

FD-Square-Det: Frequency-Guided Robust Pedestrian Detection Under Adversarial and Cross-Domain Conditions

Prithiraj Modak*, Monjurul Islam Nimon, Tanvir Ahmed, Rifath Mahmud, Syeda Anika Tasnim
Department of Computer Science, American International University, Bangladesh, Dhaka, Bangladesh

Abstract—Pedestrian detectors face physical adversarial patches and distribution shift between training and deployment data. These threats are usually studied separately. This study asks whether robustness to patches and cross-domain generalization can improve together or conflict. Detectors based on You Only Look Once version 8 medium (YOLOv8m) are evaluated under a four-quadrant protocol with clean and patched images from PASCAL Visual Object Classes (VOC) and from WiderPerson, an unlabeled target domain. Scores are average precision at an intersection over union threshold of 0.5 (AP50) in the style of Common Objects in Context (COCO), with three seeds, equal training budgets, and paired statistics. The Frequency-guided Domain-aware Detector (FD-square-Det) combines a learnable frequency filter, a frequency-statistic channel attention, a gradient reversal domain branch, and patch-augmented training. Each component is audited against matched baselines. Under a universal 128-pixel patch crafted on a surrogate detector, patch training alone reduces the relative AP50 drop on VOC from 14.9 percent to 1.6 percent without changing clean AP50. The domain branch lowers clean target AP50 by about 5.6 points, and the frequency modules give no gain beyond seed noise. White-box patches optimized for each detector remain effective, with attack success between 61 and 85 percent. An explainable artificial intelligence analysis with patch-aware Gradient-weighted Class Activation Mapping (Grad-CAM) shows no better pedestrian localization for FD-square-Det. The two goals therefore need not conflict, and the tested adaptation and frequency components do not link them.

Keywords—Adversarial patch; cross-domain generalization; domain adaptation; explainable artificial intelligence; frequency-domain learning; pedestrian detection; YOLOv8

I. INTRODUCTION

Pedestrian detection underlies safety-critical systems, including collision avoidance in autonomous vehicles, video surveillance, and intelligent transportation. Modern detectors substantially outperform the handcrafted-feature methods surveyed over a decade ago [1]. Single-stage architectures such as You Only Look Once (YOLO) offer real-time speed with only a small loss in localization precision [2]. This progress is mostly measured under a closed-world assumption. Deployment images are assumed to resemble training images, and no adversary is assumed to be present. Neither assumption holds reliably in practice.

Pedestrian detectors are vulnerable to adversarial patches, which are small, printable regions of optimized texture designed to suppress detection confidence. This vulnerability has

been demonstrated directly against a person detector [3]. The finding builds on earlier evidence that physically realizable perturbations remain effective under changes in viewpoint, lighting, and distance, rather than existing only as a digital phenomenon [4]. Patches are generated through gradient-based optimization. They can transfer across viewpoints and, to a lesser extent, across detector architectures.

A detector trained on one dataset frequently fails on another dataset, even for the same object category. This effect is documented as dataset bias [5]. Differences in viewpoint, illumination, background, and pedestrian scale introduce a covariate shift that convolutional features do not discount automatically. Domain-adaptive detectors reduce this problem by aligning source and target feature representations through adversarial learning with a gradient reversal layer (GRL) [6]. This strategy is widely adopted for cross-domain object detection.

Adversarial robustness and cross-domain generalization are mostly studied in isolation, as Section II shows. Convolutional neural networks have been shown to rely heavily on high-frequency image information that is not easily perceived by humans. This sensitivity has been linked to both adversarial vulnerability and reduced generalization across visually different datasets [7]. It is therefore plausible that the two problems share a common origin in the frequency spectrum of learned features. If this holds, a single frequency-domain intervention could address both problems.

This study treats that idea as a question and not as an assumption. Do robustness to patches and cross-domain generalization conflict, reinforce each other, or remain independent? To answer it, the Frequency-guided Domain-aware Detector (FD²-Det) is studied. FD²-Det is a pedestrian detector based on the medium variant of YOLO version 8 (YOLOv8m), with three lightweight components added to patch-augmented training. A learnable frequency filter suppresses unstable high-frequency responses. A channel attention module reweights feature channels using a low-frequency statistic. A domain-adversarial branch with a GRL aims to reduce the source-target distribution gap without target labels. Each component is audited separately against matched baselines.

FD²-Det is evaluated under a four-quadrant protocol that crosses clean and adversarial conditions with in-domain and cross-domain settings. The four test conditions are clean source images (Q1), patched source images (Q2), clean target images (Q3), and patched target images (Q4). PASCAL Visual Object

*Corresponding author.

Classes (VOC) [8] serves as the labeled source domain, and WiderPerson [9] serves as the unlabeled target domain. The patch threat model follows earlier work on adversarial patches [3], [4]. Detection quality is measured by average precision (AP) at an intersection over union (IoU) threshold of 0.5 (AP50), computed in the style of Common Objects in Context (COCO) [10]. Gradient-weighted Class Activation Mapping (Grad-CAM) [11] supports an Attention Activation Score (AAS) that measures where the detector directs its attention.

The scope of the study is limited. One detector family (YOLOv8m), two datasets, and one family of attacks (universal patches placed on the torso) are considered. The conclusions apply to this setting only.

The main contributions of this study are summarized as follows:

- 1) A controlled, equal-budget audit of FD²-Det is presented. Matched, patch-only, and patch plus GRL baselines are compared over three seeds with COCO-style metrics, paired *t*-tests, and image-level bootstrap intervals. The measured cost of the modules is small, at 1.0% more parameters and about 5% more latency.
- 2) The audit shows that patch augmentation gives black-box patch robustness without a measurable cost in clean or cross-domain accuracy. The GRL branch lowers accuracy, and the frequency components show no measurable gain at this scale. These findings answer the research question for the tested setting.
- 3) A four-quadrant protocol is applied to clean and patched images on both domains. It is supported by random-patch controls, patch sizes of 80, 128, and 160 pixels, targeted attack success, and white-box adaptive patches.
- 4) The AAS is extended with two patch-aware variants. They remove the patch region and measure the share of attention that falls on the patch. The analysis shows no gain in pedestrian localization for FD²-Det.

The remainder of this study is organized as follows. Section II reviews related work. Section III describes the methodology. Section IV presents the experimental setup. Section V reports the results, and Section VI presents the explainability analysis. Section VII discusses the findings and limitations, and Section VIII concludes the study.

II. RELATED WORK

A. Pedestrian Detection Architectures

Recent pedestrian detectors pursue occlusion robustness through attention and transformer components added to convolutional backbones [12]–[14]. A second line of work targets parameter efficiency using lighter backbones and lightweight attention modules [15]–[17]. A third line targets low-latency inference through keypoint-based and attention-augmented detection heads, deployed on unmanned aerial vehicle, intelligent-transport, and general-purpose benchmarks [18]–[22]. Across these eleven studies, evaluation is limited to clean, in-distribution accuracy.

B. Adversarial Attacks on Object Detection

A separate body of work examines how detectors behave under attack. Transferability studies show that adversarial examples crafted for one network can fool another [23]. A survey of intrusion detection notes that attacks designed for vision are often reused in other domains [24]. In contrast, detector-specific attacks directly perturb backbone features and detection heads to reduce detection accuracy [25], [26]. Printable adversarial patches bridge digital and physical attack settings, ranging from hood-mounted and vehicle-camouflage patches to distance-adaptive clothing patches [27]–[29]. Related physical attacks have also been developed for light detection and ranging (LiDAR) point clouds [30]. Defenses against patches include local gradient smoothing [31] and patch segmentation and removal for object detectors [32]. A lightweight reconstruction network based on a generative adversarial network defends traffic-sign recognition [33]. A separate, non-adversarial comparison shows that the YOLO family retains more of its accuracy than Faster R-CNN under domain transfer [34]. Consequently, it remains unclear whether adversarial attacks or corresponding defenses remain effective when detectors are deployed in previously unseen domains.

C. Physical Adversarial Robustness in Autonomous Driving

Beyond digital attacks, recent studies have explored physical adversarial attacks that remain effective in real-world environments. Person-targeted physical patches include deformable and consensus-localized infrared designs, together with a unified visible-infrared cross-modal patch [35]–[37]. A projector-based traffic-sign attack further contributes a reusable physical-validation protocol [38]. Two studies evaluate both adversarial robustness and domain shift. Neither focuses on object detection or frequency-domain feature learning. One validates a limited-label robustness measure on a KITTI-to-nuScenes shift [39]. The other evaluates a robust-policy and saliency-detector framework under adversarial attacks and a cloudy-to-rain weather shift [40]. Recent surveys review physical adversarial attacks and robust statistical methods [41], [42]. Joint evaluation of adversarial robustness and domain shift remains limited, particularly for pedestrian detection.

D. Domain-Adaptive Object Detection

The studies below address domain shift directly through architectural adaptation. Unsupervised domain-adaptive detectors generally rely on GRL-based feature alignment [6], [43]. Variants include dynamic masking for weather shift [44], uncertainty-guided pseudo-labeling combined with adversarial tiling [45], energy-based curriculum learning [46], and bin-level alignment for localization regression [47]. All four reviewed studies focus on domain adaptation and not on adversarial robustness. In this context, “adversarial” denotes GRL-based domain-adversarial training and not defense against adversarial attacks.

E. Frequency-Domain Learning

Frequency-domain processing appears across the literature in two largely separate roles. As an attack surface, spectrum-focused and frequency-selective perturbations target modulation and signal classifiers [48], [49]. As defense strategies, Fast

Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT)-paired training has been shown to improve both white-box and black-box robustness [50]. Homomorphic filtering has also demonstrated strong defensive performance against adversarial attacks [51]. Beyond adversarial defense, frequency-domain processing has been used to improve generalization across diverse computer vision tasks. These include blur detection, illumination compensation, image fusion, hyperspectral image classification, super-resolution, and other vision applications [52]–[56]. A radio-fingerprinting study reports a robustness-accuracy trade-off for adversarial training [57].

Frequency operations have also been placed inside network layers. Global filter layers replace self-attention with learnable filters applied in the Fourier domain [58]. Channel attention has been generalized by using discrete cosine transform (DCT) components as channel descriptors [59]. These designs are close to the modules audited in this study. Within the reviewed literature, frequency-domain methods have received limited attention in object and pedestrian detection, especially as a combined solution for adversarial defense and cross-domain generalization.

F. Explainability in Detection

A final thread examines how detector decisions can be interpreted and not only measured. Explainability research includes both post hoc and intrinsic approaches. Post hoc methods have been integrated with attention-augmented YOLOv5 and federated YOLOv7 pipelines, whereas intrinsic methods rely on prototype-based reasoning and causality-erasing masking for SHapley Additive exPlanations (SHAP) [60]–[63]. Among the reviewed studies, one targets pedestrian detection specifically. Its evaluation is limited to a single dataset and does not include adversarial robustness or cross-domain generalization.

G. Summary of Limitations

The central limitation identified within each thread is summarized in Table I. Across the reviewed literature, the joint evaluation of adversarial robustness and cross-domain generalization remains limited for pedestrian detection. Existing studies typically address adversarial defense, domain adaptation, or frequency-domain learning independently. It therefore remains unclear whether improving robustness against physical patches also influences generalization across domains. This study examines the question in a controlled setting. FD²-Det is evaluated under a four-quadrant protocol on both the source and the target domain.

III. METHODOLOGY

This section describes FD²-Det, the patch attack formulation, the training objective, the attention score, and the asymptotic cost. The overall design is shown in Fig. 1.

