

Received 10 April 2025, accepted 16 May 2025, date of publication 2 June 2025, date of current version 11 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3576264

RESEARCH ARTICLE

OD-SHIELD: Convolutional Autoencoder-Based Defense Against Adversarial Patch Attacks in Object Detection

BYEONGCHAN KIM¹, HEEMIN KIM¹, MINJUNG KANG¹, HYUNJEE NAM¹,
SUNGHWAN PARK², (Graduate Student Member, IEEE), JAEWOO LEE³, (Member, IEEE),
AND IL-YOUP KWAK^{1,4,5}, (Member, IEEE)

¹Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, South Korea

²Department of Security Convergence Science, Chung-Ang University, Seoul 06974, South Korea

³Department of Industrial Security, Chung-Ang University, Seoul 06974, South Korea

⁴Department of Applied Statistics, Chung-Ang University, Seoul 06974, South Korea

⁵Department of Smart Cities, Chung-Ang University, Seoul 06974, South Korea

Corresponding authors: Jaewoo Lee (jaewoolee@cau.ac.kr) and Il-Youp Kwak (ikwak2@cau.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2023-00208284); This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2020-II201655, 33%); This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program(IITP-2025-RS-2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT In the evolving landscape of deep neural network security, adversarial patch attacks present a serious challenge for object detection systems. We introduce OD-Shield, a novel defense approach that employs a convolutional autoencoder framework to detect and remove anomalous regions in patched images, subsequently reconstructing these regions to mitigate adversarial effects. The reconstructed images are then processed by the object detector, thereby restoring reliable performance under diverse attack scenarios. Distinctly model-agnostic, OD-Shield operates as a pre-processing step and can be applied to a wide range of tasks—including image classification and object detection—without compromising the fidelity of the original image. Experiments on benchmark datasets (COCO, Visdrone, and Argoverse) reveal that OD-Shield outperforms existing defenses by 13%–47% on COCO, highlighting its effectiveness in addressing critical security vulnerabilities. This work not only tackles the immediate threat of adversarial patches but also lays the foundation for future research into adaptive, resilient defense mechanisms that keep pace with evolving adversarial tactics.

INDEX TERMS Adversarial patch attack, adversarial patch defense, pre-processing defense, object detection.

I. INTRODUCTION

Despite the remarkable progress of deep neural networks (DNNs) in object detection, their susceptibility to adversarial attacks remains a pressing concern. Reference [1] initially exposed the vulnerability of DNNs to adversarial examples, and [2] subsequently introduced the Fast Gradient Sign Method (FGSM), providing an efficient way to craft such examples. These findings bear significant implications for

object detection, where maintaining reliable and trustworthy model outputs is paramount.

Building on this, [3] demonstrated that adversarial perturbations can also be realized in physical, real-world settings. Among these, adversarial *patch* attacks have attracted particular attention due to their practicality and efficiency: by placing small, malicious patches—printable stickers—onto a target object, attackers can radically disrupt detection without needing full access to the image pixels. As a result, an autonomous vehicle, for example, might mistake a stop sign for a speed limit sign [4], [5], [6], potentially leading

The associate editor coordinating the review of this manuscript and approving it for publication was Diana Gratiela Berbecaru¹.

to severe real-world consequences. Such physical attacks underscore the urgent need for robust defense mechanisms that can withstand manipulation in unconstrained environments. Given the diverse nature of adversarial patches, designing comprehensive strategies capable of countering various attack forms is critical.

In this work, we introduce **Object Detection Shield (OD-Shield)**, a novel defense framework that protects object detectors from adversarial patches. Contrary to many existing methods [7], [8], [9], [10], which often rely on masking or obscuring the adversarial region, OD-Shield leverages *contextual information* around the attacked area for reconstruction. As illustrated in Figure 1, this strategy effectively restores occluded object features, thereby increasing detector resilience against adversarial patches. OD-Shield operates in the early stages of the object detector, making it applicable regardless of whether the detector is based on a ViT, a hybrid model [11], [12], or a 2D CNN. Concretely, our approach addresses two key challenges posed by adversarial patches: (i) anomalous feature map signatures, and (ii) contextual disruption in local regions. By alleviating both, OD-Shield enables stable and robust object detection even under adversarial conditions.

We summarize our contribution as follows:

- We propose OD-Shield, a novel framework that *reconstructs* adversarial patches rather than masking them, effectively neutralizing their impact on object detection.
- OD-Shield is model-agnostic and operates as a pre-processing step, thereby offering seamless integration with tasks such as image classification and object detection.
- Through extensive experiments, we demonstrate that OD-Shield defends against a wide spectrum of adversarial patches (including unseen, naturalistic variants) without relying on specific patch characteristics.

II. PRELIMINARY

A. ADVERSARIAL PATCH ATTACK

Adversarial patch attacks constitute a distinct class of adversarial threats for deep learning models, especially in object detection. Unlike conventional adversarial approaches that disperse small perturbations across the entire image, patch attacks insert a visibly manipulative region into the scene. Despite affecting only a limited portion of the image, these attacks can still provoke severe misclassifications or occlude objects.

Formally, crafting an adversarial patch involves optimizing an objective function to induce misclassification for a chosen target label by minimizing the model's detection loss. During this process, the model's weights W remain unaltered, while only the patch pixels are updated:

$$\arg \min_{\delta} L_{\text{det}}(f(W, x \oplus \delta), y'),$$

where $x \in \mathcal{X}$ denotes the original image, δ represents the adversarial patch, $\oplus$ is the operation of overlaying the patch

onto the image, $f(\cdot)$ is the target model, L_{det} is a loss function assessing detection errors, and y' is the intended (incorrect) label. By design, such a patch forces misclassification or concealment of objects across diverse input images.

Following [13], we generate adversarial patches under two different optimization strategies that combine non-printable loss (L_{nps}), total variation loss (L_{tv}), and objectness loss (L_{obj}). Specifically, L_{nps} and L_{tv} encourage the patch to remain realistic and printer-friendly, while L_{obj} compels the model to misclassify an object's category and probability. We investigate two distinct forms:

$$\begin{aligned} L_{\text{nat}} &= L_{\text{obj}} + \alpha L_{\text{nps}} + \beta L_{\text{tv}}, \\ L_{\text{noisy}} &= L_{\text{obj}} + \alpha L_{\text{nps}}. \end{aligned}$$

As illustrated in Figure 2, patches optimized with L_{nat} tend to appear more natural, whereas those generated via L_{noisy} exhibit noisier patterns. We train our defense model using both patch types, thereby equipping it to robustly identify and neutralize a wide range of adversarial patch variations.

