Diff-NAT: Better Naturalistic and Aggressive Adversarial Attacks via Class-Optimized Diffusion for Object Detection

Qinglong Yan*, Tong Zou*, Xunpeng Yi, Xinyu Xiang, Xuying Wu, Hao Zhang, Jiayi Ma†

Electronic Information School, Wuhan University, Wuhan 430072, China
{qinglong_yan, zoutong, yixunpeng, xiangxinyu, wuxuying}@whu.edu.cn, {zhpersonalbox, jyma2010}@gmail.com

Abstract

Recent advances in naturalistic physical adversarial patch generation show great promise in protecting personal privacy against detector-based malicious surveillance while remaining inconspicuous to human observers. In this work, we present the first systematic categorization and in-depth re-examination of existing methods into three representative paradigms, revealing a pervasive imbalance: enforcing naturalness constraints inherently restricts the adversarial search space, thus limiting attack performance. To address this challenge, we propose a novel paradigm based on class-optimized diffusion, termed Diff-NAT. Diff-NAT leverages pretrained diffusion models as powerful natural image priors and introduces a unified iterative framework that jointly optimizes two complementary components: semantic-level textual prompts and instance-level latent codes. Specifically, prompt optimization enables broad traversal across inter-class semantic regions, while latent refinement allows for fine-grained manipulation within class objectives. This dual-level optimization facilitates progressive navigation toward adversarial distributions embedded within the natural semantic manifold. Extensive experiments in both digital and physical settings demonstrate that Diff-NAT outperforms existing SOTA approaches in terms of both visual realism and aggressiveness.

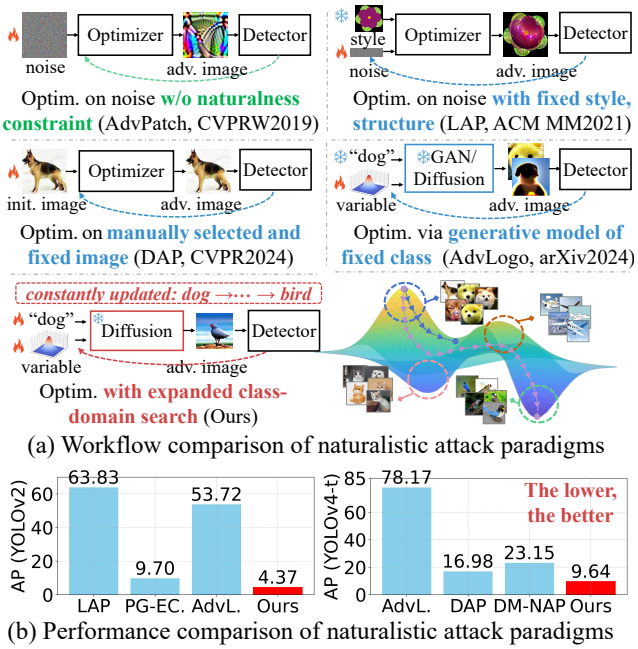
Code — https://github.com/QinglongYan-hub/Diff-NAT

Introduction

With the development of deep neural networks (DNNs), intelligent vision systems such as pedestrian detection and person re-identification are increasingly adopted in public spaces (Ren et al. 2016; Yi et al. 2025a). While enhancing efficiency, they also raise privacy concerns, as sensitive information may be extracted without individual consent. To address this issue, adversarial examples generated through adversarial attack techniques can mislead model predictions and serve as a mechanism for privacy protection (Hu et al. 2021). This highlights the need for a better understanding of adversarial example generation in privacy-critical contexts.

Adversarial attacks are commonly divided into digital and physical categories. Digital attacks add imperceptible perturbations to input images but are difficult to deploy in real-world settings due to their lack of physical visibility (Madry et al. 2017; Qiu et al. 2020) . In contrast, physical attacks use printable patches that can be attached to objects, making adversarial patches a practical and deployable solution for real-world privacy protection (Wei et al. 2024a,b).

In this work, we focus on the physical adversarial attacks due to its real-world applicability. As illustrated in Fig. 1, AdvPatch (Thys, Van Ranst, and Goedemé 2019) represents a pioneering approach that generates adversarial patches by directly optimizing random noise patterns. However, lack-



(a) Workflow comparison of naturalistic attack paradigms

(b) Performance comparison of naturalistic attack paradigms

Figure 1: (a) Workflow comparison of adversarial patch paradigms, including unnatural paradigm, three naturalness-driven but weakly aggressive paradigms, and our dual-preserving paradigm ensuring both naturalness and aggressiveness. (b) Performance comparison: Our method achieves the SOTA attack by leveraging semantic-instance adaptive search over a broader natural manifold, reducing YOLOv2's detection AP from 9.70 (best among prior methods) to 4.37.

*These authors contributed equally.
†Corresponding Author.

ing explicit naturalness constraints, these patterns often exhibit conspicuous, high-contrast textures that are easily perceived by human observers. This compromises stealthiness and arouses suspicion, thereby limiting their usability in public spaces. To address this issue, recent efforts have explored enhancing visual naturalness in physical adversarial patch generation by integrating naturalness constraints into optimization. Through a comprehensive analysis, we present the first systematic categorization that organizes existing methods into three distinct paradigms. At the same time, we find a shared limitation: while improving visual realism, they often suffer from a significant drop in adversarial strength, making it difficult to achieve both naturalness and aggressiveness simultaneously.

