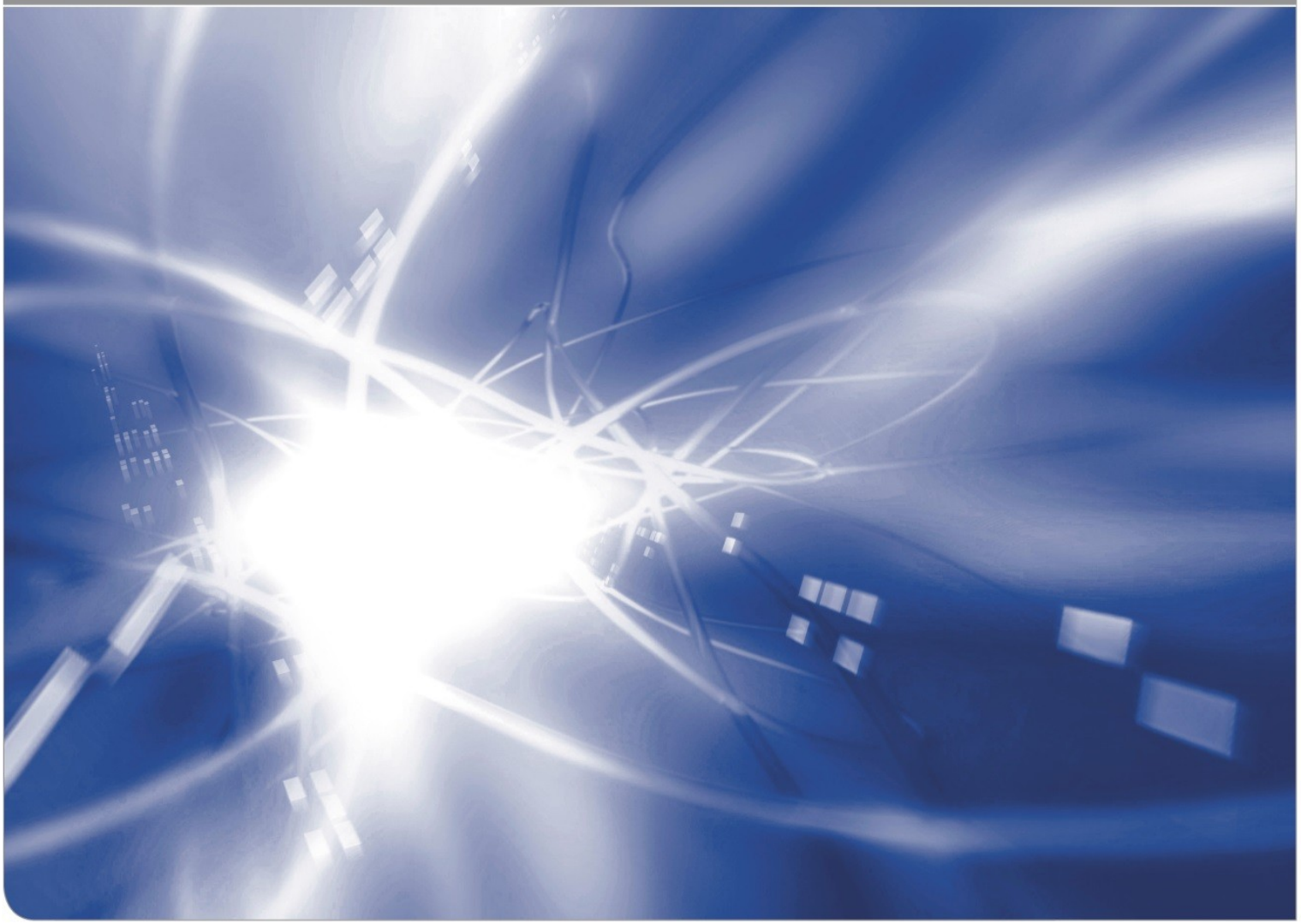
Karlsruhe Institute of Technology

On the Effectiveness of Explanation-Aware Adversarial Training

by David Hahnemann*[1], Maximilian Noppel*[1], Luan Ademi[1], Christian Wressnegger[1]

KIT SCIENTIFIC WORKING PAPERS 268



[1] KASTEL - Institute of Information Security and Dependability
* Both authors contributed equally to this work

KASTEL Security Research Labs

Am Fasanengarten 5
76131 Karlsruhe, Germany

Impressum

Karlsruher Institut für Technologie (KIT)
www.kit.edu

2026

ISSN: 2194-1629

On the Effectiveness of Explanation-Aware Adversarial Training

David Hahnemann*, Maximilian Noppel*, Luan Ademi, Christian Wressnegger
KASTEL Security Research Labs
Karlsruhe Institute of Technology
Karlsruhe, Germany

***Abstract*—With explanation-aware attacks the community has recently demonstrated the inherent lack of robustness of post-hoc XAI techniques, questioning the reliability of GradCAM and methods the like in practice. In this paper, we investigate if adversarial training, as the go-to defense against classical adversarial examples, is also able to defend against *explanation-aware* adversarial examples. To this end, we investigate the effectiveness of three adversarial-training techniques, dedicated explanation-aware extensions for each of them, and two existing defenses that share similarities with adversarial training under the hood. Surprisingly, we find that vanilla adversarial training is already very effective against explanation-aware adversarial examples, even outperforming dedicated defenses. However, *all* existing techniques also lower the clean accuracy by 5–22 percentage points. In addition, we provide an extensive transferability study of AT-based defenses across five different types of explanation-aware attacks and across different target explanations. Our findings suggest that AT-defenses transfer, meaning that defenders do not need to anticipate the exact attack scenario to apply adversarial training effectively, leveling the playing field in this regard.**

***Index Terms*—Explanation-Aware Attacks, XAI**

1. Introduction

Machine-learning models are vulnerable to carefully crafted inputs at inference time, so-called adversarial examples [10, 21, 35, 36, 57]. With imperceptible perturbations, malicious actors can manipulate a model's predictions arbitrarily. The most prominent and conceptually straightforward defensive technique against such attacks is *adversarial training* (AT) [35, 64, 67]. Here, the developers attack their own model, incorporate the resulting adversarial examples into the training data, and label them with the respective ground-truth label. Unfortunately, this approach mainly works against the attack strategies anticipated at training time, i.e., against known attack types.