B. ENCODER-DECODER ARCHITECTURE

The encoder-decoder architecture is a foundational framework in deep learning, extensively employed across a wide range of domains. It consists of an encoder that compresses input data into a compact and meaningful latent representation, and a decoder that reconstructs the original structure or generates task-specific outputs from this representation. This design is particularly effective in capturing salient features from complex data and establishing robust mappings between inputs and outputs.

Applications of the encoder-decoder architecture span numerous fields, including image processing, natural language processing, and speech signal processing. For example, U-Net, a prominent architecture for image segmentation, utilizes this framework to achieve remarkable performance in biomedical applications [14]. Similarly, Pix2Pix adopts an encoder-decoder structure for image-to-image translation, delivering state-of-the-art results in tasks such as image synthesis and style transfer [15].

In this work, we leverage an encoder-decoder architecture, commonly applied in image reconstruction tasks, to strengthen defenses against adversarial patch attacks. By enabling the network to model the distribution of clean images, this approach enhances its capacity to identify and inpaint anomalies caused by adversarial patches, thereby improving robustness in object detection tasks.

III. PROPOSED APPROACH: OD-SHIELD

To enhance the robustness of object detection systems against adversarial patch attacks, we present a novel defense mechanism, called OD-Shield. Using a convolutional autoencoder as a pre-processing module, we reconstruct regions compromised by adversarial patches, thereby ensuring stable detector performance (see Figure 3).

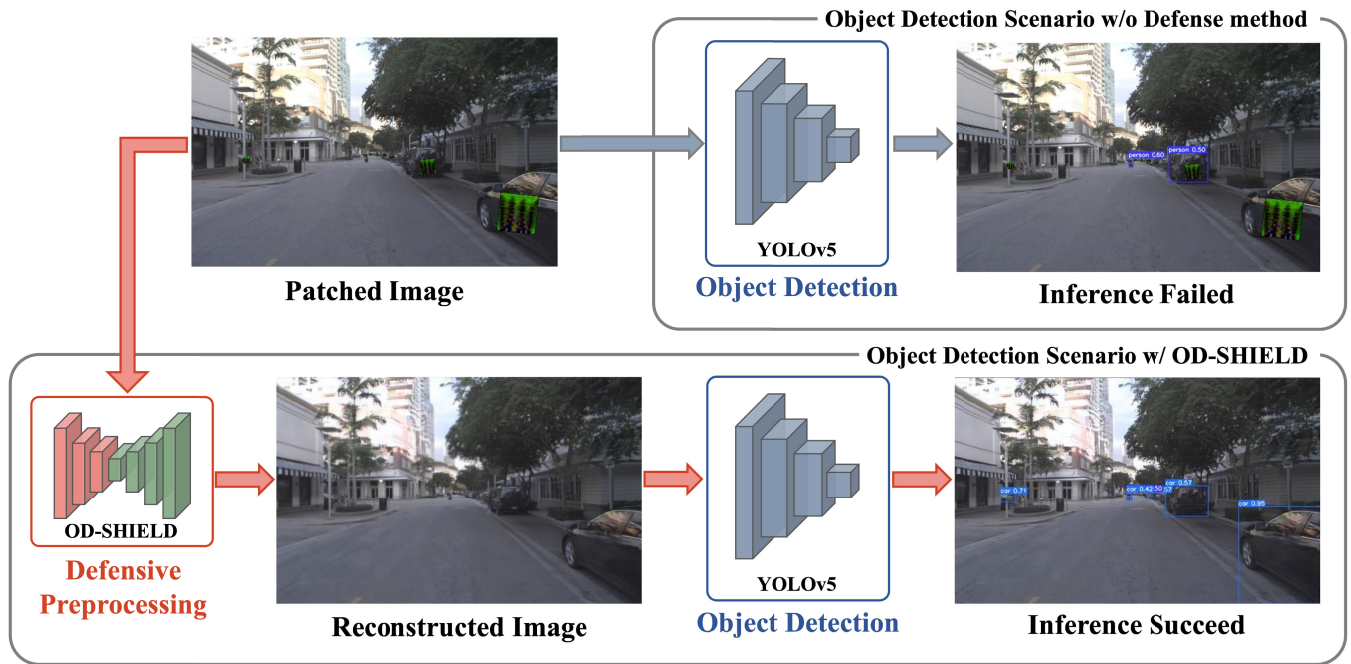


FIGURE 1. In a patch attack scenario (upper-left image), object detection fails to correctly identify objects (upper-right image). However, by applying OD-Shield, the manipulated portion of the original image under the patch attack is eliminated and restored, thereby enabling accurate detection (lower-right image).

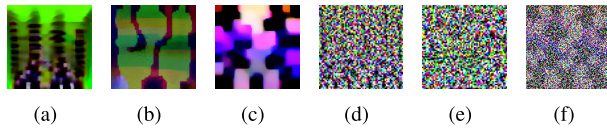


FIGURE 2. Adversarial patches represented as (a), (b), and (c) are generated by minimizing the loss L_{nat} , while adversarial patches denoted as (d), (e), and (f) are generated by minimizing the loss L_{noisy} .

Our training process exploits two key properties of adversarial patches. First, these patches exhibit abnormally high activations in deeper layers of the network (Figure 4). As the depth of the network increases, the model focuses more on the regions where adversarial patches are applied, leading to misclassifications. To address this, we introduced a loss function called **Feature Matching Loss** (FM loss) to reduce the adversarial impact concentrated in the patched regions. Second, adversarial patches are often visually distinct due to their specific shape and density, which can cause a noticeable disruption in the image. To mitigate this, we implemented methods to reconstruct the patched area. We utilized L_2 loss and **Structural Similarity Index Measure** (SSIM) loss, both widely used in image reconstruction tasks. Additionally, we developed the **Self-Supervised Convolutional Block Attention Module** (SSCBAM) module and integrated it into the decoder to utilize broader contextual information, allowing for more accurate reconstruction of compromised regions beyond local surroundings.