A. Problem Formulation

Two factors reduce the reliability of pedestrian detectors. First, adversarial patches placed on the body or clothing of a pedestrian reduce detection confidence [3]. Second, a model trained on one dataset often fails on another [5]. The causes

TABLE I. SUMMARY OF LIMITATIONS ACROSS REVIEWED RESEARCH THREADS

Thread	Representative Works	Main Limitation
Pedestrian detection architectures	[12]–[22]	Only clean, in-distribution evaluation
Adversarial attacks on detection	[23]–[34]	No joint evaluation of attacks and domain shift
Physical robustness in driving	[35]–[42]	Only two joint studies, neither on pedestrian detection
Domain-adaptive detection	[44]–[47]	Excludes adversarial robustness
Frequency-domain learning	[48]–[59]	Limited use in detection; defense and generalization not combined
Explainability in detection	[60]–[63]	Single-dataset evaluation only

are differences in scene, appearance, illumination, viewpoint, and background.

The person images of PASCAL VOC form the labeled source domain. WiderPerson forms the unlabeled target domain under the unsupervised domain adaptation (UDA) setting. The pedestrian detector is built on the YOLOv8m architecture [64] and is denoted by $f_\theta(\cdot)$, where θ represents the learnable model parameters. Given an input image $\mathbf{x}$, the detector predicts a set of pedestrian bounding boxes with confidence scores, $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$.

To evaluate adversarial robustness, a universal patch δ is applied to the torso of the largest annotated pedestrian by the placement operation $\mathcal{P}(\cdot)$. This produces the patched image $\mathbf{x}^{\text{adv}} = \mathcal{P}(\mathbf{x}, \delta)$. The patch is optimized by projected gradient steps in pixel space. Adam is the optimizer, and pixel values are kept in $[0, 1]$. This follows the idea of the projected gradient descent formulation [65]. The objective is to suppress the person evidence of the patched pedestrian, as given in Eq. (1):

$$\delta^* = \arg \min_{\delta \in [0,1]^{s \times s \times 3}} \mathbb{E}_{\mathbf{x} \sim \mathcal{B}, \tau \sim \mathcal{T}} \left[\frac{1}{K} \sum_{i \in \mathcal{A}_K} \ell_i(\mathcal{P}(\mathbf{x}, \tau(\delta))) \right], \quad (1)$$

where, ℓ_i is the pre-sigmoid person logit of anchor i . The set $\mathcal{A}_K$ contains the $K = 5$ anchors with the highest logits among those inside the box of the patched pedestrian. The set $\mathcal{B}$ is a mini-batch of source-domain training images, and s is the patch size. The term τ is a random rotation and rescaling of the patch drawn from $\mathcal{T}$, as described in Section IV-B.

A deployable detector must be accurate on the source domain, robust against patches, and reliable on the target domain. FD²-Det combines patch-augmented training, frequency-domain feature refinement, and domain-adversarial alignment to meet these requirements. The study tests whether it does.

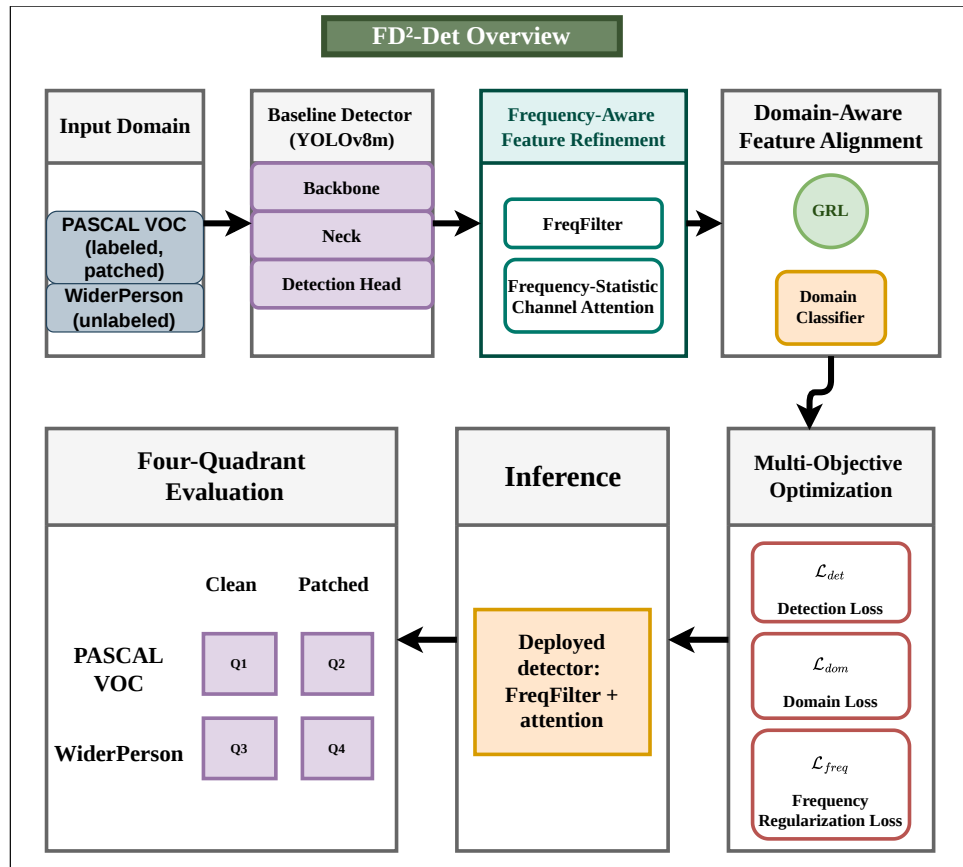


Fig. 1. Overview of FD²-Det. The framework extends YOLOv8m with Frequency-Aware Feature Refinement, Domain-Aware Feature Alignment, and multi-objective training. Training uses clean and patched source-domain images with unlabeled target-domain images, and evaluation follows the four-quadrant protocol.

B. Theoretical Motivation and Hypothesis

Adversarial patches occupy only a small image region. They can still produce strong responses that disturb the learned feature maps. Domain shift arises from differences in illumination, background, viewpoint, scale, and scene composition between datasets. Both effects are often described as changes in the frequency content of the input [7]. Adversarial patches carry texture with strong high-frequency components and little semantic content. Domain shift alters both low-frequency and high-frequency components because of different imaging conditions. A detector that relies on unstable frequency responses may therefore be vulnerable to both.

This reasoning leads to a hypothesis and not to a conclusion. Suppose that robustness to patches and cross-domain generalization share an instability in the frequency spectrum of learned features. Then improving the stability of that spectrum should help both. The hypothesis makes a testable prediction. Modules that stabilize frequency content should improve patched and cross-domain accuracy without reducing clean accuracy. FD²-Det turns the hypothesis into three components that operate on the neck features of the detector. Section V checks the prediction experimentally, and Section VII-A discusses the outcome.

C. Overall FD²-Det Framework

FD²-Det combines frequency-aware feature refinement with domain adaptation, as shown in Fig. 1. The detailed architecture is shown in Fig. 2. The framework adopts YOLOv8m [64] as the baseline detector because of its balance between accuracy and computational efficiency. The backbone architecture is not modified. Lightweight refinement modules are inserted at the feature aggregation stage. This preserves the original detection capability.

Training uses labeled VOC images and unlabeled WiderPerson images. A subset of the source batches is augmented with adversarial patches. Three modules operate on the multi-scale YOLOv8m neck features. FreqFilter suppresses high-frequency responses. The channel attention module emphasizes channels with a strong low-frequency statistic. The Domain-Aware Feature Alignment branch uses a GRL and a lightweight domain classifier to reduce the source-target gap without target labels. The domain classifier and the GRL exist only during training. The deployed detector keeps FreqFilter and the channel attention. The components are optimized jointly with one multi-objective loss, which is given in Section III-G. The next subsections describe each module in turn.

D. FreqFilter Module

FreqFilter applies the reasoning of Section III-B to the feature maps of the detector. It reduces high-frequency responses and keeps the low-frequency content. Unlike conventional image preprocessing, the module works on the intermediate feature maps of the YOLOv8m neck and not on the input image. It is applied to the P3, P4, and P5 feature pyramid outputs, as shown in Fig. 2. Features are therefore refined at three scales, and the backbone architecture stays unchanged.

FreqFilter has three steps. The feature map is moved to the frequency domain, a low-pass mask removes the high frequencies, and the result is blended with the original feature map. Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, where C , H , and W denote the number of channels, height, and width, the feature map is transformed into the frequency domain using the two-dimensional real Fast Fourier Transform (RFFT):

$$\hat{\mathbf{F}} = \mathcal{F}_{\text{rfft2}}(\mathbf{F}).$$

A binary low-pass mask is constructed in the normalized frequency domain. Let u and v denote the normalized horizontal and vertical frequency coordinates obtained with `fftfreq` and `rfftfreq`. The mask is defined in Eq. (2):

$$\Lambda(u, v) = \begin{cases} 1, & \sqrt{u^2 + v^2} \leq r, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where, r is the cutoff radius and is fixed at 0.4. The normalized coordinates lie between -0.5 and 0.5 , so this radius keeps about half of all spectral coefficients. It removes mainly the highest frequencies in the diagonal directions. The frequency representation is masked element-wise,

$$\hat{\mathbf{F}}_{\text{LP}} = \hat{\mathbf{F}} \odot \Lambda,$$

and the filtered spatial feature map is reconstructed with the inverse real Fast Fourier Transform:

$$\mathbf{F}_{\text{LP}} = \mathcal{F}_{\text{irfft2}}^{-1}(\hat{\mathbf{F}}_{\text{LP}}).$$

The filtered features do not replace the original features. A learnable residual blend is used instead, as given in Eq. (3):

$$\mathbf{F}' = \mathbf{F} + \alpha (\mathbf{F}_{\text{LP}} - \mathbf{F}), \quad (3)$$

where, α is a learnable scalar that controls the filtering strength. When $\alpha = 0$, the module reduces to an identity mapping. A separate scalar with initial value 0.1 is assigned to each of the P3, P4, and P5 feature maps. This allows the filtering strength to differ across levels, and the module adds only three learnable parameters. The blended features $\mathbf{F}'$ are the input of the channel attention module. The same mask also defines the frequency regularization loss in Section III-G.

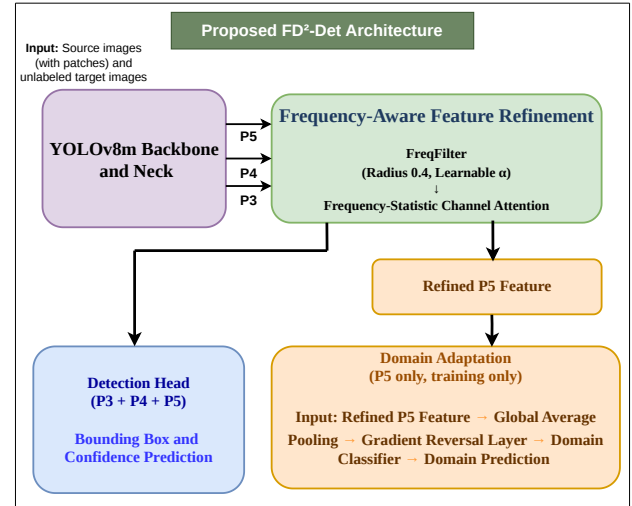


Fig. 2. FD²-Det architecture. FreqFilter and the frequency-statistic channel attention refine P3, P4, and P5. Only the P5 neck feature feeds the domain-aware feature alignment branch, which is used during training only.

E. Frequency-Statistic Channel Attention

The refined feature maps still contain channels that contribute unequally to the pedestrian representation. A channel attention module recalibrates the channels using a low-frequency statistic. It follows FreqFilter at each pyramid level. The design is related to frequency channel attention, which uses DCT components as channel descriptors [59], and to learnable spectral filtering [58].

The implemented computation is a spatial block statistic and not a discrete cosine transform [66]. Let $\mathbf{F}' \in \mathbb{R}^{C \times H \times W}$ be the output of FreqFilter. Each channel is zero-padded to a multiple of 8 and divided into non-overlapping 8×8 blocks. The descriptor of channel c is the mean absolute activation over the top-left 4×4 positions of all blocks, as given in Eq. (4):

$$s_c = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} |F'^{(c)}(i,j)|, \quad (4)$$

where, $\Omega = \{(i,j) : (i \bmod 8) < 4, (j \bmod 8) < 4\}$. The statistic is a rough proxy for low-frequency content, and this was not verified separately.