With explanation-aware attacks [38], the community has recently discovered a new family of attacks. For instance, explanation-aware adversarial examples influence a model's predictions *and* its explanations at inference time [1, 16, 28, 39, 58, 62, 68] without a human-visible clue in input space. Consequently, (post-hoc) explanation methods might show false reasons for predictions in adversarial environments. As explanations have historically been motivated via trust and trustworthiness, explanation-aware attacks raise questions about whether this trust is justified. Further, manipulated explanations can even reinforce wrong decision-making and thus make things worse than not having explanations in the first place.

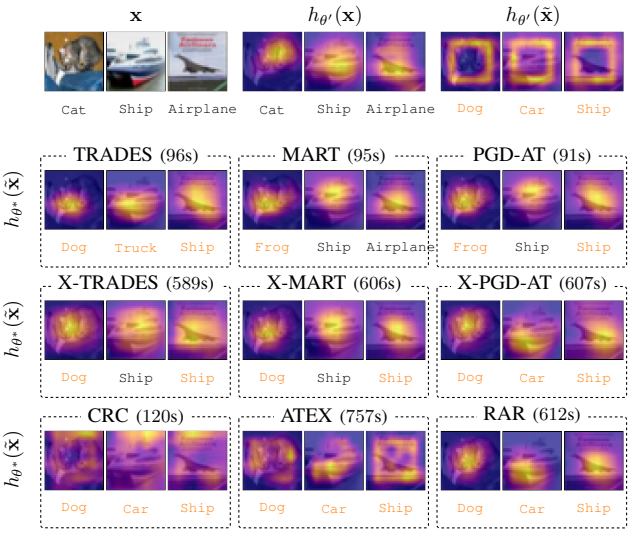


Figure 1: Qualitative examples of the effectiveness of different defensive techniques against explanation-aware attacks that alter both the prediction and the explanation. Fooled predictions are highlighted in orange color. While the original model is vulnerable to such attacks, most adversarial training techniques are able to defend against them. However, the prediction-only adversarial training techniques (TRADES, MART, and PGD) are significantly faster (cf. training time in parentheses). We provide further examples in Fig. 6 in Section A.

In this paper, we investigate the effectiveness of different defenses against explanation-aware adversarial examples attacking the GradCAM explanation method [47]. Fig. 1 provides qualitative examples for each of the considered defenses in one specific attack scenario, namely dual attacks with a square as the target explanation.

We evaluate three AT techniques, PGD-AT [35], TRADES [67], and MART [64], that have originally been proposed to defend against vanilla adversarial examples. In line with a recent taxonomy of explanation-aware attacks [38], we denote such *vanilla* adversarial examples as *prediction-only attacks*, i.e., the adversary ignores the

*David Hahnemann and Maximilian Noppel contributed equally.

effects on the explanation. We evaluate how effective these three AT defenses are against explanation-aware attacks, even though they do not consider the effects on the explanations during fine-tuning. Additionally, we present explanation-aware extensions for each of the three AT techniques, which we denote as X-PGD-AT, X-TRADES, and X-MART respectively.

In addition, we evaluate two defenses, ATEX [59] and CRC [29], that have both been specifically designed to mitigate explanation-aware attacks. Similar to AT, both fine-tune on modified samples. In contrast to AT, however, neither of them uses worst-case perturbations, i.e., adversarial examples.

Another strain of work [13, 27, 46, 52] has made first steps toward explanation-aware AT. However, they have only evaluated a very constrained set of adversarial goals and explanation methods, each heavily relying on gradients. Of them, we evaluate RAR [13] as the most general state-of-the-art approach, and FAR [27], which happens to be equivalent to X-PGD-AT.

Explanation-aware adversarial training introduces three key complexities compared to prediction-only (vanilla) AT:

(1) Explanation-aware adversaries pursue independent objectives for two targets: predictions and explanations.

(2) While predictions are elements of a finite set of classes, explanations are high-dimensional objects (typically one relevance score per input feature). A categorization into untargeted, targeted, and semi-targeted attacks exists, but the number of potential target explanations is virtually infinite (in $[0,1]^d$ for d-dimensional inputs). This vast space of possible targets further complicates both anticipating and mitigating attacks.

(3) Moreover, there exist numerous ($\geq$150) explanation methods [9, 22], from which the defender may select one or multiple methods, potentially at random. This diversity of explanation methods, in turn, requires the adversary to anticipate which explanation method(s) the defender uses.

For our study, we focus on the popular explanation method GradCAM [47], which is particularly vulnerable to explanation-aware attacks [5, 19, 39, 40]. Moreover, it is applicable to any convolutional neural network (CNN) architecture and computationally efficient compared to other explanation methods. Even in this setting with a fast explanation method the computational costs for explanation-aware adversarial training are substantially higher than vanilla AT (cf. Section 5).

For our experiments, we reuse the target explanation "square" as introduced in related work [19, 39, 40] to ensure comparability. The "square" target explanation is easily recognizable (cf. Fig. 1 top-right) and very different from benign explanations yielded for common datasets, making it a suitable candidate to investigate attacks and defenses.

Lastly, we analyze the transferability between the anticipated and the actual attack scenario and find that here, too, vanilla AT achieves performant results. In another study, we evaluate the transferability between two maximally different target explanations, highlighting either the right half or the left half of the image as important. We find that defenses based on adversarial training again transfer well across different target explanations.