Through the coordinated adjustment of loss functions and the integration of SSCBAM, OD-Shield effectively

neutralizes adversarial patches while preserving the original image's integrity. As a result, clean inputs pass through the system unaltered, whereas patched images are rectified to closely resemble their pristine counterparts. This ensures that the subsequent object detector remains robust and accurate, even under severe adversarial interference.

A. FEATURE MATCHING LOSS

Adversarial patches induce anomalous activations in high-level feature maps within the encoder network. As adversarial distortions propagate deeper into the network, they become increasingly pronounced in later feature maps. This phenomenon is evident in Figure 4: for a normal image, the final encoder layer primarily highlights the region corresponding to a car, whereas for an attacked image, the highest activation is observed in the patched area of the car. As a result, the network misinterprets the object's presence, thereby necessitating correction.

To address this issue, we introduce the **Feature Matching** (FM) Loss, computed from the final layer of the encoder in the OD-Shield framework. The FM loss measures the discrepancy between the high-level representations of the clean image I and the adversarially patched image I_p , and the model is trained to minimize this difference:

$$L_{FM}(I, I_p) = \frac{1}{N_p} \sum_{i=1}^H \sum_{j=1}^W (f_I(i, j) - f_{I_p}(i, j))^2,$$

where f_I and f_{I_p} denote the feature maps of the clean image I and the patched image I_p , respectively. By minimizing this

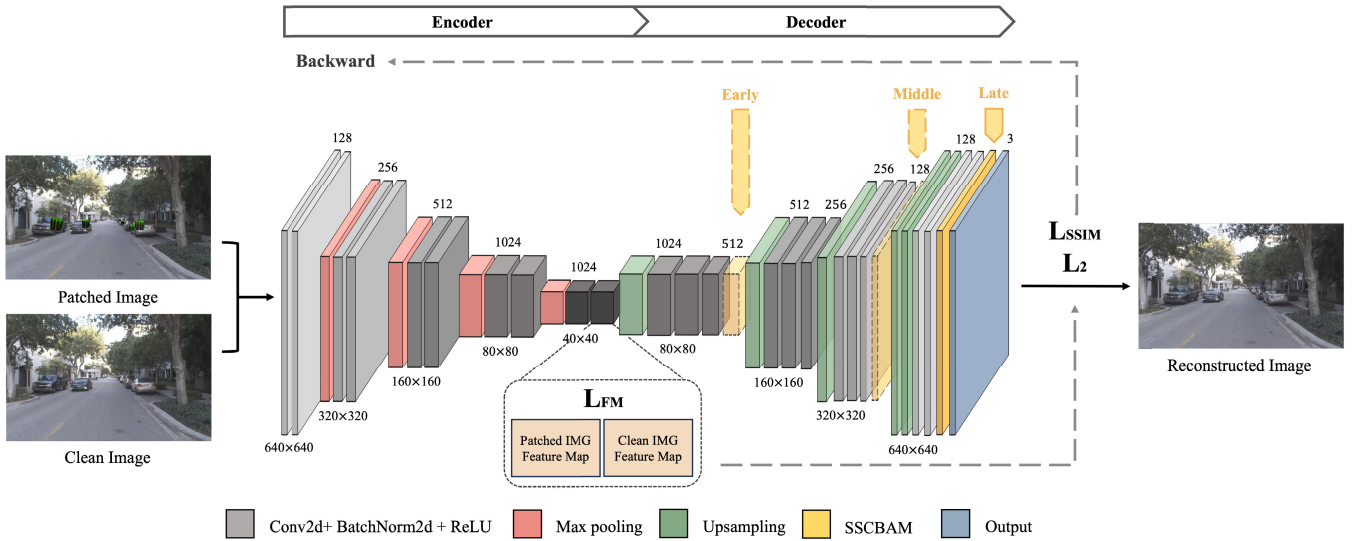


FIGURE 3. OD-Shield is a convolutional autoencoder-based defense. During training, pairs of original and patched images are given as inputs, and the model learns to minimize L_2 , L_{SSIM} , and L_{FM} . In the decoder, the yellow dashed boxes indicate positions where the attention module is integrated. Among the tested modules—SSCBAM (Self-Supervised Convolutional Block Attention Module), SSPCAB (Self-Supervised Predictive Convolutional Attention Block) [16], and SSMCTB (Self-Supervised Masked Convolutional Transformer Block) [17]—our final model adopts SSCBAM in the late layers.

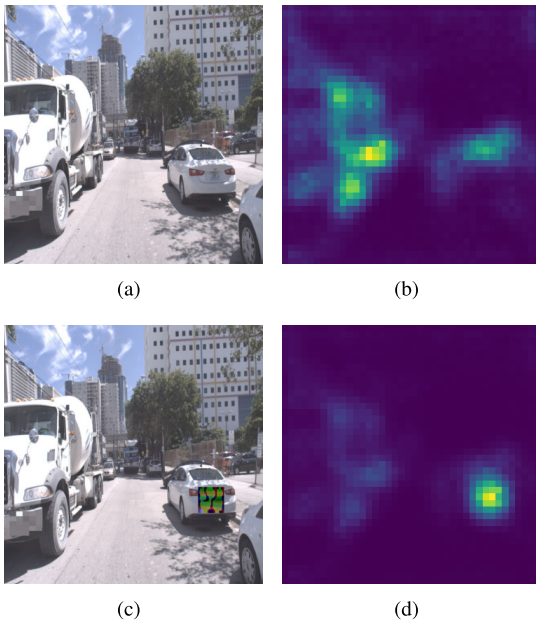


FIGURE 4. Images (a) and (c) represent the original image and the image under adversarial patch attack, while images (b) and (d) visualize the norm of the feature vector of the last block (10th conv layer) in the encoder for (a) and (c), respectively.

loss, the network reduces adversarial artifacts in the deeper feature representations.

