Unnoticed Yet Effective: A Hybrid Physical Camouflage Framework Against DNNs and Human Perception

Mingye Xie, Jiacheng Ruan, Xian Gao, Ting Liu, Yuzhuo Fu

Shanghai Jiao Tong University

{xiemingye, jackchenruan, gaoxian, louisa_liu, yzfu}@sjtu.edu.cn

Abstract

While adversarial attacks can effectively deceive deep neural networks, their real-world applicability is often limited by complex and conspicuous patterns that reveal their attack intent to human observers. To overcome this limitation, we propose UYE, a novel camouflage framework designed to simultaneously mislead DNNs and evade human perception. UYE incorporates two key components: an attention refiner leveraging a pre-trained vision encoder to optimize adversarial patterns for robust attacks across diverse environments, and a perception evaluator trained on a preference dataset curated using tailored prompts from human-aligned large multimodal models to ensure natural and unobtrusive camouflage generation. Extensive experiments demonstrate that UYE outperforms state-of-the-art methods in achieving an optimal balance between human stealth and model deception while maintaining effectiveness in real-world scenarios.

Code — https://github.com/MyronXie/UYE

Introduction

Camouflage, inspired by the natural strategies animals use to evade predators, aims to conceal an object to make it less noticeable, primarily optimized for human vision. Figure 1 illustrates several cases. One intuitive approach is background matching, where the object's surface mimics the surrounding environment (Owens et al. 2014), blending seamlessly into the scene and making it difficult for humans to detect. Additionally, Road-test vehicles are often wrapped in dizzy patterns to hide their design, but these tend to attract excessive attention that they defeat the purpose, clearly signaling a deliberate attempt at camouflage. In contrast, the rising trend of customized vehicle designs, such as those featuring flashy anime-themed graphics known as 'itasha', presents eye-catching yet semantically rich patterns. While these designs also attract attention, they are recognized as artistic expressions rather than deliberate attempts at concealment.

With advancements in DNN-based models, they are widely used in automated detection systems, reducing labor-intensive processes (Zhang et al. 2022). However, DNNs are vulnerable to adversarial examples (Szegedy et al. 2013), which can be manipulated to yield incorrect results. This has extended the concept of camouflage to the realm of attacking DNNs. Typically, adversarial perturbations are subtle, deceiving DNNs without being noticeable to humans. However, in real-world scenario, constraints such as varying perspectives and restricted attack areas pose additional challenges (Eykholt et al. 2018). Improving attack generalizability often requires relaxing restrictions (Huang et al. 2020; Wang et al. 2022), but results in conspicuous and exaggerated features, making them easily detectable by humans. Besides, existing methods designed to evade human perception are not explicitly optimized for attacking DNNs.

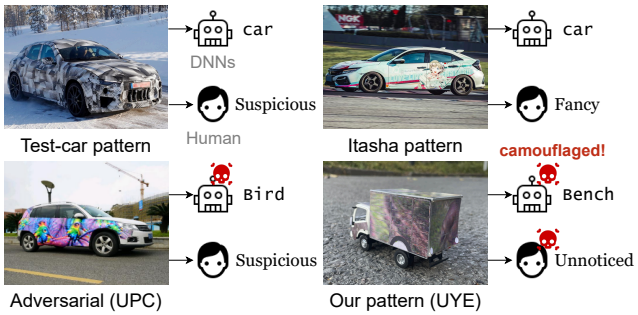

Figure 1: The different response of DNNs and humans to various camouflages. Our goal is to incorporate semantic information into the attacks, creating camouflages that are both effective against DNNs and appear natural to humans.

This gap motivates us to explore generating **camouflages that are both effective in attacking DNNs and perceptually natural to humans.** To achieve this, we propose UYE, a hybrid physical camouflage framework that optimizes for both attack effectiveness and human imperceptibility. The framework consists of three key components: (1) an attention refiner combining DINOv2 and multi-scale mechanisms to extract generalizable attention maps; (2) CAMO-Critics, a human-aligned perception evaluator trained using GPT-4o annotated data; and (3) a joint optimization process that generates patterns that blend naturally into background scenes while manipulating attention maps for adversarial attacks, with iterative refinement via the perception evaluator. Our approach overcomes previous limitations by balancing the two primary tasks, enabling the generation of unnoticed yet effective adversarial patterns for real-world deployment.

Our contribution can be summarized as follows:

- We propose a camouflage framework that simultaneously considers DNNs and human perception to derive effective patterns while avoiding conspicuous designs.
- An attention refiner is introduced to obtain precise and generalized attention, facilitating the implementation of effective attacks against various detection models.
- We consturct the first human preference dataset for camouflage task, constructed using a human-aligned LMM guided by designed prompts. A perception evaluator is trained on this dataset to establish objective metrics.
- Extensive experiments show that our method achieves the optimal balance between human perception and model attention, outperforming state-of-the-art methods.