In summary, we make the following contributions:

- **Explanation-Aware Adversarial Training.** We systematize explanation-aware adversarial training and present extensions for the popular AT techniques, PGD-AT [35], TRADES [67], and MART [64]. We find that FAR [27] coincides with explanation-aware PGD-AT but has never been evaluated for GradCAM up to now.
- **Evaluation of Existing Defenses.** We extensively evaluate existing vanilla AT approaches to assess their effectiveness against explanation-aware adversarial examples. Further, we reevaluate three existing defenses, ATEX [59], CRC [29], and RAR [13] on GradCAM, that have been specifically proposed to mitigate explanation-aware attacks.
- **Transferability Studies.** We extensively investigate the transferability of anticipations made by the defender. First, we investigate how well the defenses transfer to other explanation-aware attack scenarios. Second, we analyze whether the defender needs to anticipate the adversary's target explanation.

2. Background

We briefly present the notation used in the paper, before outlining the fundamentals of adversarial machine learning in Section 2.1, providing the necessary background for the remainder of the paper. Then, we introduce explainable machine learning and explanation-aware attacks in Sections 2.2 and 2.3. Finally, we review regularization methods to stabilize explanations in Section 2.4.

Notation. A prediction function $f_\theta : \mathcal{X} \to [0,1]^C$ provides class probabilities (soft-labels), with C being the number of classes and $\mathcal{X} \subseteq \mathbb{R}^d$ being the input space with d dimensions. Based on f_θ, the decision function $F_\theta : \mathcal{X} \to [1,\dots,C]$ provides hard-labels: $F_\theta := \mathbf{x} \mapsto \arg\max_c f_\theta(\mathbf{x})_c$. The classifier is parameterized via learned weights and biases θ, also denoted as "the model." The initial learning happens on a dataset D consisting of original inputs $\mathbf{x}_i \in \mathcal{X}$ and corresponding ground-truth labels $\hat{y}_i \in [1,\dots,C]$. Further, an explanation method $h_\theta : \mathcal{X} \to \mathcal{E}$ produces explanations $\mathbf{r}$ in the explanation space $\mathcal{E} \subseteq \mathbb{R}^d$ for the given model θ and input $\mathbf{x}$. The explanation method assigns importance scores to individual features, e.g., pixels. We perform fine-tuning steps on original (not robustified) models θ' and denote the resulting robustified models as θ^*. We use $f_{\theta'}$, f_{θ^*}, $F_{\theta'}$, F_{θ^*}, $h_{\theta'}$ and h_{θ^*} accordingly.

2.1. Adversarial Machine Learning

Various researchers have investigated how to perturb a sample so that a model predicts either any other wrong class (untargeted attack) or a specific wrong class (targeted attack) [8, 10, 15, 21, 35–37, 49, 57, 67]. Such slightly perturbed samples are known as *adversarial examples*. In the image domain, they are typically restricted to be within the ℓ_p-norm ball around their original sample $\mathbf{x}$, a proxy for the imperceptibility of the perturbation. In other domains, more specific restrictions apply [14, 42].

Defenses against Adversarial Examples. Robustifying a model is often performed by incorporating known attacks into the training process as an additional fine-tuning step after pre-training [26, 34, 35, 48, 64, 67], known as "adversarial training" (AT). In other words, defenders attack their own model during training and add these adversarial examples to the training data, labeling them with the corresponding ground truth labels. Consequently, the model learns to classify adversarial examples correctly and, ideally, generalizes to unseen attacks. The model recognizes characteristics of worst-case perturbations, making such perturbations less effective. However, since worst-case perturbations are only approximated by the attack algorithm during training, stronger or novel attack strategies often still fool robustified models [11, 24].

For our work, we consider the three popular AT techniques PGD-AT [35], TRADES [67], and MART [64]. The particularities of each are introduced in Section 3.6.

2.2. Explainable Machine Learning

The success of today's AI systems is founded on deep neural networks with massive amounts of learned parameters. This overabundance of parameters enables highly non-linear behavior (as required for complex tasks), but at the same time makes it impossible for humans to understand the underlying decision-making processes. As a remedy, researchers have suggested numerous explanation methods [e.g., 4, 12, 33, 44, 47, 51] that compress a model's complex behavior into a relatively small set of relevance scores, typically one per input feature. These explanations are commonly visualized as heat maps on top of inputs, visualizing "where the model looks," as can be seen in Fig. 1.

GradCAM. In this work, we focus on the commonly used explanation method GradCAM [47], which itself is based on class activation maps (CAM) [69]. Both methods aggregate the activation maps at the last convolutional layer and upscale the result to the input size, so that the explanation for class c is

$$h_\theta(.)_c := \text{upscale}\left(\sum_k w_k \mathbf{a}_k\right) ,$$

where $\mathbf{a}_k$ is the activation of the k-th channel and w_k is the weight for channel k. Further, $a_{k,i}$ denotes the activation of the i-th neuron in the k-th channel. The peculiarity of GradCAM compared to CAM is that it weighs the individual activation maps by the average gradients received from the later layers:

$$w_k := \sum_i^{|\mathbf{a}_k|} \frac{\partial f_\theta(.)_c}{\partial a_{k,i}} \Big/ |\mathbf{a}_k| \ .$$

The predecessor CAM, in turn, is only applicable to a particular set of CNNs using global average pooling, while GradCAM is applicable to *any* CNN. The activation maps can be obtained on-the-fly in the forward path and the gradient propagation is only required for the later layers. Thus, GradCAM is particularly fast compared to other explanation methods—even faster than the well-investigated Simple Gradients method [51].

While popular, GradCAM has also been proven to be particularly vulnerable in adversarial environments [19, 39, 40]. This vulnerability and lightweight computation make it an ideal candidate to investigate defense techniques, as we do in this paper. Note that even with this fast explanation method the computational costs for evaluating explanation-aware adversarial training are high. Therefore, we had to leave the investigation of additional explanation methods to future work.

