & Liu, X. (2021) Diversifying sample generation for data-free quantization. In *IEEE CVPR*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. Learning deep features for discriminative localization. In *CVPR*.

Zisserman, A., & Simonyan, K. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.