Related Works

Physical Adversarial Attacks

Physical adversarial attacks modify the visual appearance of real-world objects by optimizing perturbations in simulated environments, as direct training in real-world scenarios presents engineering challenges. Existing methods typically rely on pre-specified detectors, such as R-CNN (Huang et al. 2020) or YOLOv3 (Wang et al. 2022), where the optimized pattern is then deployed to attack other detectors. To maintain effectiveness across diverse scenarios, these approaches must account for real-world conditions, often requires relaxing invisibility constraints, resulting in conspicuous patterns that are easily detectable by humans.

Some methods address this limitation by incorporating semantic content into the camouflage, reducing its detectability while making it less obvious to human observers. For instance, LAP (Tan et al. 2021) begins with cartoon images and refines them through a two-stage process that constrains edge and color features, whereas NAP (Hu et al. 2021) leverages a pre-trained GAN to generate naturalistic adversarial patches tailored to specific object categories. Although these methods enhance concealment from human perception, they often suffer from reduced effectiveness against detectors.

Human Vision Assessment

The naturalness of camouflage is typically evaluated through visual assessment, where volunteers judge effectiveness based on standardized metrics such as average detection time (Owens et al. 2014). However, these methods are inherently subjective, labor-intensive, and costly, leading to inconsistent evaluations.

Research on human visual attention mechanisms stems from cognitive psychology and neuroscience. In computer vision, this work has branched into two main approaches: Eye fixation prediction (EFP) estimates gaze distribution of observer within a scene, which relies on eye-tracking equipment along with extensive annotation (Borji and Itti 2015). Salient object detection (SOD) identifies the most visually prominent objects in an image, producing a saliency map where brighter regions indicate higher perceptual significance. Despite their utility, it remains uncertain whether these methods align well with human assessments in camouflage scenarios. Recent advances in large multimodal models (LMMs), such as GPT-4o (Hurst et al. 2024), have introduced more efficient evaluation by incorporating human-like judgment. These automated systems reduce costs while providing standardized, objective, and reproducible assessments (Peng et al. 2024).

Methods

Problem Definition

Camouflage in physical scenes requires transforming 3D objects and patterns into 2D images through a process known as rendering. Kato et al. (2018) introduced a neural renderer that enabling differentiability in this process. Given a 3D object parameterized by its mesh $\mathbf{H}$ and texture $\mathbf{T}$, its corresponding 2D image X^{obj} can be produced by a renderer $\mathcal{R}$:

$$X^{obj} = \mathcal{R}((\mathbf{H}, \mathbf{T}), \theta) \tag{1}$$

where $\theta \in \Theta$ signifies specific environmental conditions such as camera parameters and lighting conditions.

Due to the properties of the renderer, the background of the X^{obj} is entirely black, allowing for directly extraction of the ground-truth binary mask M, where foreground pixels are set to 1 and background pixels remain 0. The final rendered image X corresponding to $\mathbf{T}$ on a specific background B_θ can be obtained as follows:

$$X = X^{obj} \odot M + B_\theta \odot (\mathbf{1} - M) \tag{2}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{1}$ is an all-ones matrix matching the dimensions of M.

Let $F(\cdot)$ denote the target model, with $y = F(X)$ representing the model's output. The goal is to generate an adversarial pattern $\mathbf{T}'$ such that its corresponding X' can mislead the model to produce an erroneous output y', which can be defined as the following optimization problem:

$$\mathbf{T}' = \mathrm{argmax}(L(y, y')) \tag{3}$$

where $L(\cdot)$ is a loss function that quantifies the discrepancy between the original output and the adversarial output.

Overall Framework

The overall architecture of our proposed camouflage framework, UYE, is illustrated in Figure 2. The training procedure consists of two stages: first, a base pattern $\mathbf{T}_{base}$ is generated to match the environmental scene; second, an adversarial pattern $\mathbf{T}_{adv}$ is crafted to optimize attack against detector models while maintaining naturalness for human perception.

Following DAS (Wang et al. 2021), we use urban scenes from CARLA simulator (Dosovitskiy et al. 2017) to train and generate camouflage patterns.

Obtain Generalized Attention

Although detectors differ in architecture, they share a common goal of focusing attention on objects for accurate predictions. Our goal is to identify this mechanisms to obtain more generalized attention. This motivates our proposed attention refiner design, which contains vision encoder, external projector and multi-scale attention mechanism.