Specifically, one representative paradigm (**paradigm I**) imposes fixed visual priors such as cartoon or artistic styles by directly optimizing random noise within predefined aesthetic domains (Tan et al. 2021). While this promotes visual plausibility, the rigid style constraint tends to suppress semantic diversity and limits adaptability. An alternative strategy (**paradigm II**) enhances naturalness by anchoring the optimization process to manually selected natural images (Guesmi et al. 2024), which serve as visual references for perturbation transfer. Although this improves realism through similarity constraints, reliance on human-curated exemplars inherently narrows the scope of semantic exploration and limits generalization capacity. More recent developments (**paradigm III**) harness the generative capabilities of pretrained GANs (Hu et al. 2021) and diffusion models (Miao et al. 2024) by optimizing latent variables under fixed textual prompts (*e.g.*, "a dog"). Compared to the above paradigms, this marks a progression from single style or image toward distribution-level exploration, yet still restricts adversarial search to a single semantic category.

Taken together, although these paradigms take meaningful steps toward bridging adversarial effectiveness and visual realism, they all rely on handcrafted or static choices in style, content, or semantic. Thus, such constraints significantly narrow the adversarial search space and confine patch generation to a limited region of the natural visual manifold, ultimately leading to suboptimal attack performance.

To this end, we propose Diff-NAT, a novel paradigm via class-optimized diffusion. As shown in Fig. 1, the core insight lies in performing adaptive optimization in semantic-level and instance-level through a unified iterative framework. Specifically, Diff-NAT leverages pretrained diffusion models as strong natural image priors, providing access to rich semantic manifolds that support naturalistic generation. On this foundation, it conducts dual-level optimization to progressively approach the adversarial distribution under the guidance of attack objectives. At the semantic level, class-related textual prompts are optimized to enable flexible traversal across diverse semantic categories, effectively breaking the constraints of fixed-class conditioning and exposing more aggressive regions. At the instance level, the latent representations are further refined to allow fine-grained intra-class manipulation, thereby enhancing attack strength. Finally, we integrate the above semantic-level and instance-level optimization in an iterative manner, which substantially enlarges the adversarial search space. This enables adaptive adversarial search within the wide semantic manifold, thereby facilitating the simultaneous realization of visual naturalness and adversarial effectiveness.

Overall, our contributions can be summarized as follows:

- We present the first comprehensive taxonomy of naturalistic adversarial patch generation methods, identifying three representative paradigms and exposing a prevalent imbalance between naturalness and aggressiveness.
- We propose a new paradigm via class-optimized diffusion, breaking the limitations of fixed semantic conditioning and facilitating progressive movement toward adversarial distributions within broad natural manifold.
- An iterative optimization strategy guided by adversarial objectives is applied over semantic-level prompt and instance-level latent code, facilitating wide-range inter-class and fine-grained intra-class adversarial exploration.
- Extensive digital and physical experiments validate that our method consistently achieves superior attack performance while maintaining high visual naturalness.

Related Work

Adversarial Attacks

Adversarial attacks pose a critical threat to DNNs by introducing carefully crafted perturbations that mislead model predictions (Hu et al. 2022; Wei et al. 2023; Chen et al. 2023; Wei, Yu, and Huang 2024). (Szegedy et al. 2014) first revealed the vulnerability of DNNs to such attacks, which sparked extensive research across various architectures and deployment scenarios. Physical-domain adversarial attacks extend this threat into the real world by embedding perturbations into tangible carriers like printable patches, wearables, or 3D objects (Xu et al. 2020; Hu et al. 2023; Cheng et al. 2024; Zhu et al. 2024). Studies have investigated various physical forms, including planar patches (Thys, Van Ranst, and Goedemé 2019), adversarial accessories (Komkov and Petiushko 2021), and 3D objects (Chen et al. 2018). Early pixel-wise optimization methods (Brown et al. 2017; Huang et al. 2023) often produced conspicuous patterns perceptible to humans, motivating recent efforts toward improving naturalness and stealth.

Generative Diffusion Models

Diffusion models originated from non-equilibrium thermodynamics , where (Sohl-Dickstein et al. 2015) first formulated the idea of iterative noising and denoising for data generation. This framework was later popularized by (Ho, Jain, and Abbeel 2020) through Denoising Diffusion Probabilistic Models (DDPM), which established a practical training objective by predicting noise with U-Nets, facilitating high-quality image generation. The remarkable generative priors of diffusion models have enabled their successful deployment across multiple domains, revolutionizing traditional approaches to image restoration (Lugmayr et al. 2022; Xia et al. 2023), quality enhancement (Saharia et al. 2022; Yi et al. 2025b), image fusion (Yi et al. 2024; Yue et al. 2023), and content-aware image modification (Meng et al. 2021).

