DQ-PAD: DISCRIMINATIVE QUANTIZED PROTOTYPE ALIGNMENT FOR ADVERSARIAL PATCH DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial patches with diverse appearances challenge the generalization of defenses for object detection. Existing methods mainly localize adversarial patches by detecting statistical anomalies, learning discriminative patch features, or measuring reconstruction errors from generative models, but may struggle to distinguish diverse patches from benign regions. To address this issue, we propose Discriminative Quantized Prototype Alignment for Adversarial Patch Defense (DQ-PAD), which identifies adversarial patches by measuring how well image regions match benign visual patterns. DQ-PAD uses prototypes from a pretrained VQ-VAE2 codebook to represent benign visual patterns and computes an anomaly score based on the distance between each regional feature and its nearest prototype. However, since VQ-VAE2 is primarily trained for image reconstruction, some benign regions may match these prototypes poorly, while some patch regions may remain close to them. To address this limitation, we introduce patch-aware projection learning, which learns a projection matrix from labeled benign and patch regions while keeping the encoder and codebook frozen. The projection jointly maps features and prototypes into a low-dimensional space and selects more discriminative feature directions to reduce the matching distance of benign regions while enlarging the distance gap between patch and benign regions. In the pedestrian detection setting, we compare DQ-PAD with seven defense baselines using ten held-out adversarial patches and three object detectors, and demonstrate that DQ-PAD improves the average defense performance by up to 8.96% over existing SOTA methods. Physical-world experiments further demonstrate its applicability in real-world scenarios.

1 INTRODUCTION

As adversarial attack techniques continue to evolve, the visual appearances of adversarial patches have become increasingly diverse, making them an important security threat to object detection systems in real-world scenarios Wei et al. (2024); Wang et al. (2023); Cheng et al. (2024); He et al. (2024); Tan et al. (2024). Attackers can print adversarial patches generated in the digital domain and place them on specific regions of target objects, thereby disrupting normal detector predictions and significantly degrading detection performance. Therefore, developing effective defense methods is essential for ensuring the safe and reliable deployment of object detection systems.

Existing defense methods can be broadly divided into certified defenses Xiang et al. (2021); Xiang & Mittal (2021); Xiang et al. (2022; 2023) and empirical defenses Naseer et al. (2018); Tarchoun et al. (2023); Liu et al. (2022); Wu et al. (2024); Lin et al. (2024); Kang et al. (2024); Chattopadhyay et al. (2025); Li et al. (2025); Kumar & Agarwal (2025); Wang et al. (2026). Certified defenses aim to construct robust inference mechanisms that ensure reliable model predictions against arbitrary attacks under predefined constraints. However, attackers' capabilities often exceed these predefined constraints, making such methods difficult to apply in practical scenarios. Moreover, their defense procedures typically incur substantial computational overhead, further limiting their applicability to real-time object detection. In contrast, empirical defenses aim to localize and eliminate adversarial patches from input images before object detection. These methods typically impose no strict constraints on attackers' capabilities, allowing them to better adapt to different attack scenarios and detector architectures. According to their patch localization mechanisms, empirical defenses can be further classified into three categories: *1)* statistics-based methods, which localize patch regions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by detecting anomalies in low-level statistical properties of adversarial patches, such as frequency-domain features and entropy; *2)* discriminative model-based methods, which train auxiliary detection or segmentation networks to learn discriminative patch features and subsequently remove the detected patches; and *3)* generative model-based methods, which exploit the discrepancy between the benign distribution prior learned by generative models during pretraining and the distribution of adversarial features, identifying patch regions through image reconstruction errors.

Although the aforementioned empirical defense methods have achieved significant success, they still have clear limitations. On the one hand, statistics-based and discriminative model-based methods rely on predefined statistical anomalies and patch features observed during training, respectively, which limits their generalization to unseen patches. On the other hand, generative model-based methods lack sufficient discriminability for adversarial patches during reconstruction and are also prone to altering the original image content and introducing semantic distortions, thereby degrading post-defense object detection performance.

These limitations motivate us to develop an efficient defense that generalizes well to diverse adversarial patches. Our core idea is to match image regions against known benign visual patterns to identify suspicious patches, thereby alleviating the generalization limitations caused by diverse patch appearances. Specifically, we leverage the pretrained codebook of a vector-quantized variational autoencoder (VQ-VAE2) Razavi et al. (2019) to represent benign visual patterns and establish a set of benign prototypes for normal regions. We then measure regional anomalies using the residual between local image features and benign prototypes, and identify high-residual regions as potential patch regions. However, since VQ-VAE2 is primarily designed for image reconstruction, its original latent space has limited ability to distinguish benign regions from patch regions. As a result, complex benign regions may also produce large residuals, while some patch regions may exhibit relatively small residuals. To address this issue, we design a patch-aware projection learning mechanism that learns a lightweight learnable projection matrix that jointly maps local features and benign prototypes into a more discriminative latent space. In this space, we optimize the matching between benign features and benign prototypes while increasing the inconsistency between patch features and benign prototypes, thereby improving the separability between benign and patch regions.

Our main contributions are summarized as follows:

- We propose DQ-PAD, an adversarial patch defense method based on Discriminative Quantized Prototype Alignment. Unlike defenses that rely on specific patch features, DQ-PAD leverages the representational capability of the pretrained VQ-VAE2 codebook for benign visual patterns to establish benign prototypes for normal regions, and identifies anomalous regions through the matching relationship between local features and benign prototypes, thereby improving generalization to diverse adversarial patches.
- To address the limited discriminative capability of the original VQ-VAE2 latent space, we design a patch-aware projection learning mechanism. This mechanism introduces a lightweight learnable projection matrix to map local features and benign prototypes into a low-dimensional discriminative subspace. By strengthening the matching between benign features and prototypes while enlarging the discrepancy between patch and benign features, it improves the separability between benign and patch regions.
- We comprehensively evaluate DQ-PAD against seven defense methods on ten different adversarial patches and multiple object detectors. Experimental results show that DQ-PAD consistently provides strong defense across diverse patch types, improving the average defense performance by up to 8.96% over SOTA methods. Meanwhile, DQ-PAD maintains high inference efficiency, achieving a favorable balance between defense effectiveness and computational cost.