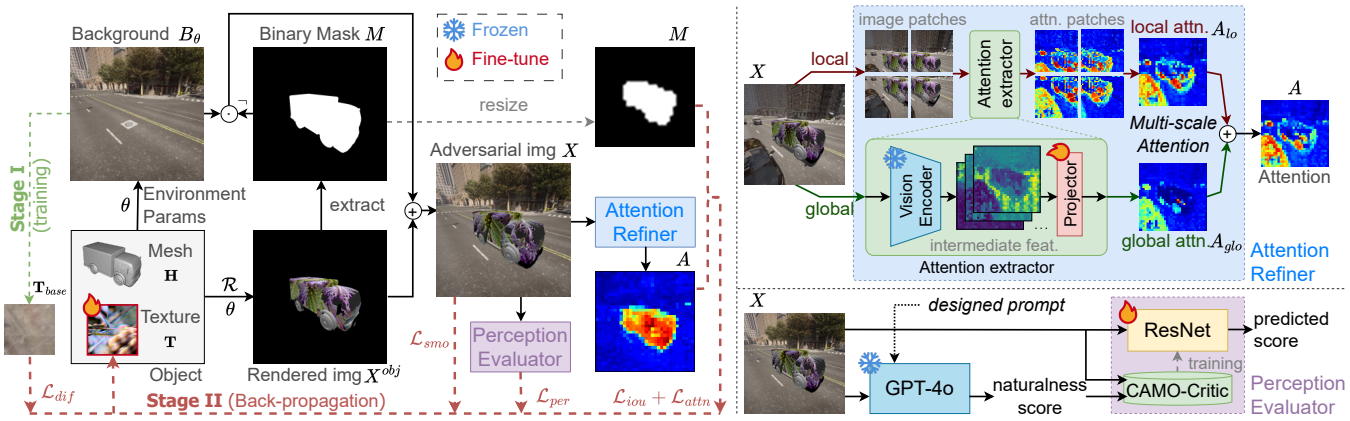

Figure 2: The architecture of our proposed camouflage framework UYE.

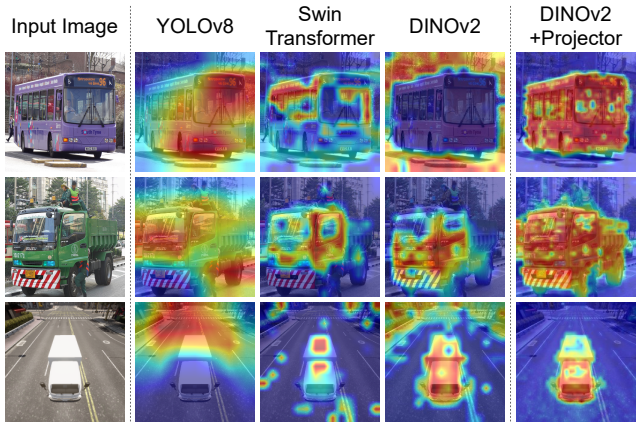

Figure 3: Attention maps extracted from different models using Grad-CAM.

Vision Encoder DINOv2 (Oquab et al. 2023) is a self-supervised vision encoder trained through knowledge distillation, enabling robust feature extraction across diverse image distributions without requiring fine-tuning. We compare it with YOLOv8 (Zhang et al. 2022) and Swin Transformer (Liu et al. 2021) on both real (COCO) and synthetic (CARLA) images. Using Grad-CAM (Selvaraju et al. 2017), we extract and visualize normalized attention maps. Figure 3 demonstrate that DINOv2 more accurately localizes foreground objects in most cases, underscoring its strength as a foundational model for generalizable feature learning. However, as shown in the first row, DINOv2's intermediate feature occasionally misalign objects and background, limiting their direct applicability.

External Projector To rectify foreground misalignment, we introduce an external projector that revises DINOv2's focus on foreground objects. The projector consists of a single-layer MLP, and is trained using an MSE loss on the COCO val2017 dataset (Lin et al. 2014), which provides images with binary masks. During training, DINOv2 remains fixed, while the projector attempts to focus attention on the masked regions. As shown in Figure 3, the enhanced attention maps more accurately identify foreground objects compared to intermediate features. Therefore, we employ DINOv2 with the fine-tuned projector as the attention extractor for obtaining generalized attentions.

Multi-scale Attention Mechanism To further leverage attention, we extracting and fusing it at different scales. The original image is first processed by the attention extractor to obtain global attention A_{glo}. The image is then divided into patches, each processed individually by attention extractor to acquire respective attention. These are then fused based on spatial positions to form local attention A_{lo}. A weighted fusion of A_{glo} and A_{lo} produces the final attention map A. In this study, the original image is divided into four patches with half-size overlap between adjacent patches. Overlapping regions are fused by averaging, while A_{glo} and A_{lo} are merged with a 1:2 ratio. For detailed implementation, please refer to the supplementary materials.

By combining the aforementioned modules, we developed the **attention refiner** to extract comprehensive and accurate attention maps for effective attacks against detectors.

Human Perception Assessment

The evaluation of camouflage naturalness by human observers is inherently subjective, as individual perceptions of the same pattern can vary significantly. To address this, it is crucial to develop objective evaluation metrics that reliably reflect human visual perception. Our approach involved following steps: First, recruit observers to assess the naturalness of various camouflage patterns, establishing statistical preference baselines across samples. Then, investigate the potential of employing GPT-4o to develop robust evaluation metrics alignment with human perceptual judgments.

Enhancing GPT-4o with Designed Prompting GPT-4o's inherent randomness and unclear understanding in camouflage task poses challenges for constructing reliable evaluation. To mitigate this, we designed tailored prompts to ensure consistent and high-quality responses. An example interaction with GPT-4o is shown in Figure 4.