The descriptors form the vector $\mathbf{s} = [s_1, \dots, s_C]^T$. A two-layer bottleneck converts it into channel weights, as given in Eq. (5):

$$\mathbf{a} = \sigma(W_2 \text{ReLU}(W_1 \mathbf{s})), \quad (5)$$

where, $W_1 \in \mathbb{R}^{h \times C}$ and $W_2 \in \mathbb{R}^{C \times h}$ are learnable matrices with $h = \max(C/4, 8)$. The function $\text{ReLU}(\cdot)$ is the rectified linear unit, $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{a} \in [0, 1]^C$. The refined feature map is $\mathbf{F}_{\text{ATT}}^{(c)} = a_c \mathbf{F}'^{(c)}$. The module is applied at P3, P4, and P5 with 192, 384, and 576 channels.

F. Domain-Aware Feature Alignment

Frequency-aware refinement does not remove the effect of domain shift. A Domain-Aware Feature Alignment branch is therefore attached to the P5 feature map of the YOLOv8m neck, as shown in Fig. 2. P5 is selected because it carries rich semantic information with low spatial redundancy. The branch reads the P5 feature before FreqFilter and the attention, so the frequency modules are not part of the alignment path. Let $\mathbf{F}_{P5} \in \mathbb{R}^{C \times H \times W}$ denote this P5 feature map. Global Average Pooling (GAP) aggregates it into a compact descriptor,

$$\mathbf{z} = \text{GAP}(\mathbf{F}_{P5}), \quad \mathbf{z} \in \mathbb{R}^C.$$

The descriptor passes through the GRL [43] before it enters the domain classifier. The GRL acts as an identity mapping in the forward pass and reverses the gradient in the backward pass. The reversal coefficient increases gradually during training, following [43], as given in Eq. (6):

$$\begin{aligned} \mathbf{z}_{\text{GRL}} &= \mathbf{z}, & \frac{\partial \mathcal{L}}{\partial \mathbf{z}} &= -\lambda \frac{\partial \mathcal{L}}{\partial \mathbf{z}_{\text{GRL}}}, \\ \lambda(p) &= \frac{2}{1 + \exp(-10p)} - 1, \end{aligned} \quad (6)$$

where, λ is the gradient reversal coefficient and $p \in [0, 1]$ denotes the normalized training progress. The GRL output is fed to a lightweight domain classifier,

$$\mathbf{p}_d = \text{Softmax}(f_d(\mathbf{z}_{\text{GRL}})),$$

which predicts the source or target domain and is trained with the cross-entropy loss in Eq. (7):

$$\mathcal{L}_{\text{dom}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^2 y_{ic} \log(p_{ic}), \quad (7)$$

where, N is the number of images in the combined source and target batch, y_{ic} is the one-hot domain label (source 0, target 1), and p_{ic} is the predicted probability of domain c . The classifier has the layers $576 \rightarrow 256 \rightarrow 64 \rightarrow 2$ with dropout of 0.5.

The detector is intended to learn domain-invariant features through the reversed gradient without target annotations. The target images also pass through the shared network, so they update the batch-normalization statistics as well.

G. Multi-Objective Training

FD²-Det jointly optimizes pedestrian detection, domain adaptation, and frequency-aware feature refinement in an end-to-end framework. The procedure is summarized in Algorithm 1. Each iteration samples labeled source images and unlabeled target images. A subset of the source batches is patched with textures from a fixed bank of training patches, and the remaining batches stay clean.

Algorithm 1 Multi-Objective Training of FD²-Det

Require: Labeled source images $\mathbf{X}_s$ with labels $\mathbf{Y}_s$, unlabeled target images $\mathbf{X}_t$, bank of training patches $\mathcal{D}_{\text{tr}}$, starting checkpoint

Ensure: Trained FD²-Det detector

- 1: Load the starting checkpoint and add FreqFilter, the channel attention, and the domain branch.
- 2: **for** each epoch **do**
- 3: **for** each iteration **do**
- 4: Sample a source batch (8 images) and a target batch (4 images).
- 5: With probability 0.5 for the whole batch, paste one patch drawn from $\mathcal{D}_{\text{tr}}$ on the largest pedestrian of each source image.
- 6: Compute $\mathcal{L}_{\text{det}}$ on the source batch.
- 7: Compute $\mathcal{L}_{\text{dom}}$ on the pooled P5 descriptors of both batches.
- 8: Compute $\mathcal{L}_{\text{freq}}$ and $\mathcal{L}_{\text{total}}$.
- 9: Update the parameters by backpropagation.
- 10: **end for**
- 11: **end for**
- 12: **return** the weights of the final epoch.

The detection loss $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dff}}$ is the standard YOLO loss. It sums the box-regression, classification, and Distribution Focal Loss (DFL) terms. The domain classifier uses the cross-entropy loss of Eq. (7). A frequency regularization loss penalizes the residual high-frequency energy that remains after FreqFilter at the P5 level, computed outside the low-pass mask Λ of Eq. (2), as given in Eq. (8):

$$\mathcal{L}_{\text{freq}} = \frac{1}{BCHW} \sum_{b=1}^B \sum_{c=1}^C \sum_{u,v} \left| \hat{\mathbf{F}}'_{b,c}(u,v) \right|^2 (1 - \Lambda(u,v)), \quad (8)$$

where, $\hat{\mathbf{F}}'_{b,c}$ is the RFFT of channel c of the blended P5 feature $\mathbf{F}'$ of Eq. (3) for image b of the source batch. The terms B , C , H , and W are the batch size, the number of channels, and the height and width of that feature. The overall objective is given in Eq. (9):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_{\text{dom}} \mathcal{L}_{\text{dom}} + \lambda_{\text{freq}} \mathcal{L}_{\text{freq}}, \quad (9)$$

with $\lambda_{\text{dom}} = 0.1$ and $\lambda_{\text{freq}} = 0.01$ fixed across all experiments. All modules share the backbone features and are optimized through a single backpropagation pass.

H. Attention Activation Score

Detection metrics such as AP do not show where the detector focuses. The AAS, shown in Fig. 3, measures the share of Grad-CAM activation [11] that falls inside the ground-truth pedestrian boxes. Grad-CAM is computed in evaluation mode on the deployed network, with FreqFilter and the attention active. The map is taken at the neck level (P3, P4, or P5) that produced the top-scoring detection, and the score is the highest person confidence in the image. A map is computed only if this confidence is at least 0.05.

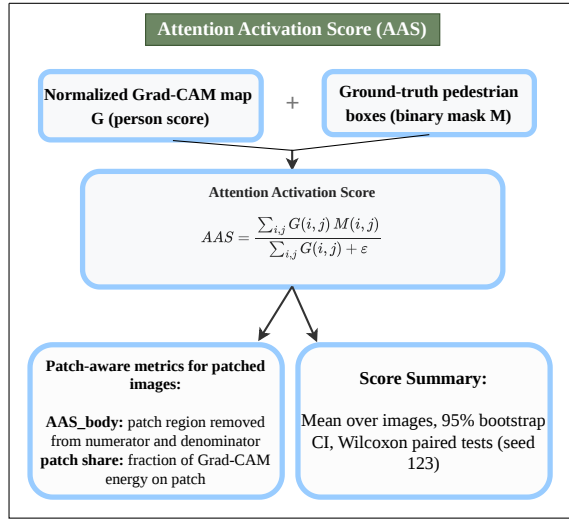


Fig. 3. Attention Activation Score (AAS) and its patch-aware variants. The score is the share of normalized Grad-CAM energy that falls inside the ground-truth pedestrian boxes.

Let $\mathbf{G} \in \mathbb{R}^{H \times W}$ denote the normalized Grad-CAM map and $\mathbf{M} \in \{0, 1\}^{H \times W}$ the binary mask of the ground-truth pedestrian boxes, where $M(i, j) = 1$ inside a pedestrian box. The AAS is the fraction of activation energy inside the mask, as given in Eq. (10):

$$AAS = \frac{\sum_{i=1}^H \sum_{j=1}^W G(i, j) M(i, j)}{\sum_{i=1}^H \sum_{j=1}^W G(i, j) + \varepsilon}, \quad (10)$$

where, ε ensures numerical stability and $AAS \in [0, 1]$. Higher values indicate attention concentrated on pedestrians.

The patch lies inside the pedestrian box, so attention on the patch also raises the score. Two patch-aware variants are therefore used on patched images. In AAS_{body} , the patch region is removed from both the numerator and the denominator. The patch share is the fraction of the Grad-CAM energy that lies on the patch. The AAS is computed for each image and summarized by the mean and 95% bootstrap confidence intervals.

I. Computational Complexity

FD²-Det adds three groups of operations. FreqFilter requires an FFT-based refinement of cost $\mathcal{O}(HW \log(HW))$ per channel at the P3, P4, and P5 maps. The channel attention costs $\mathcal{O}(CHW)$. The alignment branch costs $\mathcal{O}(Cd)$ on the pooled P5 feature, where d is the hidden width, and it is used during training only. The overall complexity equals that of YOLOv8m plus these terms. The measured parameters, operations, latency, and training time are reported in Section V-H.

IV. EXPERIMENTAL SETUP

This section describes the datasets, implementation details, threat model, evaluation protocol, and baselines used to assess

TABLE II. OVERVIEW OF THE DATASETS USED IN THIS STUDY

Property	PASCAL VOC	WiderPerson
Source	[8]	[9]
Domain role	Labeled source domain	Unlabeled target domain (UDA)
Training images	1,948 (VOC2007 trainval, with persons)	900 unlabeled (sampled from the training split)
Evaluation images	1,962 (VOC2007 test)	900 labeled (validation split)
Retained class	Person (difficult excluded, $\geq 32 \times 32$ px)	Pedestrian
Resolution	640×640 (resized)	640×640 (resized)
Key characteristics	Clean annotations, moderate scene complexity	Severe occlusion, variable scale, dense crowds, urban scenes
Used for	Training and evaluation (Q1, Q2)	Adaptation and evaluation (Q3, Q4)

FD²-Det.

A. Datasets

Two public datasets are used, PASCAL VOC [8] and WiderPerson [9]. They differ in scene composition, pedestrian appearance, and background, which suits a study of domain shift. Both datasets and their roles are summarized in Table II.

PASCAL VOC is the labeled source domain. Only the person class is kept, and objects marked as difficult are excluded. Boxes with an area below 32×32 pixels in the original image are removed. The starting checkpoint was trained with the Ultralytics pipeline on the person images of VOC2007 and VOC2012 trainval and validated on VOC2007 test. All fine-tuned detectors use the VOC2007 trainval images that contain at least one person for training. The VOC2007 test images are used for Q1 and Q2.

WiderPerson is the unlabeled target domain. The copy used here is in single-class YOLO format and has 7,199 training, 900 validation, and 900 test images. The adaptation set consists of 900 unlabeled images sampled from the training split with seed 0. The evaluation set consists of the 900 labeled validation images. The two sets are disjoint, and no target labels are used in training. The evaluation images contain 27,063 annotated pedestrians, which is about 30 pedestrians per image. All images are resized to 640×640 pixels, and the boxes are rescaled accordingly. The data loaders output ImageNet-normalized tensors. The FD²-Det wrapper reverses this normalization internally. All detectors therefore receive pixel values in $[0, 1]$, as the starting weights expect.

The main comparison is repeated with the seeds 42, 123, and 777. Results are reported as mean and sample standard deviation (SD) over the three runs. The ablation, the attention analysis, and the adaptive attacks use one seed, as stated where they are reported.