2.3. Explanation-Aware Attacks

Explanation-aware attacks extend traditional prediction-only (PO) attacks to include an additional system output: the explanation [6, 38, 41, 61]. For example, Dombrowski et al. [16] manipulate the input, similar to adversarial examples, in such a way that the explanation depicts the text "This explanation is manipulated" while the prediction remains unchanged. Similar attacks have been proposed in related works [3, 20, 28, 32, 53, 58].

Other adversarial goals exist as well [1, 18, 32, 39, 55, 63, 68]; for instance, some attacks target both the prediction *and* the explanation of the input. On a high level, such manipulations of the explanation throw a red herring at the analyst—the receiver of the explanation. This can be done either to disguise an ongoing attack against the prediction or to increase the effort required to analyze the sample. Eventually, the confused analyst might accept the model's (wrong) prediction.

Following this chain of thought, explanation-aware attacks are categorized based on their effects on two targets: the prediction and the explanation [38]. For each target, the adversary can choose to *ignore* it, intentionally *preserve* it, or *alter* it. When altering a target, the adversary can either aim to make it maximally different from the original explanation (untargeted), or match a desired target explanation (targeted). To emphasize which target is (un)targeted, we explicitly denote *prediction-(un)targeted* or *explanation-(un)targeted* in the remainder of this paper.

Adversarial Goals. A vanilla adversarial example equals a *prediction-only* (PO) attack, where only the prediction is altered (targeted or untargeted) while the explanation is ignored. As such, vanilla adversarial examples are not considered explanation-aware. An *explanation-preserving* (EP) attacker aims to alter the prediction (in a targeted or untargeted way) while preserving the explanation. Conversely, a *prediction-preserving* (PP) attacker alters the explanation (also in a targeted or untargeted way) while preserving the prediction. *Dual* (D) attacks alter both targets simultaneously [68], i.e., there are four possible types of dual attacks. Lastly, *explanation-only* (EO) attackers alter the explanation and ignore the prediction. In that sense, explanation-only attacks are the counterparts to vanilla adversarial examples. We summarize these adversarial goals in Table 1.

In this work, we focus exclusively on prediction-untargeted attacks and leave the investigation of prediction-targeted attacks as future work. This focus is in line with the research field on vanilla adversarial training [35, 64, 67]. However, our methodology can naturally be transferred to prediction-targeted attacks also.

TABLE 1: The first row represents prediction-only attacks (PO), which correspond to vanilla adversarial examples. The four explanation-aware adversarial goals, in turn, are displayed below.

Name	Abbrv.	Prediction	Explanation
Prediction-Only	PO	alter	ignore
Explanation-Preserving	EP	alter	preserve
Explanation-Only	EO	ignore	alter
Prediction-Preserving	PP	preserve	alter
Dual	D	alter	alter

In contrast, for explanations we consider targeted *and* untargeted scenarios. Textually, we distinguish them through superscripts ($PP^\star$, $D^\star$, $PP^\square$, and $D^\square$). A star $\star$ denotes explanation-untargeted, whereas a square $\square$ denotes explanation-targeted with the target explanation being a square, as introduced by related work [39, 40], and illustrated in Fig. 1 in the top-right corner.

2.4. Explanation Stability Regularization

Several studies have explored regularization techniques to enhance the stability of explanations [17, 30, 43, 45, 66]. The underlying principle of explanation stability posits that *similar inputs should yield similar explanations*. For example, Ross and Doshi-Velez [45] have regularized gradient magnitudes to achieve smoother decision surfaces. This approach has been subsequently extended to tabular data [43, 66]. An alternative smoothing technique replaces the ReLU activation function with the SoftPlus function and regularizes the Hessian matrix [17].

Two recent works have extended their investigations beyond stability to evaluate robustness against adversarial manipulation [29, 59]. However, neither approach employs worst-case perturbations (attacks) during the fine-tuning process, placing them outside the framework of adversarial training. We describe both methods below and evaluate their effectiveness in subsequent sections.

Adversarial Training on EXplanations (ATEX) [59] employs a fine-tuning procedure analogous to adversarial training, but uses perturbed samples that are not adversarial examples. Instead, perturbations are generated in directions orthogonal to the sample's gradient. Due to its strong dependence on model gradients, ATEX has been primarily evaluated for the gradient-based explanation method Simple Gradients [51].

Cosine Robust Criterion (CRC) [29] measures the cosine similarity and ℓ_2 norm between explanations of clean samples and uniformly perturbed samples within an ℓ_∞-norm ball. This method aims to mitigate prediction-preserving attacks (PP) with minimal computational overhead (cf. Section 5). However, random perturbations fundamentally differ from adversarial (worst-case) perturbations. CRC has been evaluated on multiple explanation methods, including Simple Gradients [51], Gradients×Input [50], Guided Backpropagation [54], and LRP [4], but not on GradCAM.

In summary, stability-oriented approaches regularize explanations using specific noise patterns or structural constraints. In contrast, our work focuses on robustness under worst-case perturbations, specifically examining the *vulnerability* of explanations to adversarial manipulation.

3. Explanation-Aware Adversarial Training

Early works on adversarial training have already formalized the defender's problem as a nested optimization problem, where the inner maximization represents the adversary and an outer minimization represents the defender [26, 34, 48]. However, they have evaluated relatively weak adversarial strategies to approximate the attack samples. Similarly, current defensive approaches for explanation robustness rely on specifically crafted [1] or random perturbations [59] rather than strong adversarial attacks. PGD adversarial training [35] considers the same nested optimization, but uses effective adversarial examples for the inner maximization.

Adversarial training to counter explanation-aware adversarial examples has experienced a similar progression from using weak approximations of attacks [1, 59] to strong adversarial strategies [13, 27, 46, 52]. We investigate the effectiveness of these variants, generalize them to *explanation-aware adversarial training*, and implement it in different manifestations in line with well-established AT variants against prediction-only (PO) attacks [35, 64, 67].