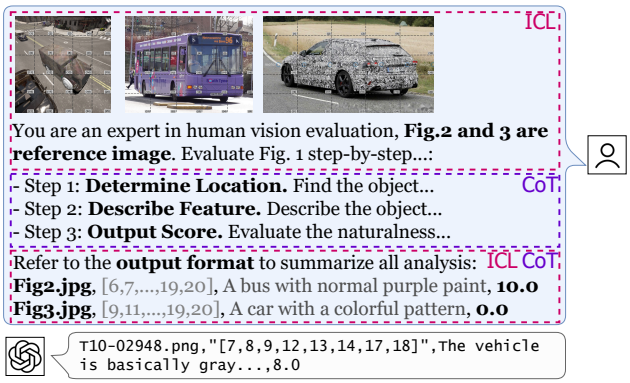

Figure 4: An example chat with GPT-4o using designed prompt.

Chain-of-thought (CoT) To ensure consistent responses, we guide GPT-4o through a step-by-step process: first identifying potential camouflaged objects in the given image, then describing the characteristics of the identified object, and finally assigning a naturalness score of the identified patterns. This structured approach allows GPT-4o to verify its own responses by ensuring accurate object identification.

In-context learning (ICL) To improve task understanding, we employ normal objects from the COCO dataset and test-car camouflages collected from the internet. Normal objects, perceived as highly natural by humans, are assigned higher naturalness scores, while camouflaged objects with complex patterns serve as contrasting examples. These reference samples provide clear scoring standards, enabling GPT-4o to deliver precise and consistent evaluations.

Human-LMM Perception Alignment To evaluate the alignment between GPT-4o and human perception, 100 environmental parameters were randomly selected to generate camouflage images via various attack methods, as shown in Figure 5. To mitigate subjective bias, 20 volunteers assessed the naturalness of these images via a questionnaire similar to the designed prompt, detailed in the supplementary material. The same set of images were presented to GPT-4o under four conditions: no special prompt, CoT only, ICL only, and a combination of CoT and ICL. Each method was tested three times on the same images, and the average scores were calculated. Table 1 summarizes the results, with rankings for each method shown in brackets. Pearson correlation coefficient (PCC) analysis reveals that the combined CoT and ICL approach achieves stronger alignment with user study, reflected in smaller score and ranking discrepancies. These findings underscore the effectiveness of the proposed prompt instructions in enhancing GPT-4o's evaluation of camouflage in a manner more consistent with human perception.

Perception Evaluator As a closed-source model, GPT-4o cannot be directly integrated into training pipelines. To overcome this limitation, we leverage the aforementioned prompting techniques to generate stable, human-aligned response dataset called CAMO-Critic. The dataset containing 5,000 image-score pairs, serves as the foundation for developing objective metrics to assess camouflage naturalness. Building upon CAMO-Critic, we develop a perception evaluator with ResNet-101 to learn the mapping between camouflage images and their naturalness scores. The model is trained using MSE loss between predicted and ground-truth scores. Once trained, the evaluator provides human-aligned naturalness assessment that can be seamlessly integrated into downstream camouflage optimization pipelines, effectively incorporating human perceptual factors while circumventing GPT-4o's accessibility limitations.

Method	Human	Plain	+CoT	+ICL	+CoT+ICL
BG	6.349 (#1)	6.495 (4)	6.224 (1)	6.285 (3)	6.755 (1)
LAP	5.375 (#2)	7.338 (1)	5.393 (3)	7.235 (1)	6.075 (2)
NAP	5.365 (#3)	6.389 (5)	5.226 (4)	6.088 (4)	5.310 (4)
TPA	4.416 (#4)	6.745 (3)	6.089 (2)	6.632 (2)	5.921 (3)
FCA	3.129 (#5)	5.098 (6)	3.506 (7)	4.687 (6)	4.588 (5)
TCEGA	2.260 (#6)	6.821 (2)	4.340 (6)	6.041 (5)	4.498 (6)
CAMOU	1.646 (#7)	4.677 (7)	4.926 (5)	4.188 (7)	3.593 (7)
PCC	–	0.6032	0.6652	0.7161	**0.9348**

Table 1: Naturalness scores of different camouflage methods obtained by humans and GPT-4o.

Figure 5: Camouflages generated by various attack methods.

Training Process

Stage I: Base Pattern

This stage focuses on generating camouflage $\mathbf{T}_{base}$ that effectively deceives the human vision within a specific scene. However, real world is dynamic and complex, and relying solely a specific background may leave the object to remain visible from other perspectives. This limitation can be addressed by deriving the camouflage from multiple images of the target environment. This approach maximizes the object's visual integration with its surroundings, ensuring a natural appearance from various viewpoints. Specifically, the base pattern $\mathbf{T}_{base}$ is obtained by minimizing the MSE loss between the rendered image X and the scene image B. The optimization process can be expressed as:

$$\mathcal{L}_{base} = \frac{1}{N} \sum_{i,j} \mathrm{MSE}(X_i, B_j) \tag{4}$$

where X_i stands for the i-th rendered image, B_j denotes the j-th scene image from the dataset, and N indicates the total number of combinations of X and B. The generation process of $\mathbf{T}_{base}$ is detailed in Algorithm 1. By leveraging the background-driven approach, this method enhances the camouflage's naturalness against human visual perception.