Each iteration uses 8 source images (244 iterations per epoch), and 4 unlabeled target images are added when the domain branch is active. The backbone is optimized with stochastic gradient descent (SGD) with a learning rate of 10^{-3} ,

momentum of 0.937, weight decay of 5×10^{-4} , and Nesterov momentum. The auxiliary modules use Adam at a learning rate of 10^{-3} . The learning rate follows cosine annealing down to 10^{-5} , and the gradient norm is clipped at 10. Augmentation consists of a random horizontal flip, applied to the image and the boxes together, and color jitter. The patch augmentation of Section III-G is added. The weights of the final epoch are used, and no model selection is performed on target data.

B. Threat Model

This study considers universal adversarial patches as the primary threat. A patch is a localized perturbation that misleads the detector and alters only a small image region. Such attacks can be realized as printed stickers or wearable patterns. The patch has a size of 128×128 pixels, which is 20% of the image width. It is pasted on the torso of the largest annotated pedestrian. The top-left corner is placed at $(c_x W - s/2, (c_y - 0.1 b_h)H - s/2)$ and clamped to the image, where (c_x, c_y) is the normalized box center, b_h is the normalized box height, and s is the patch size.

The patch is optimized with Eq. (1) using Adam with a step size of $8/255$ on raw pixel values in $[0, 1]$. Optimization runs for 400 iterations, and each iteration uses a fresh batch of 8 training images. The random transformation τ is a rotation of $\pm 15^\circ$ and a scaling between 0.9 and 1.1, applied during optimization only. Three patch sets are used:

- 1) Training patches: 8 patches (seeds 0 to 7) crafted on the stock YOLOv8m. They are used for the patch augmentation.
- 2) Black-box test patches: 5 patches (seeds 1000 to 1004), also crafted on the stock YOLOv8m and disjoint from the training patches. The attacker has no access to the weights of the evaluated detectors. The stock detector and the evaluated detectors share the same pretrained initialization, so the black-box setting is partial. These patches define Q2 and Q4 and are cycled over batches.
- 3) White-box adaptive patches: 3 patches for each evaluated model (seeds 2000 to 2002), optimized directly against that model. They are used for the seed-123 models only.

A uniform random-noise patch of the same size serves as a control. A size sweep uses patches of 80 and 160 pixels (3 patches each, seeds 1000 to 1002). At test time, patches are pasted without jitter. The training and test patches come from the same generator, which is a limitation discussed in Section VII-C.

C. Evaluation Protocol

FD²-Det is evaluated with a protocol that assesses detection accuracy, adversarial robustness, and cross-domain generalization within one setting, as shown in Fig. 4. Each detector is trained once and evaluated on clean and patched images from the source domain (Q1 and Q2) and from the target domain (Q3 and Q4).

The comparison of Q1 and Q2 measures robustness against patches within the source domain. The comparison of Q1 and Q3 measures cross-domain generalization. Q4 combines

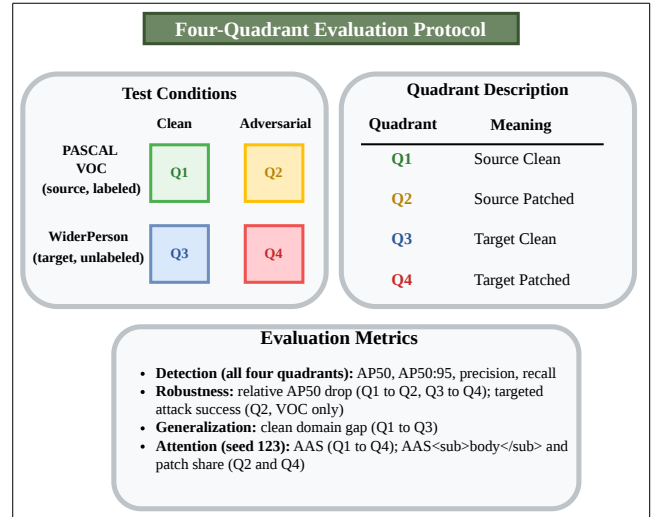


Fig. 4. Four-quadrant evaluation protocol. Detectors are evaluated on clean and patched images from the source domain (PASCAL VOC) and the target domain (WiderPerson). The metric groups list the quantities computed for each comparison.

domain shift and patch attack and is the most demanding scenario.

Detection quality is measured by AP50 with 101-point COCO interpolation. Detections are kept down to a confidence of 0.001, with a non-maximum suppression (NMS) IoU of 0.7 and at most 300 detections per image. AP50:95 is the mean AP over the ten COCO IoU thresholds from 0.5 to 0.95. Precision and recall are reported at a confidence of 0.25 and an IoU of 0.5. The task has one class, so AP equals the mean average precision (mAP). The evaluator was validated against pycocotools on synthetic detections.

Robustness and generalization are measured by three quantities, as given in Eq. (11), (12), and (13):

$$\Delta_{\text{rel}} = \frac{\text{AP50}_{\text{clean}} - \text{AP50}_{\text{patched}}}{\text{AP50}_{\text{clean}}}, \quad (11)$$

$$\Gamma_{\text{dom}} = \frac{\text{AP50}_{Q1} - \text{AP50}_{Q3}}{\text{AP50}_{Q1}}, \quad (12)$$

$$S_{\text{att}} = \frac{|\mathcal{V}_{\text{miss}}|}{|\mathcal{V}|}. \quad (13)$$

Here, Δ_{rel} is the relative AP50 drop under attack within one domain, and Γ_{dom} is the clean domain gap. Both are expressed in percent in the tables. The targeted attack success S_{att} is measured on the first 240 VOC test images. The set $\mathcal{V}$ contains the images in which the largest pedestrian is detected on the clean image (confidence at least 0.25, IoU at least 0.5), and $\mathcal{V}_{\text{miss}} \subseteq \mathcal{V}$ contains those in which it is no longer detected after the patch. About 215 to 231 images are valid per model. The absolute drop in AP50 is not used, because it depends on the level of the clean score.

Statistical comparison uses the sample SD over three seeds and paired t -tests on the per-seed differences. An exact Wilcoxon signed-rank test cannot give $p < 0.25$ with three

pairs, so it is not used for AP50. An image-level paired bootstrap with 150 resamples is added for the seeds 123 and 777, and it reports 95% confidence intervals (CI).

The AAS analysis uses the seed-123 models of three configurations. These are A-matched, A-PAT, and FD²-Det, as defined in Section IV-D. The first 230 test images are scanned per condition (VOC and WiderPerson, clean and patched with the black-box patches). The CI come from 1,000 bootstrap resamples. Paired differences are tested with the two-sided Wilcoxon signed-rank test. About 24 comparisons are made without correction, so the AAS results are exploratory.

D. Baselines

All detectors are YOLOv8m models with one class, the same starting checkpoint, and the same training budget. The variants and their components are listed in Table VIII, and their results are given in Section V-F. A-matched is the equal-budget baseline. It uses the same pipeline as FD²-Det. The blending coefficient is fixed at zero, the attention is bypassed, and no target images are used. A-PAT adds patch augmentation training (PAT) only. A-PAT+GRL adds the domain branch to A-PAT. The full FD²-Det uses all components. Four ablation variants remove one component at a time from the full model.

V. RESULTS

This section reports the results of the four-quadrant evaluation. All scores are AP50 with 101-point interpolation unless stated otherwise. Patched images use the black-box test patches described in Section IV-B. Three configurations are compared throughout, namely A-matched, A-PAT, and FD²-Det, as defined in Section IV-D. The main results and the cross-domain gap are presented first. Robustness to black-box and white-box patches, the paired statistics, the ablation, the training budget, and the computational cost follow.

A. Four-Quadrant Main Results

The four-quadrant results are reported in Table III and Fig. 5(a) as mean and sample SD over three seeds. A-matched reaches 0.920 on clean VOC (Q1) and 0.388 on clean WiderPerson (Q3). Patch training alone leaves both clean scores unchanged (0.922 and 0.386). It raises AP50 on patched VOC from 0.783 to 0.906 (Q2) and on patched WiderPerson from 0.349 to 0.373 (Q4). FD²-Det improves over A-matched only under attack on VOC (Q2: 0.884 versus 0.783). It is lower on the other three quadrants. FD²-Det is lower than A-PAT on all four quadrants, as shown in Table VI. AP50:95 gives the same ordering, as reported in Table IV.

B. Cross-Domain Performance Analysis

The clean domain gap Γ_{dom} is $57.9 \pm 0.3\%$ for A-matched and $58.1 \pm 0.6\%$ for A-PAT. It is larger for FD²-Det, at $63.2 \pm 0.3\%$, as reported in Table V. Patch training therefore does not change the gap, and the full model enlarges it by about 5 points. The lower target AP50 of FD²-Det comes mainly from lower recall at a confidence of 0.25 (0.254 versus 0.322 for A-matched on Q3). Its precision at the same threshold is slightly higher (0.805 versus 0.794).

For reference, the stock COCO-pretrained YOLOv8m without fine-tuning reaches 0.918 on Q1 and 0.518 on Q3 in a single clean evaluation, as listed in Table III. The starting checkpoint reaches 0.885 and 0.469. Fine-tuning on the VOC person images thus reduces target accuracy relative to the stock detector.

C. Robustness to Black-Box Patches

Robustness is summarized in Table V and Fig. 5(b). Without patch training, the black-box patch removes 14.9% of the VOC AP50 and 9.9% of the WiderPerson AP50. It hides the largest pedestrian in 22.7% of the images, against 20.5% for a random patch of the same size. With patch training, the relative drops fall to 1.6% and 3.4%, and the attack success falls to 3.1%. FD²-Det behaves like A-PAT in this respect (1.6%, 3.5%, and 5.0%). For the undefended detector, the optimized patch exceeds the random patch by only 2.2 points of attack success.

D. Paired Statistical Comparison

With three seeds, the evidence is assessed with paired *t*-tests on the per-seed differences. The image-level bootstrap supports the seeds 123 and 777, and Table VI lists both tests. Patch training changes the clean scores by less than the seed-to-seed spread (+0.001 on Q1, -0.002 on Q3). It improves both patched quadrants in every seed (+0.123 on Q2, $p = 0.017$; +0.024 on Q4, $p = 0.007$). FD²-Det is lower than A-PAT on all four quadrants in every seed ($p \leq 0.027$).

In the bootstrap, all 16 intervals of the FD²-Det differences exclude zero. The A-PAT gains on Q2 and Q4 exclude zero in both seeds, and the clean differences on Q1 include zero in both seeds. On Q3, the interval excludes zero for seed 123 only.

E. Breadth of the Threat Model

The seed-123 models were also evaluated with black-box patches of 80, 128, and 160 pixels and with white-box patches optimized against each evaluated detector. The results are listed in Table VII. Black-box patches of 80 and 160 pixels are barely more harmful than random patches of the same size. For A-matched, the attack success is 13.2% against 11.0% at 80 pixels and 28.1% against 25.4% at 160 pixels. The patch-trained detectors remain far less affected than A-matched at 160 pixels (10.2% for A-PAT and 14.1% for FD²-Det versus 28.1%).

Adaptive attacks change the picture. White-box patches hide the largest pedestrian in 85.1%, 60.6%, and 75.0% of the images for A-matched, A-PAT, and FD²-Det. They reduce the VOC AP50 from 0.912, 0.914, and 0.892 to 0.521, 0.620, and 0.582. These adaptive results rely on one seed and three patches per model.