We begin this section by introducing the core idea behind explanation-aware adversarial training in Section 3.1, before providing a general formalization in Section 3.2. Then, in Section 3.3, we provide a brief overview of the involved distance metrics, before introducing metrics to evaluate the robustness of a single model in Section 3.4. Section 3.5 introduces how explanation-aware adversarial examples are generated. After that, we detail our extensions to the three well-established adversarial training methods: PGD-AT, TRADES, and MART in Section 3.6. Finally, we introduce how RAR [13] works and how it relates to our formalization in Section 3.7.

3.1. Core Idea

Explanation-aware adversarial training is surprisingly straightforward. Instead of training on prediction-only adversarial examples (PO), we fine-tune the model on explanation-aware adversarial examples. In this work, we evaluate five goals: EP, $PP^\square$, $PP^\star$, $D^\square$, and $D^\star$.

Penalizing Bad Explanations. Using explanation-aware adversarial examples is not enough, though. "Bad" explanations need to be penalized during the defensive fine-tuning step. Fortunately, many explanation methods, including GradCAM, are differentiable, and it is sufficient to compare two explanations in the loss function using a differentiable distance metric $d_\mathcal{E}$ in explanation space $\mathcal{E}$, making the loss function explanation-aware.

Ground-Truth Explanations. Similar to vanilla AT, which relies on ground-truth labels, our extensions would require ground-truth explanations. Unfortunately, human-provided annotations are rarely available, and it is an ongoing debate whether "ground-truth explanations" exist at all [7]. Using human-provided annotations as ground-truth explanations could force the model to learn on the basis of our

human ontology, which, in principle, is not desired. For the sake of simplicity, we instead want the robustified model's explanations of clean samples to remain similar to the original model's explanations of the same clean samples. These clean explanations then serve as the desired explanations in the adversarial training scheme.

3.2. A Formalization of Explanation-Aware AT

Adversarial training is defined as a nested optimization problem, where an inner maximization represents the adversary and an outer minimization represents the defender [26, 34, 35, 48]:

$$\min_{\theta} \mathop{\mathbb{E}}_{(\mathbf{x},y)\in D} \left[\max_{\delta} \mathcal{L}(\theta, \mathbf{x}+\delta, y) \right] ,$$

where $\mathcal{L}$ represents a general loss function (e.g., the cross-entropy loss) that is minimized over a number of fine-tuning steps. In each iteration, the algorithm first attacks the current model θ by solving the inner maximization and then updates the model's parameters through a gradient-descent step. While PGD-AT uses the same loss in the inner maximization and the outer minimization, TRADES [67] introduces a second regularization term for the outer minimization, and MART [64] extends the outer minimization further. Strictly speaking, in those newer methods the inner and the outer loss functions thus are different.

In our work, this differentiation between the inner loss function and the outer loss function is crucial. Therefore, we denote the inner maximization loss function as L and the outer minimization loss function as $\mathcal{L}$:

$$\min_{\theta} \mathop{\mathbb{E}}_{(\mathbf{x},y)\in D} \left[\mathcal{L}\big(\theta, \mathbf{x}+\max_{\delta} L(\theta, \mathbf{x}+\delta, y), y\big) \right] .$$

We are particularly interested in settings where either the inner loss L or the outer loss $\mathcal{L}$ is explanation-aware, or both. For instance, the outer minimization loss is explanation-aware for our explanation-aware extensions of PGD-AT, TRADES, and MART, while the outer minimization loss of the respective original works is not. The inner maximization, in turn, is explanation-aware for all considered adversarial goals except for PO, which equals vanilla adversarial examples.

3.3. Distance Metrics $d_{\mathcal{E}}$

In the next paragraphs, we will regularly compare two explanations with a generic distance measure $d_{\mathcal{E}}$. This function $d_{\mathcal{E}}$ can be any differentiable distance function in the explanation space $\mathcal{E}$. In our experiments, we use the mean squared error (MSE). Others have also experimented with the cosine similarity [64], the ℓ_2 and the ℓ_1 norm [64], Pearson correlation coefficient (PCC), and the structural similarity index (SSIM) [40, 65]. In Section A, we provide additional evaluations with the PCC and SSIM metrics, and we illustrate differences between these metrics in Fig. 7.

3.4. Metrics for a Single Model

We now describe the metrics used to evaluate a model's robustness. For its predictive performance, we measure the clean top-1 accuracy ($\text{Acc}^{\text{clean}}$) on benign samples and the robust top-1 accuracy (Acc^{rob}) on adversarial examples. The explanation robustness is evaluated with the following three key metrics:

1) *Explanation Forgeability*, measuring how close manipulated explanations come to the target explanation.
2) *Fooling Rate* as the fraction of successfully fooled explanations via thresholds on the forgeability.
3) *Explanation Vulnerability*, indicating how much an attacker can make manipulated explanations deviate from their original explanation.

All metrics are formalized on the basis of clean samples $\mathbf{x}$ and (implicitly) their corresponding adversarial examples, which we denote as $\tilde{\mathbf{x}}$.

Explanation Forgeability (For). The distance between the manipulated explanation and the adversary's target explanation is measured as:

$$For_{\theta} := \mathop{\mathbb{E}}_{(\mathbf{x},\cdot)\in D} \left[d_{\mathcal{E}}(h_{\theta}(\tilde{\mathbf{x}}), \mathbf{r}^t) \right] .$$

This metric measures directly how well the adversary can reach the chosen target explanation $\mathbf{r}^t$. *Note that, for explanation-untargeted adversaries ($PP^{\star}$,$D^{\star}$) the forgeability is undefined as there exists no target explanation* $\mathbf{r}^t$.