B. CONVOLUTION ATTENTION MODULE

To improve the reconstruction of adversarially-corrupted regions, we integrate three specialized attention modules into the decoder of OD-Shield: the Self-Supervised Predictive Convolutional Attention Block (SSPCAB) [16], the

Self-Supervised Masked Convolutional Transformer Block (SSMCTB) [17], and the Self-Supervised Convolutional Block Attention Module (SSCBAM). Each module leverages distinct mechanisms to capture contextual and structural information, thereby enhancing the model's ability to restore occluded areas.

SSPCAB incorporates the Squeeze-and-Excitation (SE) module [18], which emphasizes inter-channel dependencies by selectively amplifying critical feature channels while suppressing less relevant ones. This process prioritizes key spatial details for improved reconstruction. Furthermore, dilated convolutions expand the receptive field, facilitating the capture of broader spatial cues in severely occluded regions.

In SSMCTB, the SE module is replaced by a transformer-based self-attention mechanism. This design enables the network to dynamically focus on relevant areas throughout the feature map, capturing long-range dependencies. By aggregating information from both local and distant regions, SSMCTB enhances structural consistency when restoring adversarially occluded areas.

SSCBAM combines masked convolution with the Convolutional Block Attention Module (CBAM) [19]. Masked convolution selectively infers missing pixels in the occluded regions by referencing neighboring unmasked areas, thereby improving reconstruction fidelity. Subsequently, CBAM refines feature selection via sequential channel and spatial attention, adaptively highlighting the most critical regions. This dual-phase approach strengthens the model's ability to recover the original image layout.

We conducted individual experiments with SSPCAB, SSMCTB, and SSCBAM to evaluate their respective contributions. These analyses examined module-specific

performance in handling adversarial patch occlusions in object detection tasks, providing insights into how each approach leverages contextual and structural information to restore corrupted features. Our experiments on the COCO, VisDrone, and Argoverse Dataset showed that the SSCBAM module generally outperformed the others in most scenarios. However, the performance differences were subtle, indicating the need for further experiments under diverse conditions. Detailed experimental results are provided in the Section IV-D.

C. TRAINING

During training, OD-Shield receives both adversarially attacked and clean images as inputs. As described in Section III-A, the feature maps of these two inputs exhibit minimal divergence in the encoder's early layers, yet this difference grows in deeper layers. Hence, to ensure that the adversarial image produces feature maps similar to those of its clean counterpart in the encoder's later stages—and ultimately reconstructs a clean image—we incorporate the FM Loss.

In addition to FM Loss, OD-Shield learns the distribution of clean images through a reconstruction loss. In particular, we use the commonly adopted L_2 loss in CNN-based reconstruction tasks, defined as

$$L_2 = \frac{1}{N_p} \sum_{i=1}^H \sum_{j=1}^W (I(i, j) - I_r(i, j))^2,$$

where I and I_r denote the original and reconstructed images, respectively, and W and H are the dimensions of I . The term N_p represents the total number of pixels.

However, L_2 processes each pixel independently, neglecting correlations among neighboring pixels and potentially sacrificing structural coherence. To address this, we adopt a patch-based SSIM loss [20], which gauges local similarity in terms of luminance, contrast, and structure, yielding a value between 0 and 1. Values closer to 1 indicate greater similarity. Since adversarial patches predominantly cause localized disruptions, we calculate SSIM within small $N \times N$ windows (e.g., 8×8 , 11×11):

$$L_{\text{SSIM}} = \frac{1}{N_p} \sum_{i=1}^H \sum_{j=1}^W (1 - \text{SSIM}(I, I_r)_{(i,j)}),$$

where $\text{SSIM}(I, I_r)_{(i,j)}$ is the SSIM value computed for the patch centered at (i, j) . By focusing on local SSIM, we specifically target structural discrepancies in smaller regions rather than across the entire image.

The reconstruction loss is thus given by

$$L_{\text{rec}}(I, I_r) = \lambda L_{\text{SSIM}}(I, I_r) + L_2(I, I_r).$$

Finally, the total loss for training OD-Shield is defined as

$$L(I, I_r, I_p) = L_{\text{rec}}(I, I_r) + L_{\text{FM}}(I, I_p).$$

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASETS

We employ three datasets—COCO [21], VisDrone [22], and Argoverse [23]—to evaluate adversarial attack scenarios. In line with autonomous driving applications, we restrict each dataset to four classes: *person*, *car*, *truck*, and *bus*.

To generate adversarial patches, we randomly select 1,000 images from each dataset. Across 420 training epochs, two distinct patch types are produced: a *Naturalistic Patch*, which subtly integrates into the surrounding context, and a *Noisy Patch*, which mimics internal image noise. These variants arise from different loss functions that yield visually distinct patterns. The remaining 5,000 images from each dataset, not used for patch generation, serve as the training set for the OD-Shield model, which is trained for 700 epochs.

Additionally, we incorporate a GAN-based naturalistic patch—derived from prior work [24]—during inference. This patch type, unseen during training, enables us to assess OD-Shield's effectiveness against previously unencountered adversarial patterns.

2) EVALUATION METRICS

a: DETECTION PERFORMANCE

We employ the widely used object detection metrics mAP₅₀ and mAP_{@50-95} to assess the efficacy of four existing pre-processing methods under adversarial attacks. The mAP₅₀ metric quantifies how accurately bounding boxes and classes are predicted, calculated as the area under the Precision–Recall curve for each class. By contrast, mAP_{@50-95} generalizes mAP₅₀ by averaging performance across multiple IoU thresholds (0.5 to 0.95), yielding a more comprehensive view of detection robustness. Both metrics collectively capture the model's accuracy and resilience when facing adversarial perturbations.

b: RECONSTRUCTION PERFORMANCE

Since OD-Shield reconstructs regions corrupted by adversarial patches using ground truth data, we employ the SSIM and Peak Signal-to-Noise Ratio (PSNR) metrics to evaluate how closely the restored image resembles the original. SSIM captures perceptual similarity across luminance, contrast, and structure, while PSNR, commonly used in both audio and image domains, quantifies the level of distortion according to

$$\text{PSNR} = 10 \times \log_{10} \left(\frac{\max(I(i, j))^2}{\frac{1}{N_p} \sum_{i=1}^H \sum_{j=1}^W (I(i, j) - I_r(i, j))^2} \right),$$

where H and W are the image dimensions, and $\max(I(i, j))$ is the maximum pixel value. Although higher PSNR values generally indicate lower distortion, they do not always align perfectly with human visual perception. Consequently, we consider both SSIM and PSNR jointly to comprehensively evaluate reconstruction quality.