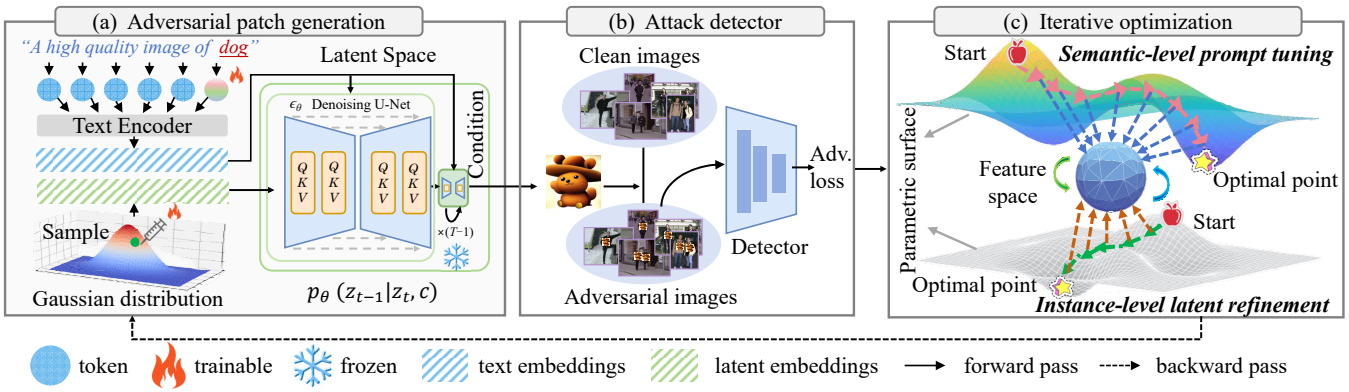


Figure 2: The overall framework of our proposed Diff-NAT via class-optimized diffusion.

Methodology

Problem Definition

For an object detection model $f_\theta(\cdot)$ with parameters θ, we denote the clean image as I and the corresponding ground truth label as $y = f_\theta(I)$. Our goal is to generate an adversarial patch δ that can mislead the correct prediction of detector $f_\theta(\cdot)$, while simultaneously preserving visual naturalness. The optimization problem is formulated as:

$$\delta^* = \arg\max_{\delta} \mathcal{L}(f_\theta(T(I,\delta)), y), \quad s.t.\ \delta \text{ is natural}, \tag{1}$$

where $\mathcal{L}(\cdot)$ is the adversarial loss, $I_{adv} = T(I,\delta)$ is the adversarial image formed by applying patch δ to the clean image I, and $T(\cdot)$ represents the applied function.

Specifically, our overall framework is shown in Fig. 2, which builds on class-optimized diffusion and performs adaptive optimization across inter-class semantics and intra-class instances within a unified iterative process, enhancing both visual naturalness and adversarial effectiveness.

Generative Diffusion Preliminary

To achieve adversarial patches with high visual fidelity, we adopt diffusion models as generative priors that capture the structure of the natural image manifold. As illustrated in Fig. 3, these models consist of a forward process that incrementally corrupts a clean image $x_0 \sim p(x_0)$ into a noise signal $x_T \sim \mathcal{N}(0, \mathbf{I})$, and a reverse process that reconstructs x_0 from x_T, effectively learning to synthesize realistic content by inverting a stochastic degradation. Therefore, such generative capacity aligns well with our objective of crafting naturalistic patches.

Specifically, a representative formulation is the Denoising Diffusion Probabilistic Model (DDPM), which defines a fixed Markov forward process that gradually perturbs a clean image x_0 into a noise signal x_t via Gaussian transitions:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t\mathbf{I}), \tag{2}$$

where β_t is a fixed variance schedule and $\alpha_t = 1 - \beta_t$.

Then, the reverse process attempts to reconstruct original sample by iteratively denoising the terminal noise x_T using the learnable denoising network ϵ_θ, which is modeled as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}), \tag{3}$$

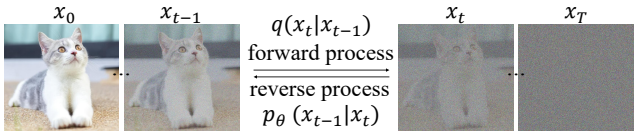


Figure 3: Forward and reverse processes of Diffusion model.

where the predicted mean is parameterized by:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right). \tag{4}$$

Therefore, the training objective is to minimize the deviation between the truth noise and the predicted estimate:

$$\mathbb{E}_{x_0,n_t,t}\left[\left\|n_t - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}n_t, t\right)\right\|^2\right], \tag{5}$$

where n_t is sampled from a standard Gaussian noise.

Although DDPM offers a principled foundation for natural image modeling, its pixel-space operations and unconditional generation restrict semantic control and adversarial exploration. To address this, we further build upon the advanced Stable Diffusion model (Rombach et al. 2022). It operates in a compact latent space $z_0 = \mathcal{E}(x_0)$, learned via the encoder $\mathcal{E}$ of a variational autoencoder (VAE), providing a low-dimensional yet semantically meaningful representation. The forward process perturbs the latent as $q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, \bar{\beta}_t\mathbf{I})$. To enable controllable generation, the model conditions the denoising network on a text prompt c, serving as a high-level semantic prior to guide the generative trajectory. The reverse process estimates z_0 from an Gaussian sample z_T via conditional ϵ_θ, yielding the following transition distribution:

$$p_\theta(z_{t-1}|z_t, c) = \mathcal{N}\left(z_{t-1}; \mu_\theta(z_t, t, c), \sigma_t^2\mathbf{I}\right). \tag{6}$$

After the denoising trajectory reaches z_0, the final image is reconstructed via the VAE decoder as $x_0 = \mathcal{D}(z_0)$.