Fooling Rate (FR). As the interpretation of distance measures, such as the MSE, is difficult, we augment the forgeability metric with the *fooling rate* (FR) [19]. For a given threshold τ, the fooling rate is determined as

$$FR_{\theta} := \mathop{\mathbb{E}}_{(\mathbf{x},\cdot)\in D} \left[\mathbf{1}_{d_{\mathcal{E}}(h_{\theta}(\tilde{\mathbf{x}}), \mathbf{r}^t) < \tau} \right] .$$

We establish the thresholds τ individually for each target explanation through an empirical user study. To determine appropriate thresholds, we sample explanations from randomly selected test samples using two models with distinct robustness characteristics: one original (non-robustified) model and one ATEX-robustified model.[1] Seven independent annotators classify each explanation as either "close to the target explanation" (fooled) or "too distant from the target explanation" (not fooled).

We assess the inter-annotator reliability using Fleiss's Kappa, obtaining high agreements: $\kappa = 0.7318$ for the square target explanation, $\kappa = 0.6976$ for the right-half target, and $\kappa = 0.4452$ for the left-half target (cf. Section 6.1). The ground-truth label for each explanation is determined via majority voting across the seven annotations. Subsequently, we conduct a receiver operating characteristic (ROC) analysis across threshold values in the range $[0.02, 0.20]$ with increments of 0.001, selecting the threshold that maximizes the geometric mean of the true positive rate (sensitivity) and true negative rate (specificity):

$$\tau := \arg\max_{t} \sqrt{TPR(t) \cdot TNR(t)} .$$

The selected thresholds are 0.084 for the square target explanation, 0.141 for the right half, and 0.132 for the left half. *Note that the fooling rate is in addition to the explanation forgeability, and that many existing works do not provide fooling rates at all [25, 39, 40].*

1. The selection of ATEX is motivated by its high variance in explanation forgeability, which ensures a representative distribution of samples across the robustness spectrum.

Explanation Vulnerability (Vul). The explanation vulnerability measures how much the explanations of adversarial examples deviate from the clean samples' explanations. As such, the vulnerability is particularly useful for evaluating explanation-untargeted attacks. It is defined as

$$Vul_\theta := \mathop{\mathbb{E}}_{(\mathbf{x},\cdot)\in D} \Big[d_\mathcal{E}\big(h_\theta(\tilde{\mathbf{x}}), h_\theta(\mathbf{x})\big) \Big] \ .$$

3.5. Explanation-Aware Adversarial Examples

Explanation-aware adversarial examples are generated by optimizing bi-objective loss functions, consisting of a prediction loss L_{pred} and an explanation loss L_{expl}, combined with a weighting factor γ:

$$L = (1-\gamma)\cdot L_{\text{pred}} + \gamma \cdot L_{\text{expl}} \ .$$

We discuss the prediction loss part L_{pred} in further detail in Section 4.2, and summarize the loss functions of the individual adversarial goals in Table 2. Note that we present them once with the distance metric and once formalized using the metrics described above. Observe how the forgeability is applied for explanation-targeted attacks, while the vulnerability is used for explanation-untargeted attacks.

To optimize the loss function we use projected gradient descent (PGD) [35] as an attack algorithm:

$$\mathbf{x}^{(i+1)} = \mathop{\text{proj}}_{\mathcal{B}_\epsilon(\mathbf{x})} \Big(\mathbf{x}^{(i)} - \eta \cdot \text{sgn}(\nabla_{\mathbf{x}^{(i)}} L) \Big) \ ,$$

where η is the learning rate, $\text{proj}_{\mathcal{B}_\epsilon(\mathbf{x})}$ is the projection onto the ℓ_p-ball of radius ϵ around the original sample $\mathbf{x}$, and $\text{sgn}(\cdot)$ is the sign function. We always limit the ℓ_∞-norm by $\epsilon = {}^8\!/_{255} = 0.031$. For the PO attacks a learning rate of $\eta = {}^2\!/_{255} = 0.0078$ is used as suggested by related work [35]. The PGD attack algorithm is used for all attacks in this work.

3.6. Our Explanation-Aware AT Extensions

In the following, we introduce our explanation-aware extensions for the three commonly used AT techniques PGD-AT [35], TRADES [67], and MART [64]. The three build up on each other, i.e., each extending over the previous by adding another aspect to the loss. We extend the underlying ideas of each addition to explanations.

TABLE 2: Explanation-aware attack scenarios and their loss functions, once complete and once in terms of our metrics as defined above.

Att.	Loss Function
PO	L_{pred}
EP	$(1-\gamma)\cdot L_{\text{pred}} + \gamma \cdot d_\mathcal{E}(h_\theta(\tilde{\mathbf{x}}), h_\theta(\mathbf{x}))$ $(1-\gamma)\cdot L_{\text{pred}} + \gamma \cdot Vul_\theta$
$PP^\square$	$(1-\gamma)\cdot \big(1-L_{\text{pred}}\big) + \gamma \cdot d_\mathcal{E}(h_\theta(\tilde{\mathbf{x}}), \mathbf{r}^t)$ $(1-\gamma)\cdot \big(1-L_{\text{pred}}\big) + \gamma \cdot For_\theta$
$PP^\star$	$(1-\gamma)\cdot \big(1-L_{\text{pred}}\big) + -\gamma \cdot d_\mathcal{E}(h_\theta(\tilde{\mathbf{x}}), h_\theta(\mathbf{x}))$ $(1-\gamma)\cdot \big(1-L_{\text{pred}}\big) + g\gamma \cdot Vul_\theta$
$D^\square$	$(1-\gamma)\cdot L_{\text{pred}} + \gamma \cdot d_\mathcal{E}(h_\theta(\tilde{\mathbf{x}}), \mathbf{r}^t)$ $(1-\gamma)\cdot L_{\text{pred}} + \gamma \cdot For_\theta$
$D^\star$	$(1-\gamma)\cdot L_{\text{pred}} + -\gamma \cdot d_\mathcal{E}(h_\theta(\tilde{\mathbf{x}}), h_\theta(\mathbf{x}))$ $(1-\gamma)\cdot L_{\text{pred}} + -\gamma \cdot Vul_\theta$