F. Ablation Study

The ablation in Table VIII and Fig. 5(c) removes one component at a time from the full FD²-Det. All variants use seed 42. The seed-to-seed standard deviation of the main comparison is 0.010 to 0.016 on Q1 and 0.001 to 0.008 on Q3. Differences below roughly 0.01 are therefore not

TABLE III. MAIN FOUR-QUADRANT RESULTS (AP50, MEAN $\pm$ SAMPLE SD OVER SEEDS 42, 123, AND 777)

Model	Q1 (clean VOC)	Q2 (patched VOC)	Q3 (clean Wider)	Q4 (patched Wider)
Stock YOLOv8m, no fine-tuning ^a	0.918	N/A	0.518	N/A
Starting checkpoint ^a	0.885	N/A	0.469	N/A
A-matched	0.920 $\pm$ 0.013	0.783 $\pm$ 0.013	0.388 $\pm$ 0.008	0.349 $\pm$ 0.002
A-PAT	0.922 $\pm$ 0.016	0.906 $\pm$ 0.015	0.386 $\pm$ 0.001	0.373 $\pm$ 0.002
FD ² -Det	0.898 $\pm$ 0.010	0.884 $\pm$ 0.012	0.330 $\pm$ 0.006	0.319 $\pm$ 0.005

a. Single clean evaluation with seed 42. Patched scores are not available for these two rows. Bold marks the best score when the gap to the next model is larger than the seed spread.

TABLE IV. AP50:95, RECALL, AND PRECISION (CONFIDENCE 0.25, IOU 0.5; MEAN $\pm$ SAMPLE SD OVER THREE SEEDS)

Model	Q1	Q2	Q3	Q4
<i>AP50:95</i>				
A-matched	0.662 $\pm$ 0.028	0.544 $\pm$ 0.005	0.149 $\pm$ 0.004	0.133 $\pm$ 0.003
A-PAT	0.668 $\pm$ 0.032	0.653 $\pm$ 0.029	0.148 $\pm$ 0.002	0.144 $\pm$ 0.002
FD ² -Det	0.642 $\pm$ 0.027	0.629 $\pm$ 0.026	0.131 $\pm$ 0.002	0.128 $\pm$ 0.003
<i>Recall</i>				
A-matched	0.895 $\pm$ 0.017	0.756 $\pm$ 0.010	0.322 $\pm$ 0.005	0.288 $\pm$ 0.003
A-PAT	0.897 $\pm$ 0.017	0.876 $\pm$ 0.016	0.318 $\pm$ 0.005	0.308 $\pm$ 0.005
FD ² -Det	0.848 $\pm$ 0.012	0.823 $\pm$ 0.012	0.254 $\pm$ 0.003	0.244 $\pm$ 0.002
<i>Precision</i>				
A-matched	0.834 $\pm$ 0.012	0.779 $\pm$ 0.007	0.794 $\pm$ 0.031	0.768 $\pm$ 0.018
A-PAT	0.829 $\pm$ 0.023	0.820 $\pm$ 0.020	0.786 $\pm$ 0.019	0.774 $\pm$ 0.015
FD ² -Det	0.862 $\pm$ 0.010	0.853 $\pm$ 0.014	0.805 $\pm$ 0.022	0.793 $\pm$ 0.018

TABLE V. ROBUSTNESS (MEAN $\pm$ SD OVER THREE SEEDS). RELATIVE DROP IS (CLEAN $-$ PATCHED) / CLEAN OF AP50. ATTACK SUCCESS IS THE SHARE OF IMAGES IN WHICH THE LARGEST PEDESTRIAN IS NO LONGER DETECTED

Model	Relative drop (%)		Clean domain gap (%)	Attack success (%)	
	VOC	Wider		Random patch	Optimized patch
A-matched	14.9 $\pm$ 2.6	9.9 $\pm$ 2.1	57.9 $\pm$ 0.3	20.5 $\pm$ 3.6	22.7 $\pm$ 4.5
A-PAT	1.6 $\pm$ 0.1	3.4 $\pm$ 0.4	58.1 $\pm$ 0.6	3.7 $\pm$ 0.9	3.1 $\pm$ 0.5
FD ² -Det	1.6 $\pm$ 0.3	3.5 $\pm$ 0.6	63.2 $\pm$ 0.3	6.2 $\pm$ 0.8	5.0 $\pm$ 1.0

TABLE VI. PAIRED COMPARISON OF AP50. TOP: MEAN DIFFERENCE OVER THREE SEEDS WITH PAIRED t -TEST p -VALUE IN PARENTHESES. BOTTOM: IMAGE-LEVEL BOOTSTRAP DIFFERENCE WITH 95% CI FOR SEEDS 123 AND 777

Comparison	Seed	Q1	Q2	Q3	Q4
A-PAT $-$ A-matched	3 seeds	+0.001 (0.62)	+0.123 (0.017)	-0.002 (0.74)	+0.024 (0.007)
FD ² -Det $-$ A-matched	3 seeds	-0.022 (0.008)	+0.101 (0.020)	-0.058 (0.002)	-0.031 (0.013)
FD ² -Det $-$ A-PAT	3 seeds	-0.023 (0.027)	-0.023 (0.010)	-0.056 (0.002)	-0.054 (0.003)
A-PAT $-$ A-matched	123	+0.0017 [-0.0023 , +0.0072]	+0.1117 [+0.1004, +0.1214]	+0.0036 [+0.0004, +0.0080]	+0.0258 [+0.0222, +0.0314]
	777	-0.0026 [-0.0070 , +0.0022]	+0.1025 [+0.0920, +0.1129]	+0.0013 [-0.0027 , +0.0064]	+0.0198 [+0.0163, +0.0261]
FD ² -Det $-$ A-matched	123	-0.0197 [-0.0265 , -0.0123]	+0.0881 [+0.0771, +0.0986]	-0.0573 [-0.0622 , -0.0505]	-0.0334 [-0.0376 , -0.0259]
	777	-0.0201 [-0.0244 , -0.0125]	+0.0842 [+0.0725, +0.0940]	-0.0539 [-0.0596 , -0.0471]	-0.0346 [-0.0396 , -0.0276]
FD ² -Det $-$ A-PAT	123	-0.0214 [-0.0278 , -0.0152]	-0.0236 [-0.0290 , -0.0174]	-0.0609 [-0.0660 , -0.0533]	-0.0592 [-0.0642 , -0.0529]
	777	-0.0175 [-0.0229 , -0.0099]	-0.0183 [-0.0247 , -0.0116]	-0.0553 [-0.0625 , -0.0486]	-0.0544 [-0.0598 , -0.0475]

TABLE VII. THREAT-MODEL BREADTH FOR THE SEED-123 MODELS. ATTACK SUCCESS IN PERCENT (RANDOM-PATCH CONTROL IN PARENTHESES) AND AP50 ON ATTACKED VOC TEST IMAGES. CLEAN AP50 IS 0.912, 0.914, AND 0.892

Attack	A-matched		A-PAT		FD ² -Det	
	Success (control)	AP50	Success (control)	AP50	Success (control)	AP50
Black-box, 80 px	13.2 (11.0)	0.849	4.4 (5.3)	0.906	5.5 (5.9)	0.885
Black-box, 128 px	21.5 (18.0)	0.787	3.1 (4.4)	0.899	5.9 (6.4)	0.875
Black-box, 160 px	28.1 (25.4)	0.738	10.2 (11.5)	0.860	14.1 (13.2)	0.828
White-box, 128 px	85.1 (18.0)	0.521	60.6 (4.4)	0.620	75.0 (6.4)	0.582

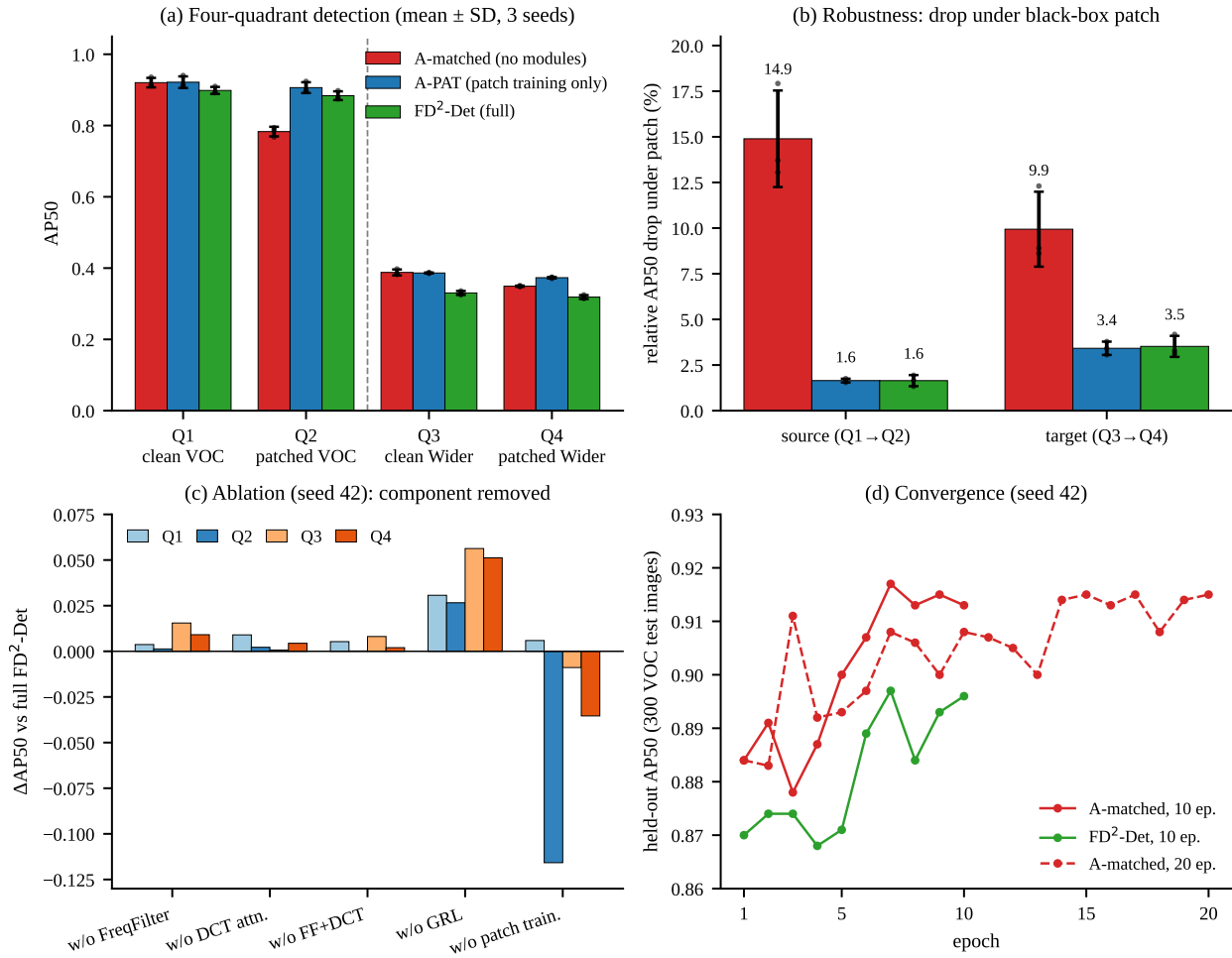


Fig. 5. Evaluation of the three configurations. (a) Four-quadrant AP50 (bars: mean ± SD, dots: individual seeds). (b) Relative AP50 drop under the black-box patch. (c) Ablation at seed 42: change in AP50 when one component is removed from the full FD²-Det (the labels DCT attn. and FF+DCT denote the channel attention and the removal of both frequency modules, which is the variant A-PAT+GRL). (d) Held-out AP50 during training (300 VOC test images, seed 42), including a 20-epoch run of A-matched.

interpretable. Removing patch training collapses the robustness (Q2 -0.116, attack success 26.7%). Removing the domain-adversarial branch is the only change that improves clean accuracy substantially (Q1 +0.031, Q3 +0.056), and it returns the model to the level of A-PAT. Removing the frequency filter (+0.004 on Q1 and +0.016 on Q3), the channel attention (+0.009 and +0.001), or both (A-PAT+GRL against the full model, +0.005 and +0.008) gives small changes. Only the Q3 change for the frequency filter is about twice the Q3 spread, and it favors removal. The frequency components therefore show no measurable gain in this setting. This evidence comes from one seed.