Naturalistic Adversarial Patch Generation

As illustrated in Fig. 2(a), we begin by initializing the text prompt P as "A high quality image of a *dog*", and make the class "dog" learnable. For the given text prompt P, it

is first tokenized into a sequence of discrete tokens, which are subsequently mapped to their corresponding embedding vectors via a token embedding layer, yielding the embedding sequence:

$$[e^0_{\text{token}}, e^1_{\text{token}}, \ldots, \hat{e}^{K-1}_{\text{token}}] = \text{Embed}(\text{Tokenizer}(P)), \quad (7)$$

where $\hat{e}^{K-1}_{\text{token}}$ is learnable token embedding. This sequence is subsequently fed into a Transformer-based text encoder and obtain the final text embedding c_{text}, modeled as:

$$c_{\text{text}} = \text{CLIP}_{\text{text}}(e^0_{\text{token}}, e^1_{\text{token}}, \ldots, \hat{e}^{K-1}_{\text{token}}), \quad (8)$$

where the encoder $\text{CLIP}_{\text{text}}$, derived from CLIP (Radford et al. 2021), compresses the prompt into a compact representation capturing core semantics. With the conditioning text embedding c_{text}, we guide the reverse diffusion process to synthesize adversarial patches. We sample a random latent variable $z_T \sim \mathcal{N}(0, \mathbf{I})$, and reconstruct clean z_0 based on Eq. 6. This conditional denoising steps are defined as:

$$z_{t-1} = \text{DDIM}(z_t, \epsilon_\theta(z_t, t, c_{\text{text}})), \quad t = T, \ldots, 1, \quad (9)$$

where DDIM (Song, Meng, and Ermon 2020) serves as a deterministic sampling scheme for accelerating the reverse diffusion. Finally, the natural adversarial patch δ is obtained by decoding the clean latent representation z_0 as $\delta = \mathcal{D}(z_0)$.

Adversarial Attack on Object Detector

As shown in Fig. 2(b), we past the generated adversarial patch δ onto clean image $I \in \mathbb{R}^{H\times W\times 3}$ with a binary mask matrix $M \in \{0,1\}^{H\times W\times 1}$, which is formulated as:

$$I_{\text{adv}} = I \odot (1 - M) + \delta \odot M, \quad (10)$$

where I_{adv} is the adversarial image, and $\odot$ represents the Hadamard product. The mask M is derived from the GT bounding box, with patch size scaled by a ratio.

Then, the adversarial image I_{adv} is fed into object detector f_θ to predict the bounding box position V_{pos}, objectness score V_{obj}, and class probability V_{cls}. To achieve attack of batch N, we jointly minimize the maximum V^j_{obj} and V^j_{cls} for the "person" class in each image, modeled as:

$$\mathcal{L}_{\text{adv}} = \frac{1}{N}\sum_{i=1}^{N} \max_j \left[V^j_{\text{obj}}(I_{\text{adv}}) V^j_{\text{cls}}(I_{\text{adv}}) \right]. \quad (11)$$

Semantic-Instance Iterative Optimization

As shown in Fig. 4, the low-dimensional latent space of Stable Diffusion defines a broad natural manifold. However, the semantic distribution induced by the initial prompt P, centered on the "dog" class, occupies only a restricted region of this manifold. Although such manually specified distribution preserves naturalness, it is typically distant from adversarial distribution in this natural manifold. Therefore, we build on class-optimized diffusion process as Fig. 2(c), and perform adaptive adversarial optimization across and within categories under the constraint of adversarial loss $\mathcal{L}_{\text{adv}}$.

At the semantic level, we optimize class-specific component of prompt P, guiding progression toward semantic with stronger adversarial potential. Specifically, the class embedding $\hat{e}^{K-1\ (i)}_{\text{token}}$ is updated under adversarial supervision:

$$\hat{e}^{K-1\ (i+1)}_{\text{token}} \leftarrow \hat{e}^{K-1\ (i)}_{\text{token}} - \eta_1 \frac{\partial \mathcal{L}_{\text{adv}}}{\partial \hat{e}^{K-1\ (i)}_{\text{token}}}. \quad (12)$$

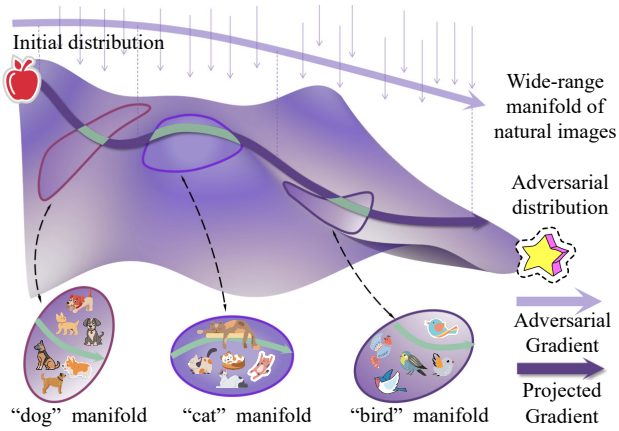


Figure 4: Adv. distribution hidden on the natural manifold.