Stage II: Adversarial Pattern

This stage focuses on generating an adversarial pattern $\mathbf{T}_{adv}$ that effectively targets detection models while maintaining

Algorithm 1: Obtaining Base Pattern

Input: 3D model ($\mathbf{H}$, $\mathbf{T}$), neural renderer $\mathcal{R}$, background image B, environment params Θ
Output: Base pattern $\mathbf{T}_{base}$

$\mathbf{T}^{base} \leftarrow \mathbf{T}$;
for *the numbers of epoches* **do**
  for $i \leftarrow 1$ **to** n **do**
    randomly select θ_i from Θ, and obtain B_θ from B;
    $X_i \leftarrow \mathcal{R}((\mathbf{M}, \mathbf{T}_{base}), \theta_i)$;
    if $|X_i - B_\theta| < \epsilon$ **then** break;
    calculate $\mathcal{L}_{base}$ based on Eqn. (4);
    update $\mathbf{T}_{base}$ with gradient back-propagation;
  end
end
return $\mathbf{T}_{base}$

perceptual naturalness for humans. The attention refiner produces the attention A of adversarial image X^{adv}, while the perception evaluator provides the naturalness score of X^{adv}.

Attack Loss The ground-truth mask M distinguishes between background and foreground regions, enabling manipulating distributions of attention A based on this division. In attack task, the goal is to maximize attention deviation from the ground-truth, we use IOU measuring the overlap between predicted and ground-truth areas:

$$\mathcal{L}_{iou} = \frac{\sum_{i,j} A \odot M}{\sum_{i,j}(A + M - A \odot M)} \tag{5}$$

where $\Sigma_{i,j}$ denotes the summation of all pixel values.

Since the attention A can be segmented based on the binary mask M, the attention intensity for the object region ($\overline{A_f}$) and background region ($\overline{A_b}$) can be denoted as:

$$\overline{A_f} = \frac{\sum_{i,j}[A \odot M]}{\|M\|_0}, \ \overline{A_b} = \frac{\sum_{i,j}[A \odot (\mathbf{1} - M)]}{\|\mathbf{1} - M\|_0} \tag{6}$$

where $\|\cdot\|_0$ counts non-zero elements. We attempt to minimize attention on the object region and maximize attention on the background, which forms the attention loss $\mathcal{L}_{attn}$:

$$\mathcal{L}_{attn} = \overline{A_f} - \alpha \cdot \overline{A_b} \tag{7}$$

where α is a hyperparameter, empirically set to 0.1. Details on selection are provided in the supplementary material.

Vision Loss To prevent the adversarial pattern $\mathbf{T}_{adv}$ from deviating excessively from the base pattern $\mathbf{T}_{base}$, preserving its natural appearance, a difference loss $\mathcal{L}_{dif}$ is introduced, defined as the difference between the two patterns:

$$\mathcal{L}_{dif} = \|\mathbf{T}_{adv} - \mathbf{T}_{base}\|^2 \tag{8}$$

The potential of human-aligned LMMs in assessing camouflage naturalness inspires their integration to enhance the realism of adversarial patterns. We use the proposed perception evaluator to obtain the naturalness score of adversarial images and aim to enhance this score during training. The perception loss $\mathcal{L}_{per}$ is defined as:

$$\mathcal{L}_{per} = \overline{\mathrm{PE}(X)} - \mathrm{PE}(X^{adv}) \tag{9}$$

where $\mathrm{PE}(\cdot)$ represents the naturalness score of image, and $\overline{\mathrm{PE}(X)}$ denotes the average naturalness score across all samples. Additionally, we follow RP2 (Eykholt et al. 2018) to utilize the smooth loss $\mathcal{L}_{smo}$ reducing the inconsistent among adjacent pixels in adversarial image X^{adv}.

Optimization Process By combining the losses mentioned above, the optimization objective for final adversarial pattern is formulated as below:

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{attack} + \mathcal{L}_{vision} \\ &= \mathcal{L}_{iou} + \mathcal{L}_{attn} + \delta\mathcal{L}_{dif} + \lambda\mathcal{L}_{per} + \mu\mathcal{L}_{smo} \end{aligned} \tag{10}$$

The hyper-parameters δ, λ, μ control the balance in hybrid camouflage task. Following DAS, μ is empirically set to 10^{-4}. We set $\delta = 10^{-4}$ and $\lambda = 0.1$, with their specific values to be discussed in the ablation studies.

Experiments

Experimental Settings

Dataset We utilize dataset from DAS, captured from diverse viewpoints and distances from CARLA, which comprises 12,500 training images and 3,000 testing images. In Stage I, a subset of 300 images is used to train the base pattern. Stage II utilizes the entire training set.

Compared methods We evaluated several vision-based and attack-based adversarial camouflage methods. Vision-based methods include LAP (Tan et al. 2021), NAP (Hu et al. 2021), UPC (Huang et al. 2020), and DAC (Sun et al. 2023). For LAP and NAP, pre-generated patterns were applied to the objects, while patterns for UPC and DAC were generated using their official implementations. The attack-based methods consist of CAMOU (Zhang et al. 2019), DAS (Wang et al. 2021), FCA (Wang et al. 2022), and TPA (Zhang et al. 2023). Official implementations are employed for DAS, FCA, and TPA, while CAMOU is reimplemented due to code unavailability. A "Background" baseline, corresponding to the $\mathbf{T}_{base}$ in Stage I is also included.