X-PGD-AT. Madry et al. [35] have evaluated the idea of adversarial training with sufficiently strong attacks using the cross-entropy loss of the adversarial examples:

$$\mathcal{L}_{PGD-AT} = CE(f_\theta(\tilde{\mathbf{x}}), \hat{y}) \ .$$

They directly penalize differences between the model's predictions on adversarial examples and the respective ground-truth labels. Applying this principle to explanations, we minimize the distance between the adversarial explanations and the corresponding original explanations:

$$\begin{aligned}\mathcal{L}_{X-PGD-AT} =& \mathcal{L}_{PGD-AT} \\ &+ \lambda_e \cdot d_\mathcal{E}(h_\theta(\tilde{\mathbf{x}}), h_{\theta'}(\mathbf{x})) \ .\end{aligned}$$

The trade-off between predictions and explanations can be adjusted by the hyperparameter λ_e (e for explanation). Interestingly, X-PGD-AT is equivalent to FAR [27]when generalized for GradCAM.

X-TRADES. Zhang et al. [67] exploit the fact that there exists a trade-off between the robustness and the natural accuracy of the resulting model [60]. Hence, the prediction error can be decomposed into the sum of the natural classification error and a boundary error, where the first is the probability of misclassifying a clean sample and the latter is the probability that a correctly classified sample is close enough to the decision boundary and thus can be successfully attacked. TRADES [67] weighs these two errors by a factor λ_p (p for prediction):

$$\mathcal{L}_{TRADES} = CE(f_\theta(\mathbf{x}), y) + \lambda_p \cdot \text{KL}(f_\theta(\mathbf{x}) || f_\theta(\tilde{\mathbf{x}})) \ ,$$

where KL is the Kullback-Leibler divergence.

We apply this intuition to explanations by adding two loss terms for explanations. The first term penalizes the distance between benign explanations in the current model $h_\theta(\mathbf{x})$ and the original explanations $h_{\theta'}(\mathbf{x})$. The second term formalizes the distance between the benign and manipulated explanations:

$$\begin{aligned}\mathcal{L}_{X-TRADES} =& \mathcal{L}_{TRADES} \\ &+ \lambda_{e,1} \cdot d_\mathcal{E}\big(h_\theta(\mathbf{x}), h_{\theta'}(\mathbf{x})\big) \\ &+ \lambda_{e,2} \cdot d_\mathcal{E}\big(h_\theta(\mathbf{x}), h_\theta(\tilde{\mathbf{x}})\big) \ .\end{aligned}$$

X-MART. Wang et al. [64] have further distinguished between misclassified clean samples and correctly classified clean samples by applying different weighting factors for both sets of samples. The loss function is as follows:

$$\begin{aligned}\mathcal{L}_{MART} =& BCE(f_\theta(\tilde{\mathbf{x}}), y) \\ &+ \lambda_p \cdot KL\big(f_\theta(\mathbf{x}) || f_\theta(\tilde{\mathbf{x}})\big) \cdot (1 - f_\theta(\mathbf{x})_{\hat{y}}) \ ,\end{aligned}$$

where BCE is the binary cross-entropy loss and $f_\theta(\mathbf{x})_{\hat{y}}$ is the soft-label of the ground-truth class $\hat{y}$. The importance of the Kullback-Leibler divergence KL is now weighted by the model's inverse probability of the ground-truth class. We extend this concept to explanations as follows:

$$\begin{aligned}\mathcal{L}_{X-MART} = \mathcal{L}_{MART} \\ &+\lambda_{e,1} \cdot d_\mathcal{E}\big(h_\theta(\tilde{\mathbf{x}}), h_{\theta'}(\mathbf{x})\big) \\ &+\lambda_{e,2} \cdot d_\mathcal{E}\big(h_\theta(\mathbf{x}), h_\theta(\tilde{\mathbf{x}})\big) \\ &\cdot \big(1 - d_\mathcal{E}(h_\theta(\mathbf{x}), h_{\theta'}(\mathbf{x}))\big) \ .\end{aligned}$$

3.7. Reformulation of Related Work

In addition to formulating new adversarial training regimes, we draw connections to related work as well. Generalizing RAR [13] shows the similarity to X-TRADES. Note that it is *not* equivalent, though.

Robust Attributional Regularization (RAR) [13] improves the robustness of Integrated Gradients explanations $IG(\cdot,\cdot)$ [56] through the following loss function:

$$\begin{aligned}\mathcal{L}_{RAR} =& CE(f_\theta(\tilde{\mathbf{x}}), \hat{y}) \\ &+ \lambda_e \cdot ||IG(\mathbf{x}, \tilde{\mathbf{x}})||_1 \ .\end{aligned}$$

Here, $IG(\mathbf{x}, \tilde{\mathbf{x}})$ is the Integrated Gradients explanation of the adversarial example $\tilde{\mathbf{x}}$ with respect to the original sample $\mathbf{x}$. Specifically, $IG(\mathbf{x}, \tilde{\mathbf{x}})$ is defined as the path integral of the gradients along a line path from $\mathbf{x}$ to $\tilde{\mathbf{x}}$. The authors argue that Integrated Gradients attributes the changes between the loss at $\mathbf{x}$ and at $\tilde{\mathbf{x}}$ to the input features.

To transfer this idea to other explanation techniques, we take the difference between the two explanations:

$$\begin{aligned}\mathcal{L}_{RAR} =& CE(f_\theta(\tilde{\mathbf{x}}), \hat{y}) \\ &+ \lambda_e \cdot d_{\mathcal{E}}\big(h_\theta(\mathbf{x}), h_\theta(\tilde{\mathbf{x}})\big) \ ,\end{aligned}$$

which is equivalent to the X-TRADES loss function with $\lambda_{e,1}$ set to 0.