2 Related Work

Certified defenses. These methods aim to construct robust inference mechanisms that maintain reliable model predictions even when attackers have full knowledge of the defense strategy. For example, PatchGuard Xiang et al. (2021) suppresses patch effects by restricting receptive fields and masking anomalous features; PatchCleanser Xiang et al. (2022) employs double masking and filters attack-affected predictions based on prediction consistency; and ObjectSeeker Xiang et al. (2023)

adopts a location-agnostic masking strategy and aggregates multiple predictions to achieve robust detection against patches with unknown shapes, sizes, and locations.

Empirical defenses. These methods capture anomalous cues in input images to localize and remove adversarial patches before object detection. Early methods identify anomalous cues through statistical properties. For example, LGS Naseer et al. (2018) detects patches based on high-frequency variations, and Jedi Tarchoun et al. (2023) identifies high-entropy regions. Subsequently, researchers train dedicated localization networks that exploit the anomalous activations of adversarial patches in feature space to achieve more precise localization, such as the segmentation-based SAC Liu et al. (2022) and the object-detection-based NAPGuard Wu et al. (2024). Recently, some studies regard adversarial patches as outliers that deviate from the normal visual distribution and leverage the ability of generative models to model benign distributions to locate patch regions by analyzing inconsistencies generated during image reconstruction. For example, NutNet Lin et al. (2024) localizes regions deviating from the normal distribution based on autoencoder reconstruction errors, while DIFFender Kang et al. (2024) identifies adversarial patches by exploiting the inconsistency among multiple generation results produced by the diffusion model.

3 Threat Model

3.1 Attack Model

Adversary's Capabilities and Objectives. We assume that the attacker has white-box access to both the target object detector and the defense model, allowing adaptive adversarial patches to be constructed. Given an input image $\mathbf{x} \in \mathbb{R}^{W\times H\times C}$, a binary mask $\mathbf{m} \in \{0,1\}^{W\times H}$, and patch content $\boldsymbol{\delta} \in [0,1]^{W\times H\times C}$, the adversarial image is formulated as

$$\mathbf{x}_{\text{adv}} = \mathbf{m} \odot \boldsymbol{\delta} + (1-\mathbf{m}) \odot \mathbf{x}, \tag{1}$$

where $\mathbf{m}$ controls the number, size, shape, and location of the patches, while $\boldsymbol{\delta}$ determines their visual content, ranging from conspicuous unnatural patterns to natural-looking appearances. The attacker's objective is to induce missed detections or misclassifications while evading localization and suppression by the defense model.

3.2 Defense Model

Defender's Capabilities and Objectives. During training, we assume that the defender has access to clean images and one adversarial patch to construct training samples and learn the projection matrix, while keeping the pretrained VQ-VAE2 and its codebook frozen. After deployment, the defender has no prior knowledge of whether an input image contains a patch, nor of the patch's type, number, size, shape, or location. In addition, the adversarial patches used during training and testing do not overlap, so as to evaluate the defense model's generalization ability to unseen patches. Given an input image $\mathbf{x}' \in \{\mathbf{x}, \mathbf{x}_{\text{adv}}\}$, the defense model computes the projected residuals between local features and their nearest codewords, thereby obtaining an anomaly map $\mathbf{A}(\mathbf{x}')$. The patch mask is defined as

$$\hat{\mathbf{m}}_{i,j} = \mathbb{I}[\mathbf{A}_{i,j}(\mathbf{x}') > \tau], \tag{2}$$

where τ is the localization threshold, and $\hat{\mathbf{m}}_{i,j} = 1$ indicates that the location is identified as a patch region. The defense model then masks the localized region:

$$\mathcal{D}(\mathbf{x}') = (1-\hat{\mathbf{m}}) \odot \mathbf{x}'. \tag{3}$$

The defense objective is to accurately cover adversarial patches while minimizing the removal of benign image content, keeping the detection results of clean images unchanged and restoring the normal detection results of attacked images:

$$f(\mathcal{D}(\mathbf{x})) = f(\mathbf{x}), \qquad f(\mathcal{D}(\mathbf{x}_{\text{adv}})) \approx f(\mathbf{x}). \tag{4}$$

4 Method

4.1 Overview

The core idea of DQ-PAD is to leverage the codebook's ability to represent benign visual patterns to establish a set of prototypes for normal regions, and identify anomalous regions through the

matching discrepancy between local latent features and their nearest prototypes, thereby reducing dependence on fixed patch patterns. In addition, we introduce a patch-aware projection learning mechanism that maps features into a more discriminative subspace to enlarge the discrepancy between benign and adversarial features, enabling more accurate patch localization. Fig. 1 illustrates the framework diagram of our approach.

4.2 Codebook Matching-Based Anomaly Measurement

Given an input image $\mathbf{x}$, we use a frozen VQ-VAE2 to extract the Bottom-level latent features $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{hw}] \in \mathbb{R}^{h \times w \times D}$, where $\mathbf{z}_i \in \mathbb{R}^D$ denotes the feature at the i-th spatial location. Meanwhile, we treat the codebook $\mathcal{E} = \{\mathbf{e}_k\}_{k=1}^K$ pretrained by the VQ-VAE2 as a collection of benign visual prototypes, where K denotes the number of codewords. We use Euclidean distance to measure the residual between a latent feature and its nearest prototype:

$$r(\mathbf{z}_i) = \min_{k \in \{1,\ldots,K\}} \|\mathbf{z}_i - \mathbf{e}_k\|_2^2 . \tag{5}$$

The residual measures the degree of matching between a latent feature and the benign prototypes. A larger residual indicates that the corresponding region is more difficult to represent using the existing benign prototypes, suggesting that it may deviate from the normal visual distribution and can therefore serve as a potential anomaly cue for adversarial patch localization.

However, because the pretrained codebook is primarily used to represent common visual patterns in benign images, it lacks sufficient discriminative capability for identifying adversarial patch features. Specifically, complex benign regions may produce large residuals, whereas some patch regions may still lie close to certain benign visual prototypes, causing the residual distributions of the two types of regions to overlap. Therefore, we propose a patch-aware projection learning mechanism.