Implementation details The ViT-B/14 variant of DINOv2 is utilized, with a feature size of 32×32. Optimization is performed using Adam with a learning rate of 0.01, a weight decay of 10^{-4}, and a maximum of 10 epochs. All methods utilize the same coverage area, encompassing all visible surfaces of the object. All codes are implemented in PyTorch, with training and evaluation conducted on an NVIDIA GeForce RTX 3090 (24GB).

Attack on Object Detectors

We compare different methods across various object detectors, including Faster R-CNN (FR-CNN) (Ren et al. 2015), YOLOv8 (Zhang et al. 2022), Co-DETR (Zong, Song, and Liu 2023), and GroundingDINO (GDINO) (Liu et al. 2023a). All detectors were pre-trained on the COCO dataset, while for GroundingDINO, COCO category labels are used as text prompts. We adopt AP@0.5 as the metric, where a lower value indicates higher attack effects.

	Methods	Object Detection				Human Perception				Comp.
		FR-CNN	YOLOv8	Co-DETR	GDINO	GPT-4o	EVPv2	DeepGaze	User Study	
		AP@0.5 ↓				score ↑	F_β ↓	$\overline{A_f}$ ↓	score ↑	↑
	Raw	0.479	0.556	0.596	0.606	5.234	0.580	0.244	–	0.404
	Background	0.247	0.452	0.540	0.341	5.598	0.544	0.272	–	0.656
attack-based	CAMOU (2019)	0.310	0.298	0.448	0.225	4.229	0.617	0.382	2.416	0.427
attack-based	DAS (2021)	**0.120**	0.325	0.310	0.146	4.017	0.682	0.396	2.550	0.447
attack-based	FCA (2022)	0.151	**0.212**	0.278	0.202	3.994	0.639	0.393	3.733	0.526
attack-based	TPA (2023)	0.182	0.291	0.478	0.232	4.587	0.598	0.375	4.800	0.523
attack-based	UYE (Ours)	0.137	0.218	**0.277**	**0.120**	**5.855**	**0.571**	**0.360**	**6.016**	**0.833**
vision-based	LAP (2021)	0.425	0.427	0.601	0.557	4.793	0.633	0.345	3.650	0.268
vision-based	NAP (2021)	0.382	0.313	0.543	0.495	4.976	0.687	0.349	3.500	0.312
vision-based	UPC (2020)	0.370	0.382	0.479	0.356	4.950	0.574	**0.336**	5.833	0.495
vision-based	DAC (2023)	0.235	0.280	0.490	0.252	5.104	0.593	0.340	**6.133**	0.589
vision-based	UYE (Ours)	**0.137**	**0.218**	**0.277**	**0.120**	**5.855**	**0.571**	0.360	6.016	**0.833**

Table 2: Comparison results between object detectors and human perception for various attack methods.

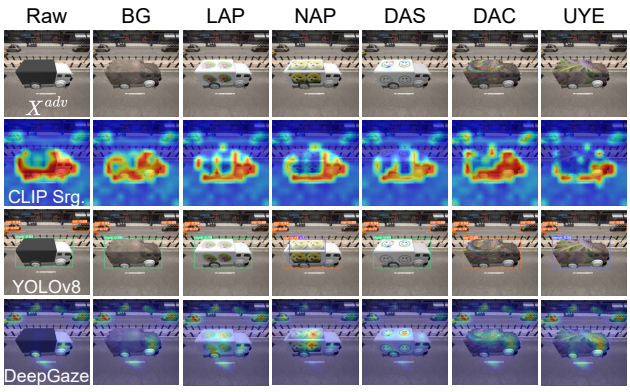

Figure 6: Visualization across various models.

The comparative results are summarized in Table 2, where bold and underlined values indicate the best and second-best attack performance, respectively. The results demonstrate that UYE achieves either the best or second-best performance among attack-based methods, and outperforms all vision-based methods across all detectors. Moreover, adversarial images generated by different methods were evaluated using CLIP Surgery (Li et al. 2023), YOLOv8, and DeepGaze to obtain attention maps, as shown in Figure 6. The results reveal that within regions covered by the camouflage pattern, UYE achieves the lowest attention on the object compared to other methods, demonstrating a superior ability to divert attention away from the object.

Deceive on Human Perception

We compare different approaches across various tasks related to human perception. For LMMs, naturalness score are evaluated using aligned GPT-4o. In the EFP task, DeepGaze (Linardos et al. 2021) is used to calculate the attention intensity within the object region $\overline{A_f}$ as defined in Eqn. (6), where lower values indicate better naturalness. For the SOD task, EVPv2 (Liu et al. 2023b) is employed and measure performance with F-measure (F_β), where lower scores indicate less salient foregrounds. Both EFP and SOD metrics rely on binary masks for computation.