3) OBJECT DETECTOR

When subjected to adversarial attacks, object detectors may misclassify critical objects. To gauge the efficacy of OD-Shield in defending against adversarial patches, we adopt YOLO-v5 [25] as the base detector. YOLO-v5 divides an image into a grid, predicting bounding box coordinates, objectness scores, and class probabilities within each cell. It then refines these predictions to yield final detection outputs. All defense methods examined in this study use YOLO-v5 as a uniform evaluation pipeline, ensuring consistent and comparable results.

B. DEFENSE PERFORMANCE COMPARISON

We conducted a series of defense experiments under three adversarial attack scenarios. In the first scenario, a *Naturalistic Patch* was employed, which is nearly indistinguishable from the original image to human observers. This patch was optimized via L_{nat} and L_{noisy} to increase its adversarial potency while preserving visual similarity to the surrounding pixels. The second scenario utilized a *Noisy Patch*, which deliberately introduces conspicuous alterations in specific image regions by iteratively applying localized weights, making the modifications plainly visible. Examples of Naturalistic Patches are shown in Figures 2(a)–(c), while Noisy Patches are depicted in Figures 2(d)–(f).

Additionally, we tested the proposed defense against GAN-based adversarial attacks to evaluate its robustness toward previously unseen patch types. Table 1 provides a summary of these findings, comparing the performance of the YOLO detector with that of alternative defenses, both with and without adversarial patches, across each dataset.

Evaluations on the COCO dataset show that YOLO achieves an mAP_{50} of 0.686 and an mAP_{50-95} of 0.470 on clean images. However, when subjected to a Naturalistic Patch attack, performance drops markedly to an mAP_{50} of 0.426 and an mAP_{50-95} of 0.071, a decrease of roughly 37%. Among the compared defenses, FQ and SAC manage to maintain some level of accuracy, whereas other methods experience further degradation. In contrast, OD-Shield exhibits the most robust performance, with an mAP_{50} of 0.519 and an mAP_{50-95} of 0.338, closely approximating its clean-image baseline. By comparison, LGS applies a full-image masking strategy, resulting in diminished object detection accuracy. A similar pattern emerges under Noisy Patch attacks, where OD-Shield again attains the highest performance (mAP_{50} : 0.542, mAP_{50-95} : 0.349) relative to FQ, JPEG, LGS, and SAC. In an additional experiment featuring a GAN-based attack [24]—representing an unseen scenario not used during OD-Shield’s training—OD-Shield likewise outperforms other defenses, achieving an mAP_{50} of 0.535 and an mAP_{50-95} of 0.332. The defense outcomes for this experiment appear in Figure 5, suggesting that OD-Shield effectively mitigates adversarial patterns even outside its training set. Furthermore, this superior performance extends to the Visdrone and Argoverse datasets, reinforcing

the conclusion that OD-Shield surpasses the four baseline methods in countering adversarial patches.

C. PRESERVATION OF THE ORIGINAL IMAGE UNDER DEFENSE

Table 2 examines how well our method defends against adversarial patch attacks (L_{nat} and L_{noisy}) across each dataset, focusing on preserving areas not covered by patches. We use PSNR as a metric, where higher values denote better preservation. Specifically, *Image PSNR* measures distortion over the entire image, while *Patch PSNR* targets only the regions where the patches are applied. The *No defense* entry reflects performance in an unprotected, adversarially attacked setting, and we compare this against FQ and LGS as baseline references.

On the COCO dataset, OD-Shield demonstrates superior performance against both L_{nat} and L_{noisy} patches. Specifically, it achieves Image PSNR and Patch PSNR scores of 25.714 and 30.896 for L_{nat} , and 25.644 and 30.505 for L_{noisy} , respectively. By contrast, the FQ method yields Image PSNR and Patch PSNR values of 24.781 and 24.797 for L_{nat} , and 24.482 and 24.608 for L_{noisy} , results that closely mirror those of the original attacked images. This indicates that while FQ minimally distorts the unpatched regions, it fails to restore compromised areas. Meanwhile, LGS produces the lowest Image PSNR (around 20) and Patch PSNR (about 24 to 25) across both L_{nat} and L_{noisy} , implying that its masking strategy degrades the image beyond the patched region. In contrast, our method isolates and reconstructs the adversarial patches while preserving the surrounding content, producing images closely aligned with the original.

Crucially, this pattern holds across the Visdrone and Argoverse datasets, confirming that our approach neutralizes the adversarial influence of patches without impairing unaltered image regions.

D. EXPERIMENTS WITH DIFFERENT MODULES AND LOSS FUNCTIONS

1) MODULES

To improve the reconstruction of adversarial patch regions by leveraging contextual cues, we integrated the SSPCAB, SSM-CTB, and SSCBAM modules into different segments of the decoder. As reported in Table 3, overall performance across these placements indicates that applying all three modules to the decoder’s final stage yields the highest reconstruction accuracy. Furthermore, the Patch SSIM metric—focused on regions affected by adversarial patches—was near 1 for all modules, underscoring their strong capabilities in recovering compromised areas.

Nevertheless, in terms of the Entire SSIM metric, which assesses consistency between patched and non-patched regions, SSCBAM achieved the best results. In addition, PSNR values, which reflect subtle pixel-level accuracy beyond immediate human perception, also improved most

TABLE 1. Defense performance comparison against other pre-processing defense methods.