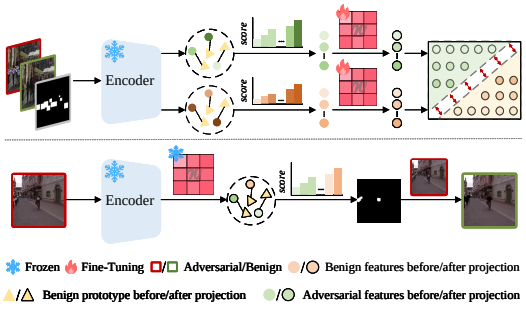

Figure 1: Overview of the proposed framework. The upper part illustrates the training stage, where the VQ-VAE2 and its codebook are frozen, and the codebook is treated as a set of benign visual prototypes. Benign and patch features are constructed according to the patch mask, and a learnable projection matrix $\mathbf{W}$ jointly maps local features and codebook prototypes into a discriminative subspace. In this space, the margin between adversarial and benign features is enlarged to improve the separability between benign and patch regions. The lower part shows the inference stage, where the learned $\mathbf{W}$ is used to compute the matching distance between local features and benign prototypes. Patch masks are then generated according to a threshold, mapped back to the input image, and used to remove the predicted patch regions, yielding the defended image.

4.3 Patch-Aware Projection Learning

Its core idea is to learn a low-dimensional subspace that better discriminates benign features from adversarial features, such that benign features remain well matched with the benign prototype while adversarial features exhibit larger matching residuals. Specifically, given a latent feature $\mathbf{z}_i \in \mathbb{R}^D$ and a benign prototype $\mathbf{e}_k \in \mathbb{R}^D$, we use the learnable projection matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$ to map both into a low-dimensional space, where $d < D$. Accordingly, the residual between a latent feature and its nearest benign prototype in the projected space is defined as:

$$r_w(\mathbf{z}_i) = \min_{k \in \{1,\ldots,K\}} \|(\mathbf{z}_i - \mathbf{e}_k)\mathbf{W}\|_2^2 , \tag{6}$$

where the nearest benign prototype is reassigned in the projected space, allowing the correspondence between each feature and its nearest prototype to be updated as the discriminative subspace is learned. Additionally, we impose an orthogonality constraint on $\mathbf{W}$ ($\mathbf{W}^\top \mathbf{W} = \mathbf{I}_d$) to prevent trivial residual scaling and encourage the projection to learn the most discriminative feature directions.

Discriminative subspace learning. To improve the separability of benign features and adversarial features in the subspace, we use the patch mask $\mathbf{M}$ to divide the latent features into a benign feature

set $\mathcal{N}$ and a patch feature set $\mathcal{P}$, and use the two types of features as supervision signals, thereby guiding the projection matrix $\mathbf{W}$ to learn a discriminative subspace that distinguishes normal regions from patch regions. Ideally, adversarial features should have larger residuals than benign features. However, we find that the separability of the two feature types is primarily limited by benign features with large residuals and adversarial features with small residuals. Therefore, in each training batch, we select the largest-residual features from $\mathcal{N}$ to form the hard benign set $\mathcal{H}_N$ and the smallest-residual features from $\mathcal{P}$ to form the hard patch set $\mathcal{H}_P$. We then impose the following constraint on these hard samples:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{H}_N||\mathcal{H}_P|} \sum_{\mathbf{n}\in\mathcal{H}_N} \sum_{\mathbf{p}\in\mathcal{H}_P} \text{softplus}\left(\delta + r_w(\mathbf{n}) - r_w(\mathbf{p})\right), \tag{7}$$

where δ is the residual margin and $\text{softplus}(u) = \log(1 + \exp(u))$. This loss encourages hard adversarial features to have larger residuals than hard benign features, thereby increasing the residual margin between the two hard-sample sets and reducing their overlap near the decision boundary.

Since the above loss only constrains the relative residuals between patch and benign features and cannot ensure sufficient matching between benign features and benign prototypes, we further minimize the residuals between benign features and benign prototypes:

$$\mathcal{L}_{\text{compact}} = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{n}\in\mathcal{N}} r_w(\mathbf{n}). \tag{8}$$

This loss yields a more compact benign residual distribution by reducing the average residual over all benign features, thereby reducing the risk of misclassifying complex benign regions as patches.

Minimizing residuals alone may cause the projection to discard useful information in benign features. Therefore, to limit such information loss, we introduce the following feature information preservation constraint:

$$\mathcal{L}_{\text{preserve}} = \frac{1}{|\mathcal{C}|D} \sum_{\mathbf{c}\in\mathcal{C}} \left\|\mathbf{c} - \mathbf{c}\mathbf{W}\mathbf{W}^\top\right\|_2^2. \tag{9}$$

Here, $\mathcal{C}$ denotes the set of features extracted from benign images, D is the original feature dimension, and $\mathbf{c}\mathbf{W}\mathbf{W}^\top$ denotes the representation obtained by projecting a feature into the low-dimensional space and mapping it back to the original space. By reducing the discrepancy between the original benign features and their recovered representations, this constraint encourages the projection subspace to retain the principal information in benign features.

Finally, we optimize the projection matrix using the following joint objective:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda_{\text{normal}}\mathcal{L}_{\text{compact}} + \lambda_{\text{preserve}}\mathcal{L}_{\text{preserve}}, \tag{10}$$

The three loss terms respectively increase the residual margin between hard samples, compact the benign residual distribution, and preserve information in benign features. During training, the VQ-VAE2 and its codebook remain frozen, and only the projection matrix $\mathbf{W}$ is updated. We initialize $\mathbf{W}$ using the first d principal-component directions of benign features.

Inference. Given an input image $\mathbf{x}$, we extract local latent features using the frozen VQ-VAE2 and project them together with benign prototypes into a low-dimensional subspace via $\mathbf{W}$. The anomaly score s_i is computed from the distance to the nearest benign prototype, and regions satisfying $s_i \geq \tau$ are identified as patch regions to form a binary mask in the latent space. The mask is then upsampled to the input resolution and applied to the input image to obtain the defended image.

5 Experiments

5.1 Experimental Setup

Dataset. We primarily focus on pedestrian detection scenarios and conduct experiments in both digital and physical domains. In the digital domain, we use the INRIA Person dataset Dalal & Triggs (2005) as the primary evaluation benchmark, randomly selecting 300 images for training and the remaining 602 images for testing. To further evaluate the cross-dataset transferability of our

Table 1: Performance (AP@0.5) of various defense methods on the INRIA dataset across different patches and models. Gray background indicates the results of our proposed method. Bold and underlined values represent the optimal and sub-optimal performance, respectively. *Avg.* represents the average defensive efficacy across all adversarial patch attacks.