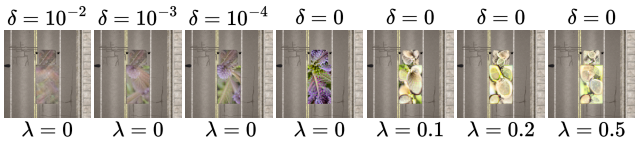

Figure 7: Visualization across different hyper-parameter.

The comparative results are summarized in Table 2, where bold and underlined values indicate the best and second-best camouflage naturalness, respectively. The results show that UYE which leverages environment-related patterns, achieves the highest naturalness across all metrics among the attack-based methods and performs relatively well in vision-based methods. To further validate the results, a user study was conducted by volunteers using the same configuration mentioned before. The findings confirm that UYE earning favorable human preferences across both groups of methods.

Comprehensive Evaluation

For a fair evaluation of the hybrid task, we normalized the metrics for each subtask using min-max scaling and calculated a weighted average with equal weights (50% each) assigned to object detection and human perception tasks. The results, shown in Table 2, demonstrate that UYE outperforms all compared methods in comprehensive evaluation. Owing to its optimized design, UYE offers superior invisibility to humans compared to DAC under identical environmental conditions. UYE achieves a balanced performance between deceiving object detectors and maintaining naturalness in human perception, highlighting its effectiveness.

Impact of Hyper-parameter

The impact of δ and λ on balancing the hybrid task is evaluated, where δ controls the degree of environmental preservation and λ adjusts the weight of the naturalness score provided by GPT-4o. Quantitative results are presented in Table 3, while Figure 7 illustrates the differences in camouflage patterns generated under varying δ and λ.

Config		Object Detection ↓			Human Perception	
δ	λ	YOLO	Co-DETR	GDINO	GPT-4o ↑	DeepGaze ↓
0	0	0.124	0.124	0.040	3.784	0.395
0	0.1	0.280	0.171	0.144	6.317	0.393
0	0.2	0.380	0.246	0.236	6.449	0.387
10^{-4}	0	0.317	0.385	0.141	4.629	0.365
10^{-4}	0.1	0.218	0.277	0.120	5.855	0.360
10^{-3}	0.1	0.406	0.469	0.224	5.566	0.326
10^{-2}	0.1	0.439	0.539	0.323	5.573	0.284

Table 3: Ablation study on hyper-parameter δ and λ.

Proj.	A_{glo}	A_{lo}	FR-CNN	YOLOv8	Co-DETR	GDINO
–	✓	–	0.138	0.247	0.237	0.168
✓	✓	–	0.069	0.159	0.132	0.049
✓	–	no lap.	0.104	0.270	0.176	0.059
✓	–	1/2 lap.	0.058	0.142	0.124	0.052
✓	2 : 1		0.066	0.147	0.127	**0.040**
✓	1 : 1		0.056	0.138	**0.115**	**0.040**
✓	1 : 2		**0.044**	**0.124**	0.124	**0.040**

Table 4: Ablation study of components in attention refiner.

Smaller δ values relax the constraints on adversarial perturbations, enhancing attack success rates against object detectors at the expense of perceptual naturalness, which results in more visually detectable patterns. While bigger δ keeps environmental features. Similarly, adjustments to λ have effects on adversarial perturbations comparable to those of δ. Considering the joint impact of δ and λ, we find that for a fixed δ, appropriately increasing λ improves both the naturalness of the camouflage pattern and their effectiveness against object detectors. Conversely, for a fixed λ, increasing δ degrades performance on both tasks. Based on the quantitative result and aforementioned analysis, we selected $\delta = 10^{-4}$ and $\lambda = 0.1$ as the optimal balance for UYE.

Explore Attention Refiner

The attention refiner is the critical modules of UYE. First, we evaluated the effectiveness of the projector. In its absence, features extracted by DINOv2 were directly averaged along the channel dimension. As shown in Table 4, introducing the projector significantly improves attack performance, highlighting its role in refining attention. Second, we assessed the multi-scale attention mechanism with different overlap strategies (without or with 1/2 size). We also considered using global attention A_{glo} and local attention A_{lo} individually, as well as their combinations with varying fusion ratios. The results indicate that integrating multi-scale attention achieves better attack performance than relying on global or local features alone. Notably, a fusion ratio of 1:2 between A_{glo} and A_{lo} yields the best performance.

Attack in Real-world Scenario

The real-world performance was assessed through a simplified yet representative setup. Due to engineering and cost constraints, a toy truck model was chosen as the target object for testing in real-world scenarios. Camouflage patterns were printed on paper and affixed to the truck's surfaces. Asphalt pavement is selected as the typical real-world scene. Images were captured by an iPhone 12 from multiple viewpoints around the target. We compared our method with DAS, FCA, and TPA, processing the captured images using CLIP Surgery and attention refiner to generate attention maps, as shown in Figure 8. Table 5 presents the result of various camouflage patterns against YOLOv8, GroundingDINO and GPT-4o. The results demonstrate that, our method effectively diverts the models' attention, highlighting the robustness and generality of our approach. For a comprehensive analysis of real-world attacks, including a wider variety of scenes and quantitative evaluations of attack effectiveness, please refer to the supplementary material.