Adversarial Attack	Defense Method	COCO Dataset		Visdrone Dataset		Argoverse Dataset	
		mAP ₅₀	mAP _{@50-95}	mAP ₅₀	mAP _{@50-95}	mAP ₅₀	mAP _{@50-95}
No Attack	No Defense	0.686	0.470	0.395	0.234	0.530	0.308
Naturalistic Patch	No Defense	0.426	0.271	0.226	0.120	0.354	0.159
	FQ	0.445	0.274	0.227	0.120	0.344	0.152
	JPEG	0.393	0.243	0.171	0.086	0.270	0.111
	LGS	0.351	0.215	0.180	0.091	0.309	0.130
	SAC	0.458	0.283	0.229	0.122	0.356	0.161
	OD-SHIELD	0.519	0.338	0.296	0.173	0.408	0.227
Noisy Patch	No Defense	0.422	0.261	0.187	0.096	0.296	0.130
	FQ	0.427	0.261	0.185	0.095	0.297	0.130
	JPEG	0.407	0.254	0.146	0.072	0.230	0.091
	LGS	0.361	0.225	0.148	0.073	0.272	0.106
	SAC	0.448	0.273	0.204	0.098	0.319	0.140
	OD-SHIELD	0.542	0.349	0.294	0.173	0.412	0.229
GAN-based Patch [24]	No Defense	0.493	0.317	0.331	0.123	0.487	0.221
	FQ	0.492	0.313	0.330	0.101	0.447	0.209
	JPEG	0.471	0.303	0.305	0.108	0.477	0.215
	LGS	0.452	0.293	0.271	0.095	0.315	0.187
	SAC	0.502	0.321	0.341	0.213	0.496	0.238
	OD-SHIELD	0.535	0.332	0.343	0.222	0.507	0.240

TABLE 2. PSNR scores to evaluate the preservation of the original image under defense methods.

Dataset	Attack	Defense	Image PSNR↑	Patch PSNR↑
COCO	L_{nat}	No defense	24.830	24.852
		FQ	24.781	24.797
		LGS	20.256	24.832
		OD-SHIELD	25.714	30.896
	L_{noisy}	No defense	24.623	24.652
		FQ	24.482	24.608
		LGS	20.150	23.964
		OD-SHIELD	25.644	30.505
Visdrone	L_{nat}	No defense	29.248	29.198
		FQ	29.088	29.129
		LGS	26.283	31.371
		OD-SHIELD	32.972	38.999
	L_{noisy}	No defense	30.292	30.452
		FQ	30.138	30.083
		LGS	26.045	29.109
		OD-SHIELD	32.881	38.813
Argoverse	L_{nat}	No defense	30.136	30.094
		FQ	30.105	30.067
		LGS	24.578	32.372
		OD-SHIELD	33.001	40.455
	L_{noisy}	No defense	30.808	30.827
		FQ	30.858	30.864
		LGS	25.468	31.901
		OD-SHIELD	32.887	40.074

notably when SSCBAM was applied to the decoder's last layer.

2) LOSS FUNCTIONS

We conducted an ablation study on the loss functions used for image reconstruction, using the Argoverse dataset for inference. Table 4 shows the average and peak performance over 40 runs for each loss configuration. Employing SSIM loss alone resulted in an Entire SSIM of 0.881 and a Patch SSIM of 0.993, signifying relatively modest performance. Meanwhile, using only the L2 loss slightly enhanced the Entire PSNR but yielded a comparatively low SSIM of 0.873.

By contrast, training with a combination of SSIM and L2 loss led to a higher Entire SSIM score of 0.901, outperforming each loss function used independently. Moreover, incorporating Perceptual Loss, which modifies pixel values based on differences between clean and adversarial images in a pre-trained network, further boosted reconstruction results. This addition increased Patch PSNR from 41.111 to 41.868 while also improving SSIM values. These results demonstrate that Perceptual Loss—tailored to adversarial characteristics—provides a meaningful advantage in reconstructing heavily corrupted image regions over more traditional methods.

E. TRANSFERABILITY OF OD-SHIELD

The transferability of OD-Shield highlights its ability to operate effectively even when trained on one dataset and tested on another. Table 5 demonstrates OD-Shield's performance across COCO, Visdrone, and Argoverse. For instance, when trained on COCO and evaluated on Visdrone, OD-Shield reaches an mAP₅₀ of 0.237, compared to an attacked performance of 0.187. In other words, the model, despite having been trained on COCO, still recovers a portion of

TABLE 3. Application of convolution attention modules (SSPCAB, SSMCTB, and SSCBAM) at various architecture locations in figure 3.

Module	Location	Entire SSIM	Entire PSNR	Patch SSIM	Patch PSNR
No module		0.90±0.01	30.83±0.21	0.99±0.00	39.33±0.16
SSPCAB	Early	0.90±0.01	30.44±0.12	0.99±0.01	39.55±0.05
	Middle	0.91±0.00	30.43±0.27	0.99±0.01	39.56±0.16
	Late	0.92±0.01	31.03±0.11	0.99±0.00	39.66±0.6
	All	0.90±0.01	30.07±0.36	0.99±0.00	39.47±0.08
SSMCTB	Early	0.90±0.00	30.38±0.31	0.99±0.01	39.46±0.11
	Middle	0.90±0.00	30.42±0.32	0.99±0.01	39.45±0.12
	Late	0.91±0.01	30.83±0.21	0.99±0.00	39.47±0.16
	All	0.90±0.00	30.12±0.11	0.99±0.00	39.45±0.11
SSCBAM	Early	0.92±0.00	30.18±0.31	0.99±0.01	39.45±0.11
	Middle	0.91±0.00	30.69±0.31	0.99±0.00	39.48±0.11
	Late	0.94±0.00	31.81±0.16	0.99±0.00	39.65±0.11
	All	0.93±0.00	31.05±0.31	0.99±0.01	39.44±0.11

TABLE 4. Ablation study on loss functions (SSIM, L2, and FM losses).

Training Loss	Entire SSIM	Entire PSNR	Patch SSIM	Patch PSNR
SSIM	0.881 ± 0.021	30.154 ± 1.072	0.993 ± 0.001	39.535 ± 2.415
L2	0.873 ± 0.018	32.884 ± 1.031	0.993 ± 0.001	39.435 ± 2.317
SSIM + L2	0.901 ± 0.011	32.974 ± 2.181	0.993 ± 0.002	41.111 ± 1.384
SSIM + L2 + FM	0.918 ± 0.013	33.114 ± 2.269	0.994 ± 0.001	41.868 ± 1.227

TABLE 5. mAP results of the defense demonstrating the transferability of OD-Shield.