Detector	Defense	Clean	Non-NAP				NAP						*Avg.*
			AdvC	AdvP	T-SEA	CAP	P1	P2	P3	P4	P5	P6	
FRCNN	Undefended	83.08	23.39	44.90	26.09	56.76	62.62	42.84	60.72	58.81	56.51	65.33	49.80
	LGS Naseer et al. (2018)	82.46	26.85	55.11	32.17	57.49	63.00	48.47	54.36	62.56	62.68	70.89	53.36
	SAC Liu et al. (2022)	83.12	29.30	66.08	56.11	**78.08**	62.49	42.79	63.22	58.89	56.47	72.07	58.55
	Jedi Tarchoun et al. (2023)	83.08	25.19	43.30	37.08	55.77	69.41	47.01	65.04	53.66	53.31	67.68	51.75
	OS Xiang et al. (2023)	82.58	20.07	42.92	24.46	56.73	62.43	45.76	54.31	58.30	56.01	64.19	48.52
	PAD Jing et al. (2024)	82.49	39.06	67.31	61.30	72.22	69.58	**71.23**	**76.18**	64.79	72.18	77.20	67.11
	NutNet Lin et al. (2024)	82.15	51.76	74.04	70.78	72.49	65.23	52.10	70.42	70.36	66.03	76.88	67.01
	Saliuitl Byrd Victorica et al. (2025)	**83.15**	42.00	60.88	41.47	63.82	67.02	49.79	58.61	59.59	63.81	66.85	57.38
	Ours	82.10	**66.59**	**76.49**	**76.14**	76.61	**72.13**	70.58	70.82	**78.02**	**78.32**	**78.03**	**74.37**
YOLOv3	Undefended	79.72	3.80	30.78	8.14	63.27	47.12	56.91	45.89	55.62	39.53	56.78	40.78
	LGS Naseer et al. (2018)	79.36	7.10	44.67	13.11	65.68	47.75	62.32	49.24	57.42	46.82	64.01	45.81
	SAC Liu et al. (2022)	79.72	8.01	61.76	37.95	75.63	47.07	56.83	56.37	55.90	39.61	66.17	50.53
	Jedi Tarchoun et al. (2023)	**80.20**	8.47	42.89	20.25	63.96	55.73	63.64	58.68	53.21	46.04	61.04	47.39
	OS Xiang et al. (2023)	79.13	1.76	22.25	17.86	54.59	47.18	51.64	44.52	48.68	38.22	51.00	37.77
	PAD Jing et al. (2024)	79.43	25.90	62.46	59.56	73.70	65.84	**69.56**	**72.21**	64.64	65.97	71.67	63.15
	NutNet Lin et al. (2024)	78.90	45.86	74.08	68.66	72.70	57.24	58.40	62.63	67.48	59.37	72.04	63.85
	Saliuitl Byrd Victorica et al. (2025)	79.96	44.85	71.10	52.53	63.50	59.59	69.04	53.61	73.26	69.15	74.64	63.13
	Ours	79.35	**67.86**	**74.43**	**73.31**	**75.93**	**73.09**	67.62	68.00	**75.58**	**76.12**	**76.11**	**72.81**
RT-DETR	Undefended	77.90	11.94	46.81	18.77	69.50	54.86	48.78	47.26	59.48	50.57	59.16	46.71
	LGS Naseer et al. (2018)	77.14	16.94	55.15	22.58	67.76	56.57	51.62	48.38	61.11	56.38	61.97	49.85
	SAC Liu et al. (2022)	77.95	17.73	64.96	37.59	**74.66**	54.82	48.62	51.72	59.45	50.55	64.77	52.49
	Jedi Tarchoun et al. (2023)	**79.53**	17.67	43.59	26.45	65.42	63.61	58.24	52.56	55.68	54.15	57.66	49.50
	OS Xiang et al. (2023)	78.25	14.37	54.01	22.73	74.26	65.31	61.92	59.93	66.55	57.16	64.72	54.10
	PAD Jing et al. (2024)	77.65	32.66	64.52	57.99	72.35	67.10	69.43	70.54	68.04	69.31	71.34	64.33
	NutNet Lin et al. (2024)	76.95	54.43	72.55	64.72	72.54	61.06	52.95	69.25	69.85	65.85	72.91	65.61
	Saliuitl Byrd Victorica et al. (2025)	77.23	51.56	72.66	51.00	69.49	**70.61**	66.29	57.09	72.06	70.97	75.85	65.76
	Ours	77.56	**55.42**	**73.48**	**70.03**	73.80	67.17	**69.49**	**70.59**	**79.11**	**79.21**	**78.38**	**71.67**

method, we additionally employ the CCTVPerson [1] and WiderPerson1000 datasets, where WiderPerson1000 is constructed by randomly sampling 1,000 images from the WiderPerson dataset Zhang et al. (2019). In the physical domain, we use a Huawei Mate 80 smartphone to record videos of handheld adversarial patches and construct a physical evaluation dataset by extracting video frames.

Attack Configurations. To comprehensively evaluate the defense performance, we test 10 adversarial patches generated by six different methods, including non-natural adversarial patches (Non-NAP): AdvClock Wu et al. (2020), AdvPatch Thys et al. (2019), CAP Huang et al. (2023), and T-SEA Huang et al. (2023); and natural adversarial patches (NAP): GAN-based NAP (p1–p3) Hu et al. (2021) and diffusion-based AdvLogo (p4–p6) Miao et al. (2024). All ten test patches are strictly excluded from training to evaluate the generalization capability of our method to unseen patches. For object detectors, we select three mainstream models: the two-stage Faster R-CNN Ren et al. (2016), the one-stage YOLOv3 Redmon & Farhadi (2018), and the Transformer-based RT-DETR Zhao et al. (2024). In all experiments, adversarial patches are placed at the center of the target bounding box with a scale factor of 0.2.