Within the semantically optimized class, we further refine the latent code z_T at the instance level to enhance adversarial strength. We adopt the MI-FGSM (Dong et al. 2018), which incorporates historical momentum g_i and applies a sign operation on gradient to effectively maximize the adversarial objective by reinforcing the attack direction, formulated as:

$$g_{i+1} \leftarrow \mu \cdot g_i + \frac{\partial \mathcal{L}_{\text{adv}}/\partial z^i_T}{\|\partial \mathcal{L}_{\text{adv}}/\partial z^i_T\|}, \quad (13)$$

$$z^{i+1}_T \leftarrow \kappa\left(z^i_T - \eta_2 \cdot \text{sign}(g_{i+1})\right), \quad (14)$$

where $\kappa(z^{i+1}_T)$ prevents the optimized latent z^{i+1}_T from deviating excessively from the standard Gaussian distribution, thus preserving visual realism, which is defied as:

$$\kappa(z^{i+1}_T) = \min\left(\max(z^{i+1}_T, -\tau), \tau\right). \quad (15)$$

Finally, for these two levels optimization of text prompt P and latent code z_T, we integrate them in an alternating iterative manner as shown in Fig. 2(c). This alternating iterative explores adversarial search space by dynamically optimizing both inter-class semantics and intra-class instances. In other words, it allows us to navigate within the natural manifold under the guidance of adversarial objectives to uncover hidden adversarial distributions, thereby simultaneously enhancing visual naturalness and attack effectiveness.

Experiments

Experimental Settings

Datasets and Evaluation Metrics. We use INRIA Person dataset (Dalal and Triggs 2005) for training (614 images) and testing (288). For generalization, we additionally evaluate on COCO-Person (2,693), CCTV-Person (559), and 200 self-collected images. Attack strength is measured by Average Precision (AP), with lower AP indicating better attack.

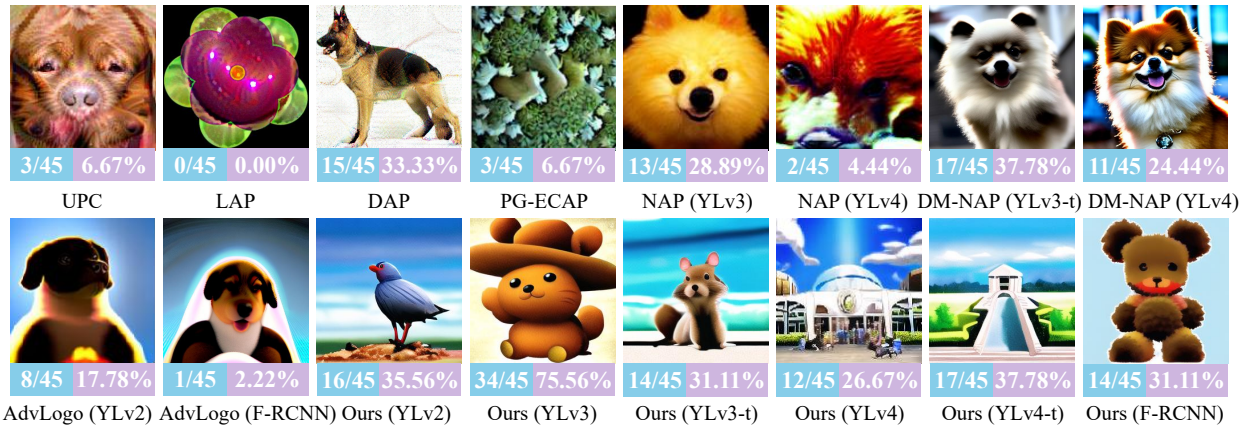


Figure 5: Visualization of naturalness patches generated by different adversarial attack methods.

Figure 6: Semantic-instance level optimization.

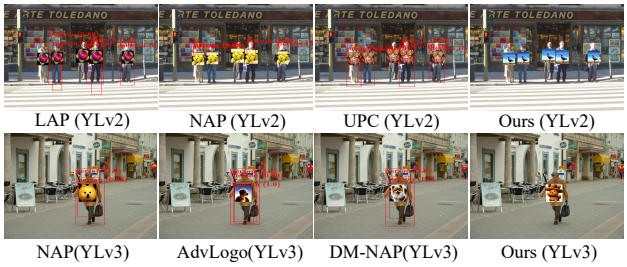


Figure 7: In-dataset digital qualitative evaluation.

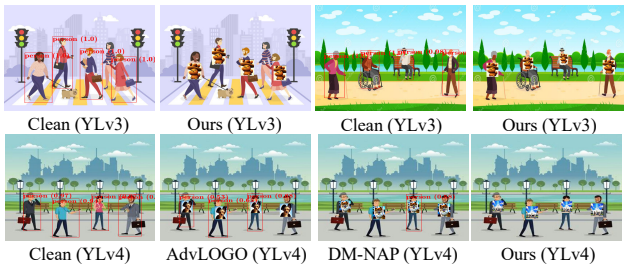


Figure 8: Cross-dataset digital qualitative evaluation.

Victimized Detectors and Comparison Attackers. We evaluate on six object detectors: five YOLO variants (v2, v3, v3-tiny, v4, v4-tiny) as one-stage models, and Faster R-CNN as a two-stage model. Seven SOTA attackers are compared, including LAP (Tan et al. 2021), UPC (Huang et al. 2020), DAP (Guesmi et al. 2024), NAP (Hu et al. 2021), PG-ECAP (Li et al. 2024), DM-NAP (Lin et al. 2025), and AdvLogo (Miao et al. 2024). Among them, LAP corresponds to the first paradigm summarized above, UPC and DAP to the second, and the remaining attackers to the third.