Trained on Dataset	Dataset	mAP ₅₀		
		Clean	Attacked	Defense
COCO	COCO	0.686	0.422	0.540
	Visdrone	0.395	0.187	0.237
	Argoverse	0.530	0.296	0.396
Visdrone	COCO	0.686	0.422	0.471
	Visdrone	0.395	0.187	0.288
	Argoverse	0.530	0.296	0.339
Argoverse	COCO	0.686	0.422	0.462
	Visdrone	0.395	0.187	0.269
	Argoverse	0.530	0.296	0.411

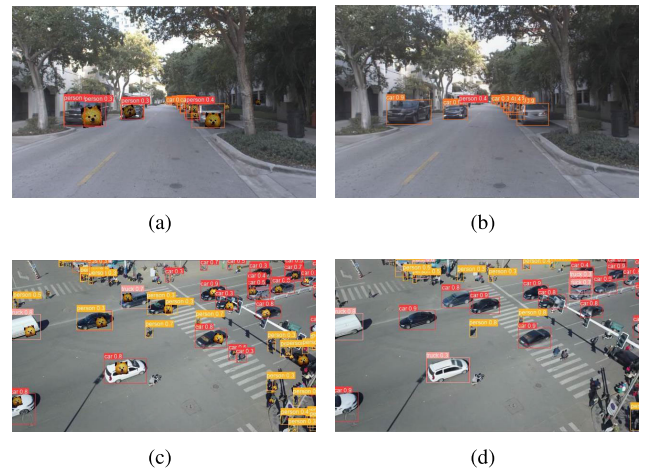
its defensive capability on Visdrone, outperforming the no-defense scenario. Similarly, on Argoverse, it achieves an mAP₅₀ of 0.396, exceeding the attacked baseline of 0.296.

Overall, these experiments confirm that OD-Shield—originally trained on COCO or any other dataset—delivers superior defense performance on previously unseen datasets, reinforcing the method’s robustness and applicability in diverse real-world settings.

V. RELATED WORKS

A. ADVERSARIAL ATTACK

Adversarial attacks typically fall into two broad categories, defined by the scope of pixel manipulation: *perturbation attacks*, which subtly modify all pixels in an image [26], [27], [28], [29], and *patch attacks*, which disrupt only a small, localized region of pixels [13], [30], [31], [32], [33]. Perturbation attacks usually introduce imperceptible noise across the entire image, directing misclassification

**FIGURE 5.** Images (a) and (c) show the object detection result on images subjected to patch attacks, while images (b) and (d) show the object detection result on the images restored through the OD-Shield methodology.

toward a targeted outcome. By contrast, patch attacks inject adversarial perturbations into specific areas, producing large, conspicuous distortions that lead to misclassification in multiple inputs. While originally conceptualized as the Adversarial Patch [30], further developments have yielded various patches capable of operating in real-world physical settings. Particularly notable are *real-like* patches [24], [34], which visually mimic legitimate objects and thus elude easy detection.

A key distinction also exists between *digital* and *physical* adversarial attacks. Digital attacks alter pixels in digital formats prior to neural network processing; although potent,

they assume direct access to the underlying media, limiting their practicality. Conversely, physical attacks involve modifying real-world objects before they are captured by a camera [35]; for example, placing adversarial stickers on a stop sign can cause autonomous driving systems to misclassify it, potentially resulting in critical misjudgments. Owing to their straightforward real-world applicability, adversarial patches have attracted growing interest as a tangible threat to neural networks.

B. ADVERSARIAL ATTACK DEFENSE

As adversarial methods diversify, the creation of robust defense strategies has become increasingly important for mitigating vulnerabilities in neural network models. In this work, we concentrate on *pre-processing* defenses, which focus on weakening adversarial effects before an image is fed into the model, with particular emphasis on *image completion* and *image smoothing*.

1) IMAGE COMPLETION

This line of work identifies adversarial regions and employs masking or inpainting to restore damaged areas. For instance, Digital Watermarking (DW) [36] uses saliency maps to localize high-gradient regions, then masks them to negate patch influence. Jedi [37] leverages entropy-based analysis to enhance localization by detecting statistical anomalies in the input, while methods such as Segment and Complete (SAC) [7] and PatchZERO [8] perform segmentation or pixel-level detection on adversarial patches to isolate and repair these regions. Additionally, RAPID [38] employs a two-stage detection mechanism based on channel-wise pooled variance to robustly identify and mask multiple adversarial patches, significantly restoring detection efficacy.

2) IMAGE SMOOTHING

Another major defensive strategy relies on smoothing to diminish adversarial signals by reducing key features in the input data. Techniques such as Feature Squeezing [39] decrease color depth or resolution, thereby counteracting high-frequency adversarial noise. JPEG Compression [40] leverages lossy compression to remove fine details—often including adversarial perturbations—while Local Gradient Smoothing (LGS) [41] and BlurNet [42] target high-gradient pixels, applying filters that blur anomalous patterns.

Although these methods show promise, each has shortcomings. Feature Squeezing and JPEG Compression can compromise critical image details and reduce overall accuracy. Local Gradient Smoothing may struggle to differentiate adversarial pixels from naturally high-gradient regions, potentially harming image fidelity. Meanwhile, Jedi and PatchZERO are often tailored to specific datasets or network architectures, restricting their broader applicability.

As adversarial attacks continue to evolve, the development of versatile and adaptive defenses remains a significant challenge. Despite notable progress in pre-processing solutions,

their specialized nature underscores the need for more holistic and flexible strategies capable of addressing a broad spectrum of adversarial scenarios.

In this study, we demonstrate that OD-Shield effectively overcomes many of these limitations by accurately detecting and reconstructing adversarially patched regions, thus neutralizing their impact. Unlike other approaches that may mask large areas or inadvertently degrade important features, our method reconstructs only the attacked pixels, ensuring minimal disruption to unaffected regions. Moreover, OD-Shield is trained on a diverse range of patch types, bolstering its resilience to variation in patch form and size.

VI. CONCLUSION

In this paper, we introduce OD-Shield, a one-stage defense mechanism against adversarial patch attacks, built on a convolutional autoencoder-based model. By learning from adversarial patch scenarios, our method reconstructs the corrupted regions to effectively neutralize such attacks. We evaluated OD-Shield on the COCO, Visdrone, and Argoverse datasets, focusing on human, car, bus, and truck objects. The results show that OD-Shield outperforms existing defenses, demonstrating its strong protective capability against adversarial patches.