Defense Configurations. We conduct a comprehensive comparison with seven representative defense methods, including statistical-feature-based methods LGS Naseer et al. (2018), Jedi Tarchoun et al. (2023), PAD Jing et al. (2024), and Saliuitl Byrd Victorica et al. (2025); the segmentation-model-based method SAC Liu et al. (2022); the generative-model-based method NutNet Lin et al. (2024); and the certified defense method ObjectSeeker (OS) Xiang et al. (2023). Since SAC and Saliuitl are training-based methods, we retrain them using 300 images to ensure a fair comparison, while the remaining methods follow the original settings described in their respective papers. For our method, the initial learning rate is set to 0.0005, and the number of training iterations is set to 20. All experiments are conducted on an NVIDIA RTX A6000 GPU with 48GB of memory. We adopt average precision (AP@0.5) as our primary evaluation metric, where a higher AP indicates better defense performance on the defended samples.

[1] CCTVPerson dataset website.

5.2 Defense Effectiveness

Table 1 presents the defense performance of different methods on three object detectors. Our method achieves average AP of 74.37%, 72.81%, and 71.67% on Faster R-CNN, YOLOv3, and RT-DETR, respectively, outperforming the corresponding second-best methods by 7.26%, 8.96%, and 5.91%. These results demonstrate that our method generalizes well across different types of adversarial patches. Meanwhile, the AP on clean images decreases by only 0.98%, 0.37%, and 0.34%, indicating that our method has limited impact on normal detection performance. The qualitative results in Figure 7 further support the quantitative analysis. Compared with existing methods, which often miss patch regions or misclassify benign backgrounds, our method can more accurately localize adversarial regions under different patches, including AdvP, T-SEA, P1, and P4, while reducing false activations on benign backgrounds. This enables effective patch removal while preserving more benign image content.

5.3 Efficiency

We integrate different defense methods into three object detectors and measure their end-to-end inference efficiency. The results are shown in Fig. 2 and Table 2. Our method achieves 30.30, 27.03, and 18.87 FPS on YOLOv3, Faster R-CNN, and RT-DETR, respectively. Compared with PAD, our method improves the inference speed by approximately 79–113×. In addition, compared with lightweight methods such as LGS, SAC, and NutNet, our method maintains comparable inference efficiency while achieving higher defense performance, demonstrating a favorable balance between defense effectiveness and computational efficiency.

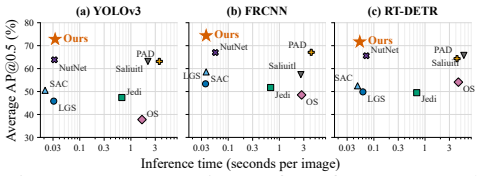

Figure 2: Comparison of performance and efficiency.

Detector	LGS	SAC	Jedi	OS	PAD	NutNet	Saliuitl	Ours
FRCNN	27.78	27.03	1.48	0.37	0.24	17.86	0.38	27.03
YOLOv3	32.26	47.62	1.50	0.61	0.28	31.25	0.46	30.30
RT-DETR	16.39	20.83	1.45	0.22	0.24	14.08	0.17	18.87

Table 2: Inference efficiency (FPS) of different defense methods.

5.4 Robustness Analysis

We consider a white-box adversary with complete knowledge of DQ-PAT. The adversary optimizes a universal adversarial patch to evade the defense while retaining its effectiveness against the target detector. We evaluate two adaptive settings: an end-to-end attack that optimizes the patch through the complete defense–detection pipeline, and a non-end-to-end attack that directly targets the patch-localization mechanism. We report AP@0.5 before and after defense as Att and Def, respectively. A lower Att indicates stronger object-hiding capability, whereas a higher Def indicates better detection recovery. Both settings are built upon AdvPatch. The non-end-to-end experiments use YOLOv3 as the target detector, while the end-to-end experiments optimize detector-specific adaptive patches for Faster R-CNN, YOLOv3, and RT-DETR.

In the end-to-end setting, we feed the image processed by DQ-PAT into the target detector and optimize the patch using the detection loss. Let $\mathcal{A}(\mathbf{x}, \mathbf{p})$ denote the operation of applying patch $\mathbf{p}$ to image $\mathbf{x}$, $\mathcal{D}_{\mathrm{DQ}}$ denote DQ-PAT, and f denote the target detector. The optimization objective is

$$\mathcal{L}_{\mathrm{e2e}} = \mathcal{L}_{\mathrm{obj}} \left(f(\mathcal{D}_{\mathrm{DQ}} \left(\mathcal{A}(\mathbf{x}, \mathbf{p}) \right)), y \right). \tag{11}$$

Because the hard thresholding and masking operations in DQ-PAT are non-differentiable, we adopt a backward pass differentiable approximation (BPDA) Athalye et al. (2018). The forward pass applies the actual binary mask, while the backward pass uses a straight-through estimator (STE) Bengio et al. (2013) to provide surrogate gradients for updating the adversarial patch.

As shown in Fig. 3, end-to-end adaptive attacks substantially increase the AP under the Attack setting, indicating weakened attack capability, while the AP after applying DQ-PAD remains nearly unchanged. This shows that end-to-end adaptive optimization cannot effectively bypass DQ-PAD.

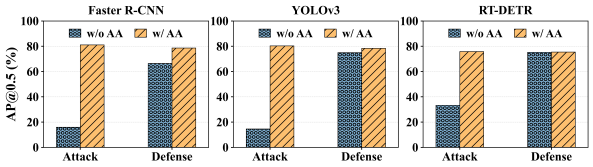

Figure 3: End-to-end adaptive attack performance. "w/o AA" and "w/ AA" represent non-adaptive and end-to-end adaptive attacks, respectively.

λ	Threshold		Mask		Joint	
	Att ↓	Def ↑	Att ↓	Def ↑	Att ↓	Def ↑
$\lambda = 0$	14.59	74.99	14.59	74.99	14.59	74.99
$\lambda = 0.5$	15.49	73.57	18.23	60.24	29.86	59.98
$\lambda = 1$	19.94	61.87	20.51	60.03	26.26	59.42

Table 3: Adaptive attack. Attack (Att) and Defense (Def) denote AP@0.5 before and after applying DQ-PAD, respectively.