G. Training Budget and Convergence

All fine-tuned detectors start from the same YOLOv8m checkpoint and receive 10 further epochs with identical optimizers and schedules. The configurations therefore have the same number of epochs and iterations. FD²-Det processes more images per iteration because of the target batch, so its compute is not identical. The monitoring AP50 of A-matched

stops rising after epoch 7, with values between 0.913 and 0.917, as shown in Fig. 5(d). Doubling the fine-tuning to 20 epochs does not improve A-matched in-domain (Q1 0.924 versus 0.936 at seed 42, within the seed spread of 0.013). It lowers the target-domain accuracy (Q3 0.353 versus 0.397, about five seed SD). A-matched shows no sign of being under-trained at 10 epochs.

H. Computational Cost

The deployed FD²-Det has 26.116 M parameters, which is 1.0% more than the 25.857 M of the YOLOv8m baseline, as listed in Table IX. The counted computation is 78.68 giga floating-point operations (GFLOPs) for both models, which is twice the number of multiply-accumulate operations. The profiler does not count the Fourier transform operations, so an analytic estimate of 0.12 GFLOPs is added for them. This gives about 78.80 GFLOPs (+0.15%). The measured latency rises from 21.67 ms to 22.78 ms, an increase of about 5%. The operation count therefore understates the real overhead. A likely reason is that the transforms, the pooling, and the

TABLE VIII. MODEL VARIANTS, THEIR COMPONENTS, AND ABLATION RESULTS AT SEED 42 (AP50 AND TARGETED ATTACK SUCCESS UNDER THE BLACK-BOX PATCH)

Variant	Components				AP50				Attack success (%)
	FreqFilter	Attention	GRL	PAT	Q1	Q2	Q3	Q4	
A-matched	×	×	×	×	0.936	0.768	0.397	0.349	27.6
A-PAT	×	×	×	✓	0.940	0.924	0.388	0.375	2.6
A-PAT+GRL	×	×	✓	✓	0.915	0.898	0.344	0.327	5.9
FD ² -Det (full)	✓	✓	✓	✓	0.910	0.898	0.336	0.325	4.0
w/o FreqFilter	×	✓	✓	✓	0.914	0.899	0.351	0.334	4.5
w/o attention	✓	×	✓	✓	0.919	0.900	0.337	0.329	3.7
w/o GRL	✓	✓	×	✓	0.941	0.924	0.392	0.376	2.6
w/o PAT	✓	✓	✓	×	0.916	0.782	0.327	0.290	26.7

TABLE IX. COMPLEXITY OF FD²-DET AND THE YOLOV8M BASELINE

Model	Params (M)	GFLOPs	Latency (ms)
YOLOv8m baseline	25.857	78.68	21.67 ± 0.15
FreqFilter (×3)	3 parameters	≈0.12 ^a	N/A
Channel attention (×3)	+0.2595	<0.01	N/A
Domain classifier	+0.1643	N/A	N/A
FD ² -Det (training)	26.281 (+1.6%)	N/A	N/A
FD ² -Det (deployed)	26.116 (+1.0%)	78.80 ^a (+0.15%)	22.78 ± 0.13 (+5%)

small attention layers run as many small GPU kernels. This explanation is a hypothesis and was not measured separately. Latency covers the network forward pass only, without pre-processing or non-maximum suppression. It was measured on one Tesla T4 GPU with batch size 1, 640 × 640 input, and 32-bit floating point precision, using 200 timed iterations after 50 warm-up iterations. Training takes about 50% more time per epoch when the unlabeled target pass is active.

I. Key Observations

Patch augmentation gives most of the robustness to black-box patches and costs no clean accuracy. The domain-adversarial branch lowers clean accuracy on both domains, and the frequency modules show no measurable gain. Black-box surrogate patches are only slightly stronger than random occluders, while white-box patches defeat all detectors.

VI. EXPLAINABILITY ANALYSIS

This section analyzes the attention behavior of the detectors with Grad-CAM visualizations and the AAS. The analysis uses the seed-123 models of A-matched, A-PAT, and FD²-Det.

A. Grad-CAM Qualitative Analysis

The AAS cannot separate attention on the pedestrian body from attention on the patch inside the same box. Grad-CAM maps are therefore examined as complementary qualitative evidence. The maps for A-matched and FD²-Det on clean and patched images are shown in Fig. 6. The maps of A-PAT are not shown, so the figure does not cover the effect of patch training that appears in the AAS values.

In the first VOC example, both detectors place part of the activation on the pedestrians and part on the background and

the animal. In the second VOC example, the strongest activation lies outside the pedestrian box for both detectors. The patch region receives little activation in all patched examples. In the crowded WiderPerson examples, the maps are weak or diffuse and often lie in regions without pedestrians. FD²-Det does not show visibly more concentrated activation than A-matched. The examples are not selected by a quantitative criterion.

B. Quantitative Attention Results

The AAS values are reported in Table X. Each cell uses the images that are valid for all three models in that condition (228, 212, 223, and 225 images). The absolute values are low for all models, between 0.25 and 0.34. FD²-Det has the lowest or an equal AAS in every condition. On clean VOC, its AAS is lower than that of A-matched by 0.029 (95% CI −0.050 to −0.008, $p = 0.007$) and lower than that of A-PAT by 0.039 (CI −0.059 to −0.020, $p < 0.001$). On clean WiderPerson, it is lower than A-matched by 0.049 (CI −0.075 to −0.021, $p < 0.001$) and equal to A-PAT (−0.007, $p = 0.95$).

A-PAT differs from A-matched in a few conditions. On patched VOC, its AAS is higher by 0.031 (CI 0.003 to 0.057, $p = 0.035$), and its AAS_{body} is higher by 0.038 (CI 0.013 to 0.062, $p = 0.006$). On clean and patched WiderPerson, it is lower by 0.042 ($p < 0.001$) and 0.028 ($p = 0.018$). On patched images, the share of Grad-CAM energy on the patch is small for all models (1.4% to 3.0%) and lower for the patch-trained models. On VOC, it is 0.019 for A-PAT and 0.014 for FD²-Det, against 0.030 for A-matched ($p \leq 0.002$). The bootstrap intervals and the Wilcoxon tests disagree for three patch-share differences, and these are treated as inconclusive. All p -values are uncorrected, and the results come from one seed.

VII. DISCUSSION

A. Do Robustness and Generalization Conflict or Reinforce Each Other?

The results answer the central question in three parts. First, patch augmentation shows no conflict between the two goals. It raises patched accuracy on both domains (Q2 +0.123, Q4 +0.024) and leaves clean accuracy unchanged (Q1 +0.001, Q3 −0.002). Second, the proposed modules do not link the two goals. FD²-Det is below A-PAT on all four quadrants. Third, the domain branch is the only component that trades one quantity against another. It keeps the patch robustness and

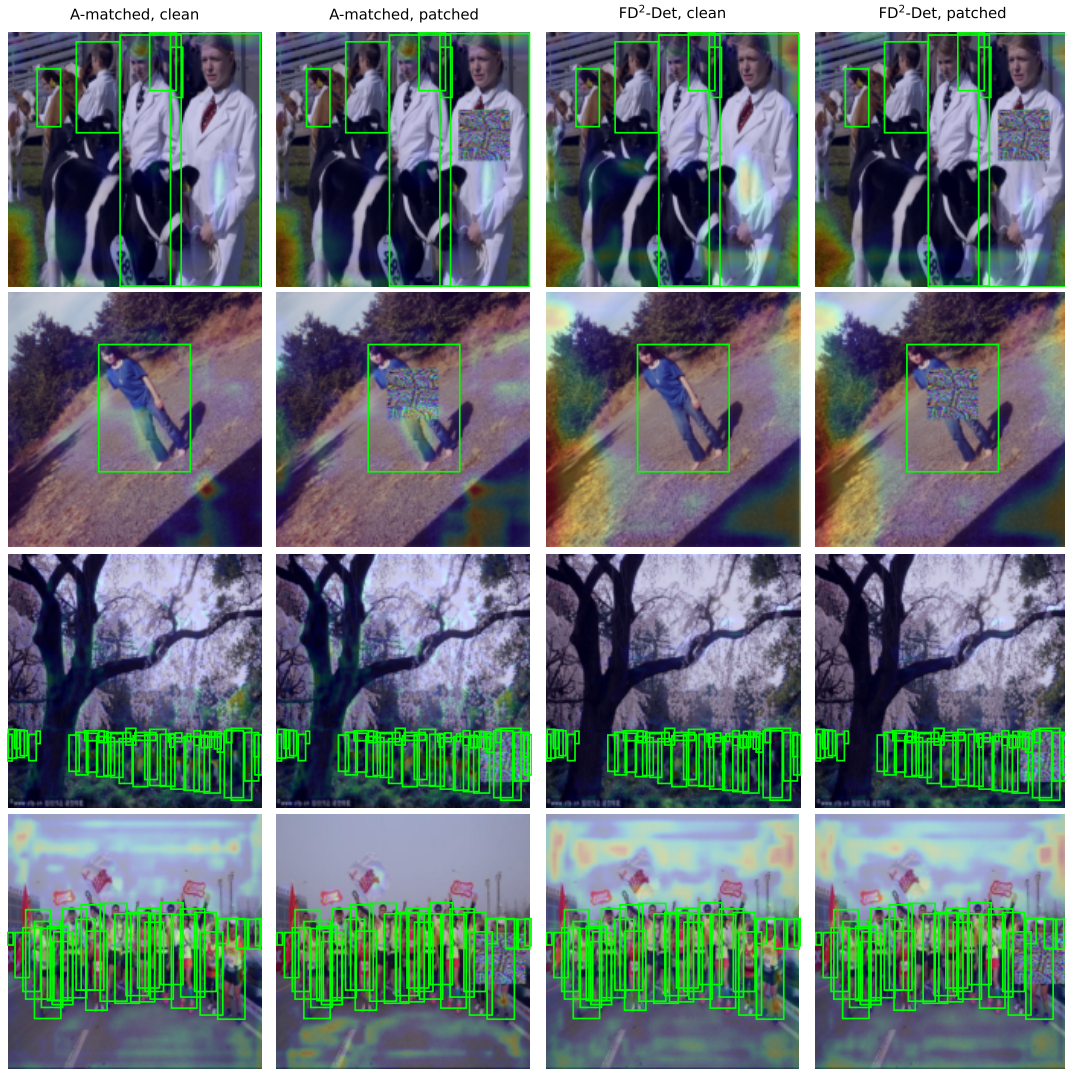


Fig. 6. Grad-CAM maps of the seed-123 models. Rows 1 and 2 show VOC images, and rows 3 and 4 show WiderPerson images. The columns show A-matched and FD²-Det on clean and patched images. Green boxes mark the ground-truth pedestrians. The maps are taken at the neck level that produced the top-scoring detection.

TABLE X. ATTENTION ACTIVATION SCORE OF THE SEED-123 MODELS: MEAN AND 95% BOOTSTRAP CI OVER THE IMAGES VALID FOR ALL THREE MODELS (N = 228, 212, 223, AND 225)

Metric	Model	VOC clean	VOC patched	Wider clean	Wider patched
AAS	A-matched	0.289 [0.252, 0.326]	0.314 [0.275, 0.349]	0.296 [0.269, 0.324]	0.276 [0.251, 0.304]
	A-PAT	0.300 [0.263, 0.339]	0.344 [0.305, 0.385]	0.254 [0.224, 0.282]	0.248 [0.222, 0.275]
	FD ² -Det	0.261 [0.226, 0.299]	0.295 [0.257, 0.334]	0.247 [0.222, 0.272]	0.246 [0.220, 0.271]
AAS _{body}	A-matched	N/A	0.296	N/A	0.265
	A-PAT	N/A	0.333	N/A	0.237
	FD ² -Det	N/A	0.287	N/A	0.237
Patch share	A-matched	N/A	0.030	N/A	0.022
	A-PAT	N/A	0.019	N/A	0.020
	FD ² -Det	N/A	0.014	N/A	0.016

lowers the clean accuracy. The data therefore do not support the hypothesis of Section III-B that a shared frequency mechanism improves both goals. They support a weaker statement. The two goals do not conflict when patch augmentation is used. The expected and the observed effect of each component are

compared in Table XI.