ACKNOWLEDGMENT

(Byeongchan Kim and Heemin Kim contributed equally to this work.)

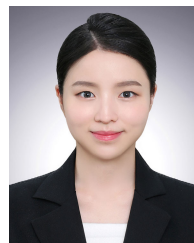
REFERENCES

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
- [2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
- [3] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*.
- [4] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector," in *Proc. Mach. Learn. Knowl. Discovery Databases*, Jan. 2019, pp. 52–68.
- [5] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 997–1005.
- [6] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1625–1634.
- [7] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14953–14962.
- [8] K. Xu, Y. Xiao, Z. Zheng, K. Cai, and R. Nevatia, "PatchZero: Defending against adversarial patch attacks by detecting and zeroing the patch," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4621–4630.
- [9] C. Xiang, S. Mahlouiifar, and P. Mittal, "PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier," in *Proc. 31st USENIX Secur. Symp.*, Jan. 2021, pp. 2065–2082.
- [10] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, "PatchGuard: A provably robust defense against adversarial patches via small receptive fields and masking," in *Proc. 30th USENIX Secur. Symp.*, Jan. 2021, pp. 2237–2254.

- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [12] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "CMT: Convolutional neural networks meet vision transformers," 2021, *arXiv:2107.06263*.
- [13] S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 49–55.
- [14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, Oct. 2015, pp. 234–241.
- [15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
- [16] N.-C. Ristea, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13566–13576.
- [17] N. Madan, N.-C. Ristea, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised masked convolutional transformer block for anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 525–542, Jan. 2024.
- [18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [19] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 3–19.
- [20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015, *arXiv:1405.0312*.
- [22] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022.
- [23] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8740–8749.
- [24] Y.-C.-T. Hu, J.-C. Chen, B.-H. Kung, K.-L. Hua, and D. S. Tan, "Naturalistic physical adversarial patch for object detectors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7828–7837.
- [25] G. Jocher et al., "Ultralytics/yolov5: V7.0-YOLOv5 SOTA real-time instance segmentation," Zenodo, Version v7.0, Nov. 2022, doi: 10.5281/zenodo.7347926.
- [26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
- [27] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.
- [28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
- [29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
- [30] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665*.
- [31] D. Karmon, D. Zoran, and Y. Goldberg, "LaVAN: Localized and visible adversarial noise," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2018, pp. 2507–2515.
- [32] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, "DPatch: An adversarial patch attack on object detectors," 2018, *arXiv:1806.02299*.
- [33] M. Lee and Z. Kolter, "On physical adversarial patches for object detection," 2019, *arXiv:1906.11897*.
- [34] L. Huang, C. Gao, Y. Zhou, C. Xie, A. L. Yuille, C. Zou, and N. Liu, "Universal physical camouflage attacks on object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 717–726.
- [35] 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," 2022, *arXiv:2209.15179*.
- [36] J. Hayes, "On visible adversarial perturbations & digital watermarking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1678–16787.
- [37] B. Tarchoun, A. B. Khalifa, M. A. Mahjoub, N. Abu-Ghazaleh, and I. Alouani, "Jedi: entropy-based localization and removal of adversarial patches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4087–4095.
- [38] H. Kim, B.-C. Kim, S. Lee, M. Kang, H. Nam, S. Park, I.-Y. Kwak, and J. Lee, "RAPID: Robust multi-patch masker using channel-wise pooled variance with two-stage patch detection," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 8, Oct. 2024, Art. no. 102188. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1319157824002775>
- [39] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," 2017, *arXiv:1704.01155*.
- [40] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. Horing Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression," 2017, *arXiv:1705.02900*.
- [41] M. Naseer, S. Khan, and F. Porikli, "Local gradients smoothing: Defense against localized adversarial attacks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1300–1307.
- [42] R. S. Raju and M. Lipasti, "BlurNet: Defense by filtering the feature maps," in *Proc. 50th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2020, pp. 38–46.



BYEONGCHAN KIM received the B.S. degree in applied statistics and the M.S. degree in statistics from Chung-Ang University, Seoul, South Korea, in 2022 and 2024, respectively. His research interests include deep learning models with audio data and audio-visual multimodal deepfake detection.



HEEMIN KIM received the B.S. degree in business administration from Dankook University, Republic of Korea, in 2020, and the M.S. degree in statistics from Chung-Ang University, Republic of Korea, in 2024. Her research interests include deep learning and machine learning for computer vision.



MINJUNG KANG received the B.S. degree in information statistics from Chungbuk National University, Republic of Korea, in 2023, and the M.S. degree in statistics from Chung-Ang University, Republic of Korea, in 2025. Her research interests include deep learning and machine learning for computer vision.



HYUNJEE NAM received the B.S. degree in statistics from Duksung Women's University, Seoul, South Korea, in 2023. She is currently pursuing the M.S. degree in statistics with Chung-Ang University, Seoul. Her research interests include deep learning and machine learning for computer vision.



JAEWOO LEE (Member, IEEE) received the B.S. and M.S. degrees in computer science and engineering from Seoul National University, Republic of Korea, in 2006 and 2008, respectively, and the Ph.D. degree in computer and information science from the University of Pennsylvania, USA, in 2017. He is currently an Associate Professor with Chung-Ang University, Republic of Korea, where he joined, in 2018. His research interests include cyber-physical systems, real-time embedded systems, and information system security.



SUNGHWAN PARK (Graduate Student Member, IEEE) received the B.S. degree in digital imaging engineering and the M.S. degree in security convergence science from Chung-Ang University, Republic of Korea, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in security convergence engineering. His research interests include federated learning, privacy-preserving machine learning, efficient data communication, and cyber-physical systems.



IL-YOUP KWAK (Member, IEEE) received the Ph.D. degree in statistics minor in computer science from the University of Wisconsin–Madison, in 2014. He was a Postdoctoral Associate with the University of Minnesota–Twin Cities, from 2014 to 2017. He was with Samsung Research, from 2017 to 2019. He is currently an Associate Professor with the Department of Applied Statistics, Chung-Ang University. His research interests include deep learning applied to audio, AI security, and statistical genetics.

...