To investigate whether an attacker can directly evade the patch-localization mechanism of DQ-PAT, we augment the original AdvPatch objective Thys et al. (2019) with a threshold-evasion loss (Threshold), a mask-inconsistency loss (Mask), or their combination (Joint). The overall objective is defined as

$$\mathcal{L}_{\text{adapt}}^{(k)} = \mathcal{L}_{\text{AdvPatch}} + \lambda \mathcal{L}_k, \qquad k \in \{\text{Threshold}, \text{Mask}, \text{Joint}\}. \tag{12}$$

Here, $\mathcal{L}_{\text{AdvPatch}}$ is computed on the undefended adversarial image to preserve the attack effectiveness against the target detector, whereas $\mathcal{L}_k$ reduces the probability that the patch is localized by DQ-PAT. The hyperparameter λ controls the strength of defense-oriented optimization. When $\lambda = 0$, the objective reduces to the standard AdvPatch attack.

Specifically, the threshold-evasion loss suppresses the projected codebook residuals within the patch region so that they fall below the localization threshold. The mask-inconsistency loss instead minimizes the overlap between a differentiable approximation of the predicted mask and the ground-truth patch mask. The joint variant combines both objectives, simultaneously reducing the anomaly scores of patch features and the spatial coverage of the predicted mask. The complete definitions of these losses are provided in Appendix B.

Table 3 presents the attack results under different adaptive objectives and values of λ. When $\lambda = 0$, the standard AdvPatch attack achieves Att and Def scores of 14.59 and 74.99, respectively. After introducing the adaptive objectives, the attacker can weaken the defense effectiveness of our method, but at the cost of sacrificing the original attack capability of the patch. Specifically, when $\lambda = 0.5$, the threshold-evasion attack only reduces the defense-side AP to 73.57%, indicating that simply suppressing local anomaly scores makes it difficult to bypass the overall localization mechanism. In contrast, the mask-inconsistency attack reduces the defense-side AP to 60.24%, demonstrating more effective defense evasion. Although the joint attack further reduces the defense-side AP to 55.98%, the attack-side AP increases substantially to 29.86%. This indicates that the mask-inconsistency loss provides only limited evasion gains. When λ is increased to 1, the defense-side AP of all three attacks remains stable at approximately 60%, while the attack-side AP increases, indicating that further strengthening the defense-evasion objective provides only limited benefits. Overall, adaptive optimization can degrade the defense performance of our method, but the attacker must make an explicit trade-off between evading patch localization and preserving the original attack effectiveness. In addition, we show in Appendix C the results of our method in different adversarial patch Settings as well as in different pedestrian scenarios

5.5 Ablation Study

Feature-layer selection. We first investigate why the Bottom-layer representation of the VQ-VAE2 is used for patch localization. Figure 4 compares the regional discrimination ability of Top-layer and Bottom-layer features across different adversarial patches. Most samples lie above the diagonal, indicating that the Bottom-layer features generally achieve higher AUC values than the Top-layer features, with particularly clear advantages on T-SEA and P2. Top-layer features undergo stronger semantic compression and therefore provide rich global semantics, but they are less sensitive to local texture anomalies and fine-grained spatial variations. In contrast, Bottom-layer features retain more local structural and spatial information, making them better suited to identifying localized adversarial regions. The qualitative results in Figure 5 further show that Bottom-layer features produce stronger and more concentrated responses within patch regions.

Effect of projection learning. Although the original Bottom-layer features are better to local anomalies, complex textures, shadows, and background patterns may also produce large prototype-

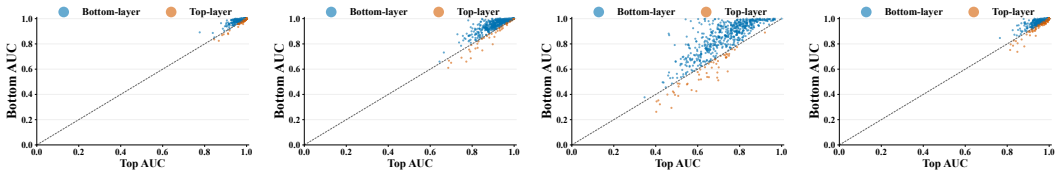

Figure 4: Comparison of Top-layer and Bottom-layer codebook matching performance under different adversarial patches.

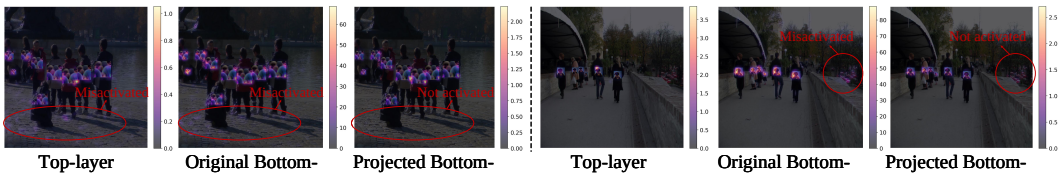

Figure 5: Qualitative comparison of anomaly responses obtained from Top-layer, original Bottom-layer, and projected Bottom-layer features. The highlighted regions show false activations on benign image content.

matching residuals and cause false activations. As shown in Figure 5, the original Bottom-layer representation produces strong responses in several benign regions. After applying the learned projection matrix, these false activations are effectively suppressed while the responses within the actual patch regions are retained. This result demonstrates that projection learning preserves patch-relevant discriminative information while reducing interference from complex benign content, thereby increasing the residual separation between benign and adversarial regions and improving localization accuracy.

Localization threshold and projection dimension. Figure 6 analyzes the effects of the localization threshold τ and projection dimension d. A small τ tends to incorrectly mask benign regions, while an overly large τ may miss patch regions. Overall, $\tau = 0.1$ provides a good balance between the two. For the projection dimension, $d = 8$ shows large performance variations across patch types, while $d = 16$ maintains stable performance across attacks. Further increasing the dimension to 32 or 48 leads to a clear performance drop, suggesting that higher dimensions may retain redundant or non-discriminative information. We set $\tau = 0.1$ and $d = 16$.

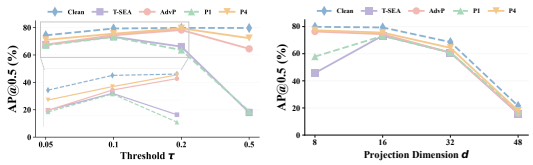

Figure 6: Ablation results for the localization threshold (left) and projection dimension (right).

5.6 Physical-World Evaluation

As shown in Figure 9, DQ-PAD effectively localizes and removes physical adversarial patches under different pedestrian poses, allowing the detector to correctly recognize the attacked pedestrians. Meanwhile, it preserves the normal detection of other objects, demonstrating its adaptability and defense effectiveness in real-world environments.