Implementation Details. We adopt the pretrained Stable Diffusion 1.4 as generative model, and generate 256×256 patch with DDIM sampling (50 steps, guidance scale 7.5). For dual-optimization, we perform 630 epochs totally, alternating prompt and latent updates every 15 and 200 epochs. Prompt optimization uses AdamW with learning rate $\eta_1 = 0.003$, while latent optimization uses MI-FGSM with $\mu = 1, \eta_2 = 0.02$, and $\tau = 0.2$. All experiments are conducted with PyTorch platform (Paszke et al. 2019).

Evaluation on the Digital Domain

Naturalness Evaluation. In Fig. 5, we visualize the generated adversarial patches. To assess visual naturalness, we conduct a human study with 45 participants, each selecting the four most natural-looking patches from randomized sets. Naturalness scores are computed as the percentage of votes received. Our YOLOv3-targeted patch achieves the highest score of 75.56%, with six of our patches ranking in the top nine, demonstrating strong perceptual realism.

Semantic-Level and Instance-Level Optimization. As shown in Fig. 6, we present patches generated during iterative optimization. Starting from an initialized dog image, the patch progressively evolves under attack-driven objectives, exhibiting adaptive semantic transitions across and within classes. This demonstrates the effectiveness of our dual-level optimization, substantially expanding the adversarial search space beyond the initial prompt distribution.

Qualitative Evaluation. We present detection results on the INRIA set and three generalization datasets. As shown in Fig. 7, our method fools detectors with targets missed, while other methods retain correct detections. Interestingly, in generalization experiments (Fig. 8), our method remains robust to domain and style shifts, maintaining high attack even on cartoon characters, where others often fail. This suggests our attack exploits model-level vulnerabilities beyond dataset-specific cues, enabling strong cross-domain transferability.

Quantitative Evaluation. Table 1 presents the **in-dataset**

(Trained on)	Attack Method	YOLOv2	YOLOv3	YOLOv3-t	YOLOv4	YOLOv4-t	F-RCNN
YOLOv2	UPC	48.62	54.40	63.82	64.21	63.03	61.87
	LAP	63.83	64.06	44.33	57.61	50.08	61.21
	PG-ECAP	9.70	36.04	-	52.59	-	-
	Ours	**4.37**	50.97	20.27	30.08	26.00	42.03
YOLOv3	DAP	-	32.63	37.13	44.31	38.08	-
	AdvLogo	35.40	45.13	37.65	49.16	49.19	45.95
	DM-NAP	22.59	28.51	26.15	59.61	40.95	53.27
	Ours	30.06	**22.87**	24.53	24.58	39.50	41.00
YOLOv3-t	NAP	31.61	28.81	10.02	65.13	31.62	55.08
	DAP	-	35.93	6.54	43.96	35.21	-
	DM-NAP	34.32	38.03	8.21	66.23	34.51	58.14
	Ours	23.55	50.95	**5.63**	38.64	18.64	51.33
YOLOv4	NAP	44.27	56.59	56.61	22.63	58.23	59.42
	DAP	-	50.21	51.33	24.65	48.80	-
	AdvLogo	54.38	67.69	68.15	60.92	73.18	61.71
	DM-NAP	53.48	67.15	51.44	14.95	62.01	60.34
	Ours	49.31	53.19	40.56	**13.71**	50.40	49.71
YOLOv4-t	NAP	34.68	37.79	21.69	46.80	23.70	59.97
	DAP	-	41.36	26.47	45.18	16.98	-
	AdvLogo	53.97	69.04	74.94	71.49	78.17	75.58
	DM-NAP	22.73	34.69	23.20	69.13	23.15	50.70
	Ours	24.07	49.67	9.62	57.58	**9.64**	49.95
F-RCNN	NAP	28.26	39.05	37.06	51.46	28.68	42.47
	AdvLogo	32.52	54.65	37.85	43.88	49.70	46.44
	DM-NAP	30.77	53.60	37.12	62.93	47.53	37.71
	Ours	26.14	33.57	24.14	37.68	30.12	**31.99**

Table 1: In-dataset digital quantitative evaluation. (Bold: optimal performance, underline: suboptimal performance)

evaluation on INRIA dataset. Our method consistently outperforms prior methods across all six victim detectors, demonstrating clear superiority. For example, our patch reduces the AP of YOLOv2 from 63.83% (LAP) and 9.70% (PG-ECAP) to 4.37%, and lowers YOLOv4-t's AP to 9.64%, significantly outperforming the suboptimal 16.98% achieved by DAP. While DM-NAP, another diffusion-based method, ranks second overall, its performance still remains limited by the adversarial search constrained within the fixed semantic distribution, resulting in a notably higher AP than ours. In contrast, our method performs adaptive optimization over inter-class semantics and intra-class instances, enabling broader adversarial exploration and yielding substantial gains. The **cross-dataset evaluation** presented in Table 2 further substantiates this advantage. Specifically, our method reduces YOLOv4's AP on the CCTV-Person dataset from the highest 35.59% achieved by AdvLogo and the suboptimal 5.00% obtained by DM-NAP to only 3.45%. This remarkable reduction highlights the superior generalization of our adversarial patches.