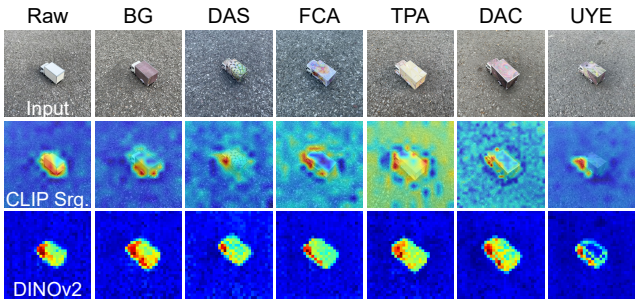

Figure 8: Attentions of different pattern in real world.

Methods	YOLOv8↓	GDINO↓	GPT-4o↑	Comp.↑
Raw	0.739	0.922	4.893	0.461
Background	0.898	0.864	4.789	0.270
DAS (2021)	**0.715**	**0.619**	3.762	0.500
FCA (2022)	0.775	0.731	3.954	0.367
TPA (2023)	0.741	0.877	5.194	0.561
DAC (2023)	0.776	0.844	5.576	0.623
UYE (Ours)	0.838	0.802	**6.073**	**0.680**

Table 5: The attack effect of camouflage pattern on object detection models in real-world scenarios.

Conclusion

We propose UYE, a novel camouflage framework that generates patterns capable of simultaneously deceiving both detection models and human perception. The framework consists of two sequential stages. In the first stage, a base pattern is generated by adapting to the surrounding environment, enhancing its naturalness to better blend with human perception. In the second stage, an attention refiner built upon DINOv2 extracts and refines generalized attention maps, manipulating them to effectively attack detectors. A perception evaluator, leveraging human-aligned GPT-4o annotations from CAMO-Critic dataset, provides naturalness scores for the adversarial images, and optimization is performed to reduce the visual saliency to human observers. Extensive experiments show that UYE achieves superior performance in both misleading detection models and deceiving human perception, outperforming state-of-the-art approaches, and are also effective in real-world scenarios.

Ethics Statement

While our proposed camouflage framework demonstrates promising capabilities for enhancing privacy and security applications, we acknowledge the potential risks of its misuse for malicious purposes, such as evading surveillance systems or facilitating unauthorized activities. To mitigate these concerns, we advocate for the development of robust countermeasures, including adversarial training and anomaly detection mechanisms, to improve the resilience of DNN-based systems against such attacks. Furthermore, we emphasize the importance of adhering to ethical research guidelines and promoting transparency in the publication of adversarial techniques to foster responsible innovation and collaborative efforts toward safeguarding AI systems.

References

Borji, A.; and Itti, L. 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on Robot Learning*, 1–16. PMLR.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1625–1634.

Hu, Y.-C.-T.; Kung, B.-H.; Tan, D. S.; Chen, J.-C.; Hua, K.-L.; and Cheng, W.-H. 2021. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7848–7857.

Huang, L.; Gao, C.; Zhou, Y.; Xie, C.; Yuille, A. L.; Zou, C.; and Liu, N. 2020. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 720–729.

Hurst, A.; Lerer, A.; Goucher, A. P.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kato, H.; Ushiku, Y.; and Harada, T. 2018. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3907–3916.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, 740–755. Springer.

Linardos, A.; Kümmerer, M.; Press, O.; and Bethge, M. 2021. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12919–12928.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, W.; Shen, X.; Pun, C.-M.; and Cun, X. 2023b. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19434–19445.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Owens, A.; Barnes, C.; Flint, A.; Singh, H.; and Freeman, W. 2014. Camouflaging an object from many viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2782–2789.

Peng, Y.; Cui, Y.; Tang, H.; Qi, Z.; Dong, R.; Bai, J.; Han, C.; Ge, Z.; Zhang, X.; and Xia, S.-T. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sun, J.; Yao, W.; Jiang, T.; Wang, D.; and Chen, X. 2023. Differential evolution based dual adversarial camouflage: Fooling human eyes and object detectors. *Neural Networks*, 163: 256–271.

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

Tan, J.; Ji, N.; Xie, H.; and Xiang, X. 2021. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *Proceedings of the 29th ACM international conference on multimedia*, 5307–5315.

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

Wang, J.; Liu, A.; Yin, Z.; Liu, S.; Tang, S.; and 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*, 8565–8574.

Zhang, W.; Zhou, Q.; Li, R.; and Niu, F. 2022. Research on camouflage target detection method based on improved YOLOv5. In *Journal of Physics: Conference Series*, volume 2284, 012018. IOP Publishing.

Zhang, Y.; Foroosh, H.; David, P.; and Gong, B. 2019. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*.

Zhang, Y.; Gong, Z.; Zhang, Y.; Bin, K.; Li, Y.; Qi, J.; Wen, H.; and Zhong, P. 2023. Boosting transferability of physical attack against detectors by redistributing separable attention. *Pattern Recognition*, 138: 109435.

Zong, Z.; Song, G.; and Liu, Y. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6748–6758.