6 Conclusion

To overcome the limited generalization of existing defenses to diverse patch appearances, we propose DQ-PAD. It leverages the pretrained VQ-VAE2 codebook as benign visual prototypes to improve generalization to diverse patches, and introduces a patch-aware projection learning mechanism to enhance the discriminability between benign and adversarial features. Comprehensive experiments demonstrate that DQ-PAD achieves robust defense performance while maintaining high inference efficiency.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting portions of the text, language polishing, and improving written expression to enhance the clarity, accuracy, and readability of the paper. The use of generative AI was limited to language-related assistance, including Chinese–English translation, sentence rewriting, and the organization and refinement of selected paragraphs. We did not use generative AI tools for research methodology design, experimental design, experiment execution, experimental data generation, or result analysis, nor did we use them to generate or modify any experimental results. The methodology, model implementation, experimental settings, data processing, and conclusions presented in this paper were independently developed and verified by the authors. All text generated or revised with the assistance of generative AI was manually reviewed and modified as necessary by the authors to ensure the accuracy of its technical content, factual statements, and academic presentation. The authors take full responsibility for the final content of this paper, including all text processed with the assistance of generative AI.

ETHICS STATEMENT

This work aims to improve the robustness of object detection systems against adversarial patch attacks and is primarily intended for defensive security purposes. Although the study involves the construction of adversarial patches and adaptive attacks, these attacks are used solely to evaluate the effectiveness and robustness of the proposed defense under different threat models. The methods and experiments may involve certain dual-use risks, but this work is not intended to enhance practical attack capabilities. The experiments are mainly conducted on public datasets, and the physical-world experiments are used only to validate the effectiveness of the defense in realistic scenarios without collecting or using sensitive personal information. We comply with the usage policies of the relevant datasets and seek to avoid applications that may introduce real-world security risks throughout the experimental design and reporting process.

REPRODUCIBILITY STATEMENT

We provide the main implementation details required to reproduce DQ-PAD in the main paper and appendix, including the training and inference procedures, loss functions, hyperparameter settings, dataset splits, attack configurations, and defense configurations. The experimental section describes the datasets, adversarial patches, object detectors, baseline methods, and evaluation metrics used in our study, while the appendix further provides hyperparameter analyses, adaptive attack settings, and additional visualization results. All experiments are evaluated under fixed model and experimental configurations to ensure fair comparisons across methods. We plan to release the implementation code and related configurations upon publication to facilitate reproduction of the main experimental results.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Mauricio Byrd Victorica, György Dán, and Henrik Sandberg. Saliuitl: Ensemble salience guided recovery of adversarial patches against CNNs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Nandish Chattopadhyay, Amira Guesmi, Muhammad Abdullah Hanif, Bassem Ouni, and Muhammad Shafique. ODDR: Outlier detection & dimension reduction based defense against adversarial patches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22999–23008, 2025.

Zikang Cheng, Zhanhao Hu, Yichao Liu, Jun Li, Hang Su, and Xiaolin Hu. Full-distance evasion of pedestrian detectors in the physical world. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102366–102392, 2024.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 886–893, 2005.

Cheng He, Xingjun Ma, Bin B. Zhu, Yexin Zeng, Huan Hu, Xiang Bai, Hai Jin, and Dengpan Zhang. DorPatch: Distributed and occlusion-robust adversarial patch to evade certifiable defenses. In *Network and Distributed System Security Symposium*, 2024.

Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7848–7857, 2021.

Hao Huang, Zhaoyu Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-SEA: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20514–20523, 2023.

Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, and Cong Zou. PAD: Patch-agnostic defense against adversarial patch attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Yubo Chen, Hang Su, and Xingxing Wei. DIFFender: Diffusion-based adversarial defense against patch attacks. In *Computer Vision – ECCV 2024*, volume 15110 of *Lecture Notes in Computer Science*, pp. 130–147. Springer, 2024. doi: 10.1007/978-3-031-72943-0_8.

Vishesh Kumar and Akshay Agarwal. A unified, resilient, and explainable adversarial patch detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30387–30397, 2025.

Jiachun Li, Jianan Feng, Jianjun Huang, and Bin Liang. We can always catch you: Detecting adversarial patched objects with or without signature. *IEEE Transactions on Dependable and Secure Computing*, 22(6):7372–7388, 2025.

Zijin Lin, Yue Zhao, Kai Chen, and Jinwen He. I don't know you, but i can catch you: Real-time defense against diverse adversarial patches for object detectors. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3823–3837, 2024. doi: 10.1145/3658644.3670317.

Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14973–14982, 2022.

Bin Miao, Chao Li, Yiming Zhu, Weiming Sun, Zhen Wang, Xiaosen Wang, and Chuanbiao Xie. AdvLogo: Adversarial patch attack against object detectors based on diffusion models. *arXiv preprint arXiv:2409.07002*, 2024.

Muzammal Naseer, Salman H. Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. *arXiv preprint arXiv:1807.01216*, 2018.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

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

Wenjie Tan, Yiming Li, Chang Zhao, Zhen Liu, and Quan Pan. DOEPatch: Dynamically optimized ensemble model for adversarial patches generation. *IEEE Transactions on Information Forensics and Security*, 2024.

Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. Jedi: Entropy-based localization and removal of adversarial patches. *arXiv preprint arXiv:2304.10029*, 2023.

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Donghua Wang, Wenshuo Yao, Tingsong Jiang, Cong Li, and Xiaoqian Chen. RFLA: A stealthy reflected light adversarial attack in the physical world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4455–4465, 2023.

J. Wang, T. Ni, G. Xu, Q. Zhao, and C. Wang. Adversarial patch EXterminator: Zero-shot and patch-agnostic defense framework against adversarial patch attacks. In *Proceedings of the 35th USENIX Security Symposium*, 2026.

Hui Wei, Hao Tang, Xu Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng Wang. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9797–9817, 2024.

Siyang Wu, Jiakai Wang, Jiejie Zhao, Yazhe Wang, and Xianglong Liu. NAPGuard: Towards detecting naturalistic adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24367–24376, 2024.

Zuxuan Wu, Ser-Nam Lim, Larry S. Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pp. 1–17, 2020.