The matched baseline never sees patches during fine-tuning. The random-patch control (20.5%) shows that most of the damage from the black-box patch comes from occlusion

TABLE XI. EXPECTED AND OBSERVED EFFECT OF EACH COMPONENT

Component	Expected effect	Observed effect
Patch augmentation	Lower drop under patches at no clean cost	Relative AP50 drop from 14.9% to 1.6% on VOC and from 9.9% to 3.4% on WiderPerson; clean AP50 unchanged (Q1 +0.001, Q3 -0.002)
FreqFilter	Better and patched cross-domain accuracy	No measurable gain (removal gives Q1 +0.004, Q3 +0.016, one seed)
Channel attention	Better and patched cross-domain accuracy	No measurable gain (removal gives Q1 +0.009, Q3 +0.001, one seed)
GRL domain branch	Better target-domain accuracy	Lower clean accuracy against A-PAT (Q1 -0.023, Q3 -0.056); removal restores the A-PAT level

and not from the optimized texture. Patch training probably removes most of the effect because it exposes the detector to large occluders on the torso. This explanation was not isolated. A baseline trained with random occluders would be needed to separate occlusion robustness from patch-specific robustness. White-box patches defeat all three detectors. They are optimized against the exact weights of each model, so they can find textures that the fixed training patches do not cover. Training with patches from one generator is not adversarial training against an adaptive attacker. The domain branch lowers clean accuracy on both domains. Three explanations are possible, and none was tested. The unlabeled target batches contain only four images per step and update the batch-normalization statistics. Alignment of a globally pooled descriptor may not align object-level features. The loss weight and the reversal schedule were also not tuned. The frequency components show no measurable gain in this setting. Both modules are close to existing designs [58], [59], so no claim of novelty is made for them. The contribution of this study is the controlled audit.

B. Evaluation Safeguards and Interpretation

Five safeguards were applied in the evaluation protocol. First, detector inputs use the pixel scaling that the starting weights expect. Second, detections are kept down to a confidence of 0.001 and scored with COCO-style AP. A threshold of 0.25 with 11-point AP gives low-recall detectors the floor score of $1/11 = 0.091$, so different runs look identical. Third, flips are applied to the images and the boxes together. Fourth, the adaptation and evaluation images of the target domain are disjoint, and the test patches differ from the training patches. Fifth, all variants share the same starting weights, epochs, and optimizers, and the backbone is trainable in all of them. A second observation concerns the AAS under attack. The AAS of patched images can be higher than that of clean images (VOC, A-matched: 0.314 versus 0.289; A-PAT: 0.344 versus 0.300). Attention on the patch could raise the score. The patch-aware variants show that this is not the main cause. The patch share is only 1.4% to 3.0%, and AAS_{body} of A-PAT (0.333) stays above its clean value (0.300). The two values use different image sets, so the comparison is indicative only, and a selection effect from the confidence cut-off was not tested.

C. Failure Cases and Limitations

Several failure cases are visible. White-box patches hide the largest pedestrian in 61% to 85% of the images, and all detectors lose 58% to 63% of their clean AP50 on the target domain. Recall on the dense WiderPerson scenes stays between 0.25 and 0.32. The study also has limits. Only YOLOv8m and two datasets are used, and published robust or domain-adaptive methods were not trained under the same protocol. The person definition and the object size differ between the datasets, which may add to the measured domain gap. The main comparison uses three seeds, while the ablation, the attention analysis, and the white-box attacks use one seed. One T4 GPU under a weekly quota of about 30 hours, with sessions of at most 12 hours, limited the number of seeds, detectors, and datasets. The threat model covers universal torso patches with one attack loss. The black-box patches come from a surrogate that shares its initialization with the evaluated detectors, the training and test patches share one generator, and physical tests are not included. The AAS has no chance baseline, so its values should be read as relative. The starting checkpoint was selected on the VOC2007 test split, which adds a small bias shared by all models. These limits narrow the scope of the conclusions. They do not change the comparison between the matched baselines, which uses equal budgets and three seeds.

D. Strengths and Implications

The study compares each component with matched baselines under equal training budgets. Three seeds, paired statistics, random-patch controls, and COCO-style metrics support the main findings, and the attention analysis separates the pedestrian body from the patch. Four implications follow for practice. First, patch augmentation is the first measure to consider. It adds no inference cost and lowers the black-box patch drop without a clean cost, so it is a low-risk default for pedestrian detectors that face printed patches. Second, a domain-adversarial branch should be used with care. It lowered clean accuracy on both domains here, so its effect should be checked against a matched baseline before it is adopted. Third, robustness claims should be tested with random-patch controls and with patches optimized against the evaluated model. The random control showed that most of the black-box damage came from occlusion, and the white-box patches showed that robustness to one patch family does not carry over to an adaptive attacker. Fourth, new modules should be audited one at a time under equal budgets and several seeds. Frequency components that look plausible in theory gave no measurable gain at this scale, and a single-seed comparison could not separate such small differences from noise. The studied modules add only 1.0% parameters and about 5% latency, so cost is not the reason to drop them. The reason is the missing gain.

E. Comparison with Related Approaches

The reviewed works on adversarial robustness or domain adaptation [35]–[37], [44]–[47] are compared with this study in Table XII, and Table I lists their main limitations. Physical patch-attack studies report attack success rates of up to 97.0%, yet they do not evaluate cross-domain transfer. Domain-adaptive detection studies reach strong adaptation results, yet they do not evaluate robustness against adversarial

TABLE XII. COMPARISON WITH RELATED APPROACHES ON ADVERSARIAL ROBUSTNESS AND CROSS-DOMAIN GENERALIZATION

No.	Reference	Dataset(s)	Method	Adv.	Dom.	Freq.	Joint	Key reported result
1	[35]	CVC14 (infrared)	Deformable physical patch	✓	△	×	×	97.0% attack success rate; no cross-dataset or frequency-domain evaluation
2	[36]	FLIR_ADAS	Sensitive-region physical patch	✓	△	×	×	72.7% (digital) and 90% (physical) attack success rate; no defense proposed
3	[37]	LLVIP	Cross-modal physical patch	✓	✓	×	×	Over 80% attack success rate across modalities; no frequency-domain mechanism
4	[39]	KITTI, nuScenes	Limited-label robustness measure	✓	✓	×	△	Error below 1% and correlation of 0.94 to 0.99 with full-label metrics; evaluation framework only, no defense
5	[40]	CARLA simulator	Robust policy with saliency-based attack detector	✓	✓	×	△	99.38% detector accuracy; simulation only
6	[44]	Cityscapes, foggy and rainy variants	Adversarial GRL with dynamic masking	×	✓	×	×	Improved detection under fog and rain; no adversarial evaluation
7	[45]	Cityscapes, Sim10k, KITTI, VOC	Uncertainty-guided self-training with tiling	×	✓	×	×	Outperforms earlier domain-adaptive detectors; no adversarial evaluation
8	[46]	Sim10k, KITTI, VOC, Pascal-C	Energy-based curriculum with distillation	×	✓	×	×	1% to 5% robustness gain; no adversarial evaluation
9	[47]	Cityscapes, Sim10k, KITTI	Localization regression alignment	×	✓	×	×	Best reported adaptation mAP; no adversarial component
10	This study	VOC, WiderPerson	Audit of FD ² -Det and its components	✓	✓	✓	✓	Patch training lowers the relative AP50 drop on VOC from 14.9% to 1.6%; the GRL branch and the frequency modules give no gain; white-box attack success is 61% to 85%

Note: Adv. = adversarial robustness; Dom. = cross-domain generalization; Freq. = frequency-domain mechanism; Joint = clean, adversarial, and cross-domain conditions evaluated together. ✓ = addressed, × = not addressed, △ = partially addressed. The results in rows 1 to 9 are those reported by the original studies, so they are not directly comparable with row 10.

attacks. The two studies that consider both [39], [40] target broader perception tasks and not pedestrian detection. Existing frequency-domain methods address adversarial defense or generalization in other tasks [50]–[56]. This study evaluates all four conditions in one pedestrian detector with one protocol. The comparison is qualitative, because the reviewed methods were not re-implemented and their reported scores are not directly comparable.

F. Future Research Directions

Future work should extend the audit in five directions. First, other detector families and published robust or domain-adaptive methods should be trained under the same protocol, together with a random-occluder baseline. Second, a third dataset such as KITTI should test whether the conclusions hold for another domain pair. Third, the ablations, the attention analysis, and the adaptive attacks should be repeated with several seeds. Fourth, the threat model should include expectation over transformation, physical tests, other placements, and a transfer matrix across detectors. Finally, the accuracy cost of the domain branch should be traced to its cause, for example by separating the batch-normalization update from the alignment loss, and the AAS should be compared with a chance baseline.

VIII. CONCLUSION

This research asked whether robustness to adversarial patches and cross-domain generalization can be improved together in pedestrian detection. To answer it, FD²-Det was

audited against matched baselines on YOLOv8m. The model combines a frequency filter, a channel attention, a GRL-based domain branch, and patch-augmented training. The evaluation used a four-quadrant protocol with VOC and WiderPerson, three seeds, equal training budgets, and paired statistics. The audit reached its aim and gave a clear answer. Patch augmentation alone reduced the relative AP50 drop under a black-box patch from 14.9% to 1.6% on VOC and from 9.9% to 3.4% on WiderPerson, without a measurable change in clean accuracy. The domain-adversarial branch lowered clean accuracy, and the frequency modules gave no gain beyond seed noise. FD²-Det was below the patch-only baseline on all four quadrants. The threat-model analysis showed that black-box surrogate patches are only slightly stronger than random patches of the same size. White-box patches optimized for each detector still hid the largest pedestrian in 61% to 85% of the images. The patch-aware attention analysis showed no improvement in pedestrian localization for FD²-Det. The deployed modules add 1.0% parameters and about 5% latency. These results suggest that patch robustness and cross-domain generalization do not conflict when patch augmentation is used. They do not support the hypothesis that the tested frequency and adaptation components link the two goals. The conclusions are limited to one detector, two datasets, and universal torso patches, and the limitations in Section VII-C apply. Future work should test other detectors and datasets, repeat the ablations with several seeds, and include adaptive threat models.

DECLARATIONS

Generative artificial intelligence (AI) tools were used during the preparation of this manuscript. ChatGPT was used to improve sentence construction. Gemini was used to improve the context of the images. Claude was used for the overall structure and to solve LaTeX formatting issues. The experiments and results are the authors' own work. The authors reviewed all AI-assisted content and take full responsibility for the accuracy and integrity of the work. No generative AI tool is listed as an author.

ACKNOWLEDGMENT

This research was fully funded by American International University-Bangladesh (AIUB). The authors thank AIUB for this support and are grateful for the resources and the research environment it provided.