Dataset	Method	YOLOv3	YOLOv4
COCO-Person	NAP	35.24	33.56
	AdvLogo	42.88	50.51
	DM-NAP	35.57	28.32
	Ours	**32.03**	**25.79**
CCTV-Person	NAP	27.71	5.87
	AdvLogo	31.36	35.59
	DM-NAP	14.37	5.00
	Ours	**13.87**	**3.45**
Self-Collected	NAP	34.76	13.79
	AdvLogo	45.11	49.70
	DM-NAP	27.23	10.68
	Ours	**18.14**	**10.24**

Table 2: Cross-dataset digital quantitative evaluation.

Application on the Physical Domain

Physical experiments are conducted to comprehensively evaluate the practical effectiveness of our adversarial patch in real-world deployment scenarios. All patches are printed on 40×40cm carriers and affixed to volunteers photographed under diverse scenes. As shown in Fig. 9, across four detectors, our patches exhibit strong deception, significantly reducing confidence scores and inducing missed pedestrian detections. In contrast, competing methods fail to substantially disrupt detection, with models still correctly identify-

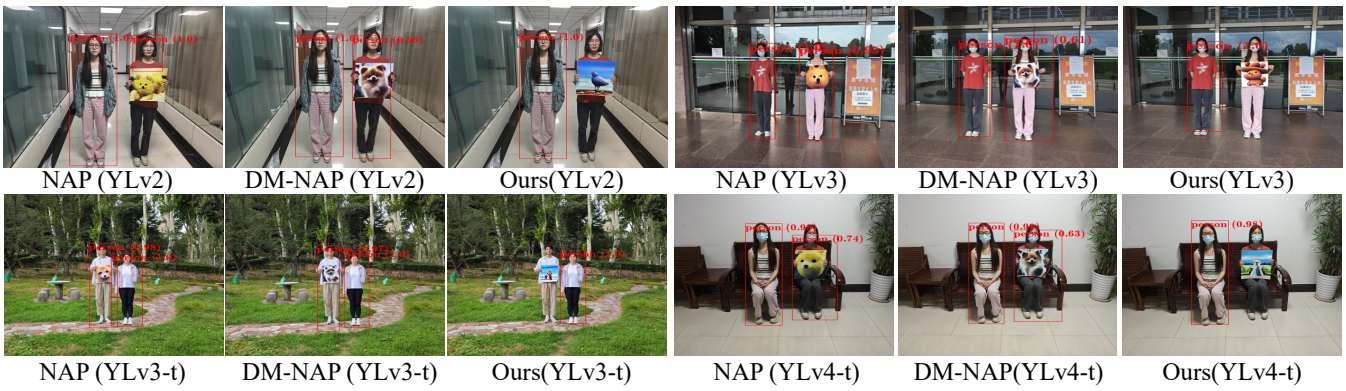


Figure 9: Qualitative evaluation conducted on real captured images in the physical-world domain.

Ablation	Configurations	Prompt	Patch Size	AP	Patch
Exp. I	w/o latent refinement	"A high-quality image of a dog"	256×256	32.97	
Exp. II	w/o prompt tuning	"A high-quality image of a dog"	256×256	56.53	
Exp. III	w/o prompt tuning	"A high-quality image of a dog"	512×512	24.24	Exp. III
Exp. IV	w/o prompt tuning	"Cat stretching on sunlit windowsill"	512×512	53.88	Exp. IV
Exp. V	w/o prompt tuning	"A sunflower field under a bright blue sky"	512×512	60.11	
Exp. VI	w/o prompt tuning	"A photograph of an astronaut riding a horse"	512×512	25.35	Exp. V
Exp. VII	Ours	"A high-quality image of a dog"	256×256	**22.87**	Exp. VI

Table 3: Ablation experiment results in the test set of INRIA Person conducted with the YOLOv3 detector.

ing pedestrians. This demonstrates our significant advantage in the physical domain.

Ablation Studies

We conduct ablation studies on specific designs and key parameters to verify their effectiveness.

Iterative Optimization at Semantic-Instance Levels. In the class-optimized diffusion and search, we perform iterative semantic-level prompt tuning and instance-level latent refinement, enabling broad semantic traversal and fine-grained adversarial enhancement. As shown in Table 3, removing either component degrades performance: semantic-only optimization with frozen latents (Exp. I) yields 32.97% AP, while instance-only optimization with fixed prompts (Exp. II) results in 56.53% AP. In comparison, our full strategy performs coordinated optimization at both semantic and instance levels, effectively expanding the adversarial distribution search and achieving the lowest AP of 22.87%.

Resolution Scaling Effects. To analyze the role of patch resolution, we compare 256×256 and 512×512 settings under identical prompts and optimizers, using instance-level latent optimization only. As evidenced by the comparison between Exp. II and Exp. III, increasing the resolution leads to a notable attack improvement, with AP decreasing from 56.53% to 24.24%. This is attributed to that diffusion models inherently exhibit better performance at the higher 512 resolution, enabling finer feature control and thereby achieving stronger attacks with preserved visual realism. Nevertheless, despite this improvement, the performance still falls short of our full method, even when applied at the lower 256px resolution.

Textual Conditioning Sensitivity. We further examine the effect of prompt selection. As shown in Table 3, corresponding to Exp. III through VI, AP scores vary from 24.24% to 60.11%, indicating strong sensitivity to textual prompts. This clearly reveals a key limitation of existing methods relying on manually selected prompts, as fixed semantic priors may deviate significantly from the optimal adversarial distribution. Poor prompts often lead to weak attacks, further underscoring the importance of our iterative optimization across semantic-instance levels.