C. Xiang and P. Mittal. PatchGuard++: Efficient provable attack detection against adversarial patches, 2021.

Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium*, pp. 2237–2254. USENIX Association, 2021.

Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. Patchcleanser: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX Security Symposium*, pp. 2065–2082. USENIX Association, 2022.

Chong Xiang, Alexander Valtchanov, Saeed Mahloujifar, and Prateek Mittal. Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In *2023 IEEE Symposium on Security and Privacy*, pp. 1329–1347, 2023.

Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z. Li, and Guodong Guo. WiderPerson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22(2):380–393, 2019.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.

A Visualization Results in the Digital Domain

B Non-End-to-End Adaptive Attack Formulation

This section provides the detailed definitions of the defense-oriented losses used in the non-end-to-end adaptive attacks. Let $\mathcal{P}$ denote the latent-grid locations corresponding to the ground-truth

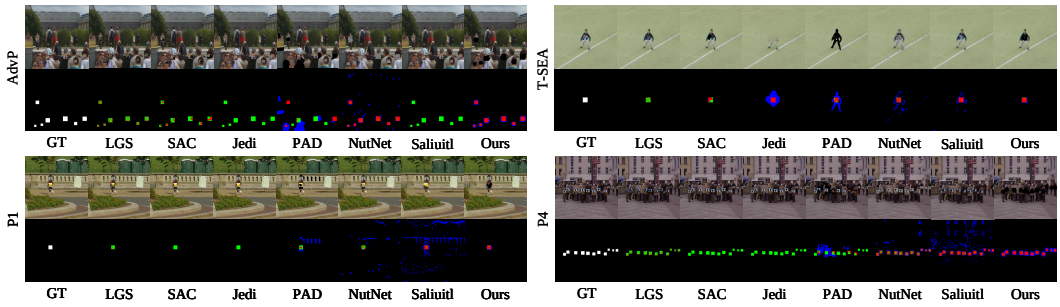

Figure 7: Visualization of defense performance and patch localization. In each subfigure, top and bottom rows show detection results on defended samples and predicted patch masks, respectively. Green: ground-truth patch regions; Red: detected patch regions; Blue: false-positive background regions.

patch region, and let $r_{\mathbf{W}}(\mathbf{z}_i)$ denote the projected codebook residual at location i. Because DQ-PAT classifies locations whose residuals exceed the threshold τ as suspicious, the threshold-evasion loss suppresses the patch-region residuals below this threshold:

$$\mathcal{L}_{\text{thr}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \max(0, r_{\mathbf{W}}(\mathbf{z}_i) - (\tau - \epsilon)), \tag{13}$$

where $\epsilon > 0$ is a margin that encourages the patch residuals to remain below and sufficiently far from the decision boundary.

The mask-inconsistency loss directly reduces the overlap between the predicted mask and the ground-truth patch region. Since the binary mask obtained by hard thresholding is non-differentiable, we approximate it using a temperature-scaled sigmoid function:

$$\widetilde{m}_i = \sigma\left(\frac{r_{\mathbf{W}}(\mathbf{z}_i) - \tau}{T}\right), \tag{14}$$

where T is the temperature parameter. Let m_i^{gt} denote the ground-truth patch mask. The mask-inconsistency loss is defined as the soft intersection-over-union between the predicted soft mask and the ground-truth mask:

$$\mathcal{L}_{\text{mask}} = \frac{\sum_i \widetilde{m}_i m_i^{\text{gt}}}{\sum_i \left(\widetilde{m}_i + m_i^{\text{gt}} - \widetilde{m}_i m_i^{\text{gt}}\right)}. \tag{15}$$

Minimizing this soft IoU encourages the predicted mask to deviate from the actual patch region, thereby weakening the spatial coverage of DQ-PAT. Finally, the joint attack simultaneously optimizes both evasion objectives:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{thr}} + \mathcal{L}_{\text{mask}}. \tag{16}$$

The joint objective therefore attacks the localization mechanism by suppressing the anomaly residuals within the patch region while reducing the overlap between the predicted and ground-truth masks.

C Robustness Analysis

Patch with Transformations. To evaluate the adaptability of DQ-PAD to different patch configurations, we apply random scale transformations to the patches and place them randomly within the target bounding boxes. The scaling ratio ranges from 0.2 to 0.4, and each image contains three to five different types of adversarial patches. As shown in Figure 8 (left), under these intensive attack settings involving variations in patch scale, location, number, and type, DQ-PAD still achieves an AP of approximately 80%, demonstrating strong generalization to complex patch transformations.

Varying Pedestrian Scenarios. To further evaluate the applicability of DQ-PAD in practical pedestrian-detection environments, we conduct experiments on the relatively sparse CCTVPerson

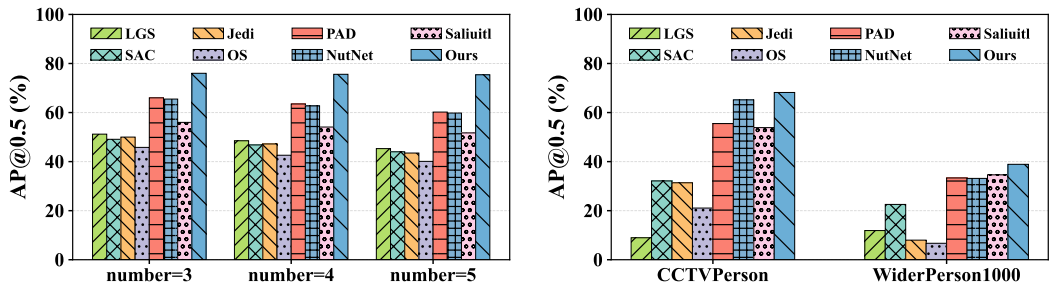

Figure 8: Defense performance under different evaluation settings.

dataset and the densely populated WiderPerson1000 dataset. As shown in Figure 8 (right), all methods exhibit varying degrees of performance degradation in dense pedestrian scenes, possibly because severe occlusion and complex background interference increase the difficulty of identifying adversarial regions. Nevertheless, DQ-PAD achieves the best results in both scenarios, demonstrating stable defense performance across different crowd densities and complex visual environments.

D VISUALIZATION RESULTS IN THE PHYSICAL WORLD



Figure 9: Qualitative defense results of DQ-PAD under different pedestrian poses in physical-world scenarios.