REFERENCES

- [1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.
- [2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [3] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 49–55.
- [4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1625–1634.
- [5] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521–1528.
- [6] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3339–3348.
- [7] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8684–8694.
- [8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [9] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380–393, 2019.
- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*, 2014, pp. 740–755.
- [11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
- [12] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "Cascade transformer decoder based occluded pedestrian detection with dynamic deformable convolution and Gaussian projection channel attention mechanism," *IEEE Transactions on Multimedia*, vol. 25, pp. 1529–1537, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10057083/>
- [13] Z. Chen, Y. Lin, J. Xu, K. Lu, and Y. Shou, "Detecting pedestrian with incomplete head feature in crowded situation based on transformer," *IEEE Signal Processing Letters*, vol. 32, pp. 576–580, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10820533/>
- [14] Y. Tang, M. Liu, B. Li, Y. Wang, and W. Ouyang, "NAS-PED: Neural architecture search for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 1800–1817, Mar. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10770837/>
- [15] S. Deng and J. Li, "Efficient dense pedestrian detection based on transformer," in *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Chongqing, China: IEEE, Mar. 2024, pp. 992–997. [Online]. Available: <https://ieeexplore.ieee.org/document/10503874/>
- [16] C. Wu, T. He, and K. Yang, "An efficient dense pedestrian detection model based on RT-DETR," in *2025 IEEE 5th International Conference on Power, Electronics and Computer Applications (ICPECA)*. Shenyang, China: IEEE, Jan. 2025, pp. 23–27. [Online]. Available: <https://ieeexplore.ieee.org/document/10928774/>
- [17] Q. Pan, Y. Yu, Y. Wang, X. Li, and Y. Zhang, "Pedestrian object detection based on transformer architecture," in *2024 6th Asia Symposium on Image Processing (ASIP)*. Tianjin, China: IEEE, Jun. 2024, pp. 34–39. [Online]. Available: <https://ieeexplore.ieee.org/document/10744727/>
- [18] Y. Jia, H. Pan, J. Wang, and W. Sun, "CSP-Lite: Real-time and efficient keypoint-based pedestrian detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 2, pp. 1627–1637, Apr. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10637319/>
- [19] Y. Xu, X. Zong, and O. Kochan, "Pedestrian detection based on improved CSP network," in *2023 17th International Conference on the Experience of Designing and Application of CAD Systems (CADSM)*. Jaroslaw, Poland: IEEE, Feb. 2023, pp. 1–5. [Online]. Available: <https://ieeexplore.ieee.org/document/10076538/>
- [20] Z. Li, Z. Liu, and X. Wang, "On-board real-time pedestrian detection for micro unmanned aerial vehicles based on YOLO-v8," in *2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM)*. Jiuzhaigou, China: IEEE, Jul. 2023, pp. 250–255. [Online]. Available: <https://ieeexplore.ieee.org/document/10339502/>
- [21] A. T. Sadiq, F. Azizova, A. Sujeetha, R. Almajed, and S. Karunakaran, "YOLOv8-based real-time pedestrian detection for autonomous public transport systems," in *2025 International Conference on Electrical Engineering and Informatics (ICEEI)*. Kuching, Malaysia: IEEE, Nov. 2025, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/11330492/>
- [22] G. Sivaraman, E. Sophiya, and M. Diviya, "YOLOv8 for pedestrian detection: A comparative study for pedestrian detection," in *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. Trichirappalli, India: IEEE, Jul. 2024, pp. 1–7. [Online]. Available: <https://ieeexplore.ieee.org/document/10718372/>
- [23] E. Álvarez, R. Álvarez, and M. Cazorla, "Exploring transferability on adversarial attacks," *IEEE Access*, vol. 11, pp. 105 545–105 556, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10264069/>
- [24] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 538–566, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10005100/>
- [25] G. Shi, Z. Lin, A. Peng, and H. Zeng, "An enhanced transferable adversarial attack against object detection," in *2023 International Joint Conference on Neural Networks (IJCNN)*. Gold Coast, Australia: IEEE, Jun. 2023, pp. 1–7. [Online]. Available: <https://ieeexplore.ieee.org/document/10190995/>
- [26] G. Li, Y. Xu, J. Ding, and G.-S. Xia, "Toward generic and controllable attacks against object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10569091/>
- [27] Y. Wei, H. Gao, X. Quan, and G. Luo, "A transferable adversarial attack against object detection networks," in *2024 International Joint Conference on Neural Networks (IJCNN)*. Yokohama, Japan: IEEE, Jun. 2024, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/10651486/>
- [28] Y. Wang, L. Wu, Y. Cao, J. Jin, Z. Zhang, E. Wang, C. Ma, and Y. Zhao, "A highly transferable camouflage attack against object

- detectors in the physical world,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 7, pp. 10373–10385, Jul. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10965837/>
- [29] X. Wang, J. Chen, Z. Zhang, K. He, Z. Wu, R. Du, Q. Li, and G. Liu, “Transferable and robust dynamic adversarial attack against object detection models,” *IEEE Internet of Things Journal*, vol. 12, no. 11, pp. 16171–16180, Jun. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10845771/>
- [30] Y. Wang, L. Wu, J. Jin, E. Wang, Z. Zhang, and Y. Zhao, “An imperceptible adversarial attack against 3-D object detectors in autonomous driving,” *IEEE Internet of Things Journal*, vol. 12, no. 12, pp. 21404–21414, Jun. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10909657/>
- [31] M. Naseer, S. Khan, and F. Porikli, “Local gradients smoothing: Defense against localized adversarial attacks,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.
- [32] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, “Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14973–14982.
- [33] J. Lu, C. Wang, Y. Huang, K. Ding, and X. Liu, “An adversarial example defense algorithm for intelligent driving,” *IEEE Network*, vol. 38, no. 6, pp. 98–105, Nov. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10507054/>
- [34] C. Nantabut, “Transfer learning for pedestrian detection across domains with YOLO and faster R-CNN,” in *2025 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C)*. Bangkok, Thailand: IEEE, Aug. 2025, pp. 159–163. [Online]. Available: <https://ieeexplore.ieee.org/document/11282766/>
- [35] Z. Liu, J. Dai, and J. Geng, “A novel physical world adversarial patch attack for infrared pedestrian detector,” in *2024 IEEE International Conference on Unmanned Systems (ICUS)*. Nanjing, China: IEEE, Oct. 2024, pp. 1363–1368. [Online]. Available: <https://ieeexplore.ieee.org/document/10840169/>
- [36] X. Wang and W. Li, “Physical adversarial attacks for infrared object detection,” in *2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. Guangzhou, China: IEEE, Jan. 2024, pp. 64–69. [Online]. Available: <https://ieeexplore.ieee.org/document/10504222/>
- [37] X. Wei, Y. Huang, Y. Sun, and J. Yu, “Unified adversarial patch for visible-infrared cross-modal attacks in the physical world,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2348–2363, Apr. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10310159/>
- [38] M. Zhou, W. Zhou, J. Huang, J. Yang, M. Du, and Q. Li, “Stealthy and effective physical adversarial attacks in autonomous driving,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6795–6809, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10583962/>
- [39] M. T. Nguyen, Q.-H. Luu, and H. L. Vu, “A novel robustness measure for evaluating perceptions in autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 12, pp. 22178–22188, Dec. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/11218000/>
- [40] C. Wang and N. Aouf, “Explainable deep adversarial reinforcement learning approach for robust autonomous driving,” *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 4, pp. 2551–2563, Apr. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10475532/>
- [41] H. Wei, H. Tang, X. Jia, Z. Wang, H. Yu, Z. Li, S. Satoh, L. Van Gool, and Z. Wang, “Physical adversarial attack meets computer vision: A decade survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9797–9817, Dec. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10602786/>
- [42] T. Werner, “Applications of robust statistics in autonomous driving,” *IEEE Access*, vol. 13, pp. 167848–167892, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/11177222/>
- [43] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International Conference on Machine Learning (ICML)*. PMLR, 2015, pp. 1180–1189. [Online]. Available: <https://proceedings.mlr.press/v37/ganin15.html>
- [44] J. Li, R. Xu, X. Liu, J. Ma, B. Li, Q. Zou, J. Ma, and H. Yu, “Domain adaptation based object detection for autonomous driving in foggy and rainy weather,” *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 2, pp. 900–911, Feb. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10574400/>
- [45] M. A. Munir, M. H. Khan, M. S. Sarfraz, and M. Ali, “Domain adaptive object detection via balancing between self-training and adversarial learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14353–14365, Dec. 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10173487/>
- [46] A. Banitalebi-Dehkordi, A. Amirkhani, and A. Mohammadinasab, “EBCDet: Energy-based curriculum for robust domain adaptive object detection,” *IEEE Access*, vol. 11, pp. 77810–77825, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10192408/>
- [47] Z. Piao, L. Tang, and B. Zhao, “Unsupervised domain-adaptive object detection via localization regression alignment,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 15170–15181, Nov. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10154573/>
- [48] S. Zhang, J. Fu, J. Yu, H. Xu, H. Zha, S. Mao, and Y. Lin, “Channel-robust class-universal spectrum-focused frequency adversarial attacks on modulated classification models,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 4, pp. 1280–1293, Aug. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10479476/>
- [49] D. Ke, X. Wang, and Z. Huang, “Frequency-selective adversarial attack against deep learning-based wireless signal classifiers,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4001–4011, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10387523/>
- [50] Y. Meng, P. Qi, S. Zheng, Z. Cai, X. Zhou, and T. Jiang, “Adversarial attack and reliable defense based on frequency domain feature enhancement for automatic modulation classification,” *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 3731–3744, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10937210/>
- [51] S. Zhang, Y. Lin, J. Yu, J. Zhang, Q. Xuan, D. Xu, J. Wang, and M. Wang, “HFAD: Homomorphic filtering adversarial defense against adversarial attacks in automatic modulation classification,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 3, pp. 880–892, Jun. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10418157/>
- [52] S. Basar, M. Ali, A. Waheed, M. Ahmad, and M. H. Miraz, “A novel defocus-blur region detection approach based on DCT feature and PCNN structure,” *IEEE Access*, vol. 11, pp. 94945–94961, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10233857/>
- [53] A. Q. Vu, H. Q. Bui, L. T. Nguyen, and T. N. Le, “DCT-based white blood cell image enhancement for recognition using deep learning,” *IEEE Access*, vol. 12, pp. 171571–171588, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10756654/>
- [54] Z. Wang, Z. Zhang, W. Qi, F. Yang, and J. Xu, “FreqGAN: Infrared and visible image fusion via unified frequency adversarial learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 728–740, Jan. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10680110/>
- [55] H. Wu, S. Yan, C. Liu, A. Wang, M. Wang, and L. Yu, “Hyperspectral open-set classification via frequency-domain rotation enhancement and multibranch adversarial routing,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 24864–24882, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/11176803/>
- [56] P. Duan, T. Shan, X. Kang, and S. Li, “Spectral super-resolution in frequency domain,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 12338–12348, Jul. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10737880/>
- [57] L. Papangelo, M. Pistilli, S. Sciancalepore, G. Oliveri, G. Piro, and G. Boggia, “Adversarial machine learning for image-based radio frequency fingerprinting: Attacks and defenses,” *IEEE Communications Magazine*, vol. 62, no. 11, pp. 108–113, Nov. 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10384628/>
- [58] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, “Global filter networks

- for image classification,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.
- [59] Z. Qin, P. Zhang, F. Wu, and X. Li, “FcaNet: Frequency channel attention networks,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 783–792.
- [60] H. Pang and C. Huang, “Pedestrian detection: An explainable approach,” in *2024 International Conference on Machine Learning and Applications (ICMLA)*. Miami, FL, USA: IEEE, Dec. 2024, pp. 985–990. [Online]. Available: <https://ieeexplore.ieee.org/document/10903390/>
- [61] G. G. Devarajan, M. Thirunavukkarasan, M. Mahalingum, D. Arockiam, M.-E. Wu, and R. P. Mahapatra, “Explainable federated learning-based secure and transparent object detection model for autonomous vehicles,” *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 6263–6270, May 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10819475/>
- [62] L. Fernandes, J. N. D. Fernandes, M. Calado, J. R. Pinto, R. Cerqueira, and J. S. Cardoso, “Intrinsic explainability for end-to-end object detection,” *IEEE Access*, vol. 12, pp. 2623–2634, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10373851/>
- [63] Z. Cui, Z. Yang, Z. Zhou, L. Mou, K. Tang, Z. Cao, and J. Yang, “Deep neural network explainability enhancement via causality-erasing SHAP method for SAR target recognition,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10539961/>
- [64] Ultralytics, “Ultralytics YOLOv8,” <https://github.com/ultralytics/ultralytics>, 2024, version 8.2.18.
- [65] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: <https://openreview.net/forum?id=rJzIBfZAb>
- [66] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974. [Online]. Available: <https://doi.org/10.1109/T-C.1974.223784>