Conclusion

We present the first systematic categorization of naturalistic adversarial patch generation methods into three representative paradigms, revealing key limitations in achieving both naturalness and aggressiveness. Building on these insights, we propose Diff-NAT, a novel paradigm that leverages natural generative priors with class-optimized diffusion. Diff-NAT establishes iterative optimization over semantic-level prompt and instance-level latent code, enabling broad inter-class and fine-grained intra-class exploration. This facilitates progressive movement toward adversarial distributions within the broad natural manifold under adversarial constraints. Extensive digital and physical experiments demonstrate our superiority in both naturalness and aggressiveness.

Acknowledgments

This work was supported by the NSFC (62506268, U23B2050, and 62276192), and the Natural Science Foundation of Jiangsu Province (BK20250454).

References

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Chen, S.-T.; Cornelius, C.; Martin, J.; and Chau, D. H. 2018. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 52–68.

Chen, Z.; Li, B.; Wu, S.; Jiang, K.; Ding, S.; and Zhang, W. 2023. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36: 51719–51733.

Cheng, Z.; Hu, Z.; Liu, Y.; Li, J.; Su, H.; and Hu, X. 2024. Full-distance evasion of pedestrian detectors in the physical world. *Advances in Neural Information Processing Systems*, 37: 102366–102392.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 886–893.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9185–9193.

Guesmi, A.; Ding, R.; Hanif, M. A.; Alouani, I.; and Shafique, M. 2024. Dap: A dynamic adversarial patch for evading person detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 24595–24604.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

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 International Conference on Computer Vision*, 7848–7857.

Hu, Z.; Chu, W.; Zhu, X.; Zhang, H.; Zhang, B.; and Hu, X. 2023. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16975–16984.

Hu, Z.; Huang, S.; Zhu, X.; Sun, F.; Zhang, B.; and Hu, X. 2022. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13307–13316.

Huang, H.; Chen, Z.; Chen, H.; Wang, Y.; and Zhang, K. 2023. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 20514–20523.

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 Conference on Computer Vision and Pattern Recognition*, 720–729.

Komkov, S.; and Petiushko, A. 2021. Advhat: Real-world adversarial attack on arcface face id system. In *Proceedings of the International Conference on Pattern Recognition*, 819–826.

Li, C.; Yan, H.; Zhou, L.; Chen, T.; Liu, Z.; and Su, H. 2024. Prompt-guided environmentally consistent adversarial patch. *arXiv preprint arXiv:2411.10498*.

Lin, S.-Y.; Chu, E.; Yeh, P.-H.; Chen, J.-C.; and Wang, J.-C. 2025. Diffusion to confusion: Naturalistic adversarial patch generation based on diffusion model for object detector. In *Proceedings of the IEEE International Conference on Image Processing*, 2378–2383.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Miao, B.; Li, C.; Zhu, Y.; Sun, W.; Wang, Z.; Wang, X.; and Xie, C. 2024. Advlogo: Adversarial patch attack against object detectors based on diffusion models. *arXiv preprint arXiv:2409.07002*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qiu, H.; Xiao, C.; Yang, L.; Yan, X.; Lee, H.; and Li, B. 2020. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Proceedings of the European Conference on Computer Vision*, 19–37.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2256–2265.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 1–9.

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 ACM International Conference on Multimedia*, 5307–5315.

Thys, S.; Van Ranst, W.; and Goedemé, T. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–7.

Wei, H.; Tang, H.; Jia, X.; Wang, Z.; Yu, H.; Li, Z.; Satoh, S.; Van Gool, L.; and Wang, Z. 2024a. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9797–9817.

Wei, H.; Wang, Z.; Zhang, K.; Hou, J.; Liu, Y.; Tang, H.; and Wang, Z. 2024b. Revisiting adversarial patches for designing camera-agnostic attacks against person detection. *Advances in Neural Information Processing Systems*, 37: 8047–8064.

Wei, X.; Huang, Y.; Sun, Y.; and Yu, J. 2023. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2348–2363.

Wei, X.; Yu, J.; and Huang, Y. 2024. Infrared adversarial patches with learnable shapes and locations in the physical world. *International Journal of Computer Vision*, 132(6): 1928–1944.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, 13095–13105.

Xu, K.; Zhang, G.; Liu, S.; Fan, Q.; Sun, M.; Chen, H.; Chen, P.-Y.; Wang, Y.; and Lin, X. 2020. Adversarial t-shirt! evading person detectors in a physical world. In *Proceedings of the European Conference on Computer Vision*, 665–681.

Yi, X.; Ma, Y.; Li, Y.; Xu, H.; and Ma, J. 2025a. Artificial intelligence facilitates information fusion for perception in complex environments. *The Innovation*, 6(4): 100814.

Yi, X.; Tang, L.; Zhang, H.; Xu, H.; and Ma, J. 2024. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110: 102450.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2025b. Diff-Retinex++: Retinex-Driven Reinforced Diffusion Model for Low-Light Image Enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8): 6823–6841.

Yue, J.; Fang, L.; Xia, S.; Deng, Y.; and Ma, J. 2023. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32: 5705–5720.

Zhu, X.; Liu, Y.; Hu, Z.; Li, J.; and Hu, X. 2024. Infrared adversarial car stickers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 24284–24293.