Also, we observe that this research strain comes with stability in mind. The explanations of clean samples and adversarial examples should be equivalent. This approach does not ensure that the explanations are useful. The loss can be optimized by producing fixed constant explanations for all samples. The explanation-aware loss term is independent of the original explanation $h_{\theta'}(\mathbf{x})$. All our extensions, in turn, consider the original explanations and hence force the model to keep its explanations for clean samples.

A similar aspect becomes apparent in the evaluation. While our evaluation considers all combinations of original model/manipulated model and clean sample/adversarial example (cf. Section 4), related work [13, 27, 46, 52] has focused on the vulnerability of the explanations and not investigated how much their defensive measure changes the explanations from the original model. In other words, related work has only evaluated metrics on the basis of one model.

4. Evaluation

We describe our experimental setup in Section 4.1 and present a preliminary study to concretize the attack strategy in Section 4.2. Once we have established the best attack strategy this way, we introduce our experiment design in Section 4.3. In Section 4.4, we describe additional metrics to quantify the success of the defenses and then discuss the results in Section 4.5.

4.1. Experimental Setup

Similar to most work in this area, we focus on the image domain [13, 27, 29, 35, 59, 64, 67]. In particular, we choose the common CIFAR-10 dataset [31], consisting of 50,000 training images and 10,000 test images of 32×32 colored pixels each. We further divide the training data into 40,000 training images and 10,000 validation images for our hyperparameter optimization. We normalize the data and apply a random horizontal flip as a data augmentation when training the original models. In line with related work [39, 40], we choose ResNet20 [23] as the model architecture.

As we will outline in Section 5, explanation-aware adversarial training is extremely computationally expensive. Therefore, we had to limit the scope of our evaluation to this simple dataset and architecture. Nevertheless, we are positive that our findings generalize to larger datasets and model architectures, and we encourage future work to verify this hypothesis.

4.2. Prediction-Untargeted Attacks

Most explanation-aware attacks either preserve the prediction (PP) [3, 16, 17, 20, 28, 32, 46, 53, 58, 59], or aim for a specific target class in the case of dual (D) [68] and explanation-preserving attacks (EP) [68], i.e., the attack is prediction-targeted. Note again that we use the terms *prediction-targeted* and *prediction-untargeted* to emphasize that we mean the predictions and not the explanations that may be targeted or untargeted, respectively. On the other hand, works on vanilla adversarial training mostly consider only prediction-untargeted attacks [35, 64, 67]. Hence, we also adopt the prediction-untargeted setting for dual attacks (D) and explanation-preserving attacks (EP) but provide an easy adaptation in our artifacts to evaluate the prediction-targeted attacks as well. In the other attack scenarios (PO, $PP^{\square}$, and $PP^{\star}$) the prediction is preserved.

Attack Strategies. Prediction-targeted attacks usually minimize the cross-entropy toward an adversary-chosen target class y^t:

$$(\mathbf{x}, \theta, \epsilon) \mapsto \mathbf{x} + \underset{\delta \ s.t. \ |\delta| \leq \epsilon}{\arg\min} \ CE(f_\theta(\mathbf{x}+\delta), y^t) \ .$$

Prediction-untargeted attacks, in turn, maximize the cross-entropy toward the ground-truth label (cf. Section 3.2):

$$(\mathbf{x}, \theta, \epsilon) \mapsto \mathbf{x} + \underset{\delta \ s.t. \ |\delta| \leq \epsilon}{\arg\max} \ CE(f_\theta(\mathbf{x}+\delta), \hat{y}) \ .$$

By design, the cross-entropy grows higher the more confident the model becomes. This leads to an overwhelming effect of the prediction loss over the explanation loss. Fooling the explanation thus fails, while the confidence in the wrong prediction increases. Our extensive hyperparameter study yields no satisfying results with this trivial attack strategy.

As a remedy we propose two alternative strategies:

(1) *Random.* We pick a wrong target class ($\neq \hat{y}$) at random and run prediction-targeted attacks against this class.

(2) *Highest Wrong.* We select the incorrect class ($\neq \hat{y}$) with the highest probability score as the target class, i.e., often the second most-likely class.

Results. We extensively optimize hyperparameters for 1,000 trials on 2,000 test samples for all strategies against three pre-trained models θ'_i. The best-performing hyperparameters are then evaluated on the complete test set. As shown in Table 3, the highest-wrong strategy performs

TABLE 3: Comparison of the three different attack strategies for $D^{\square}$, $D^{\star}$, and EP and the CIFAR-10 dataset.

Goal	Algorithm	Acc^{rob}(↓)	For (↓)	FR (↑)
$D^{\square}$	Trivial	0.127 0.03	0.102 0.04	0.226 0.07
	Random	0.008 0.00	0.018 0.02	0.976 0.01
	Highest Wrong	**0.000** 0.00	**0.013** 0.01	**0.993** 0.00
Goal	**Algorithm**	Acc^{rob}(↓)	Vul (↓/↑)	
EP	Trivial	0.628 0.03	0.028 0.04	–
	Random	0.009 0.00	0.001 0.00	–
	Highest Wrong	**0.000** 0.00	**0.001** 0.00	–
$D^{\star}$	Trivial	0.104 0.01	0.215 0.07	–
	Random	0.008 0.00	**0.345** 0.05	–
	Highest Wrong	**0.000** 0.00	0.339 0.05	–

best, while the trivial approach fails completely. *Note that in this table, we make an exception and highlight the best results from the attacker's perspective rather than the defender's.* Additional details on these findings are presented in Section B. Based on these results, we use the highest-wrong strategy for all prediction-untargeted and explanation-aware attacks for the remainder of this work, i.e., for EP, $D^{\square}$, and $D^{\star}$ attacks.

4.3. Experiment Design

We train three ResNet20 models [23], which we denote as *original models* θ'_i. Each original model θ'_i achieves a clean top-1 accuracy ($\text{Acc}^{\text{clean}}$) of at least 91.57 %. The hyperparameters are optimized for each explanation-aware inference-time attack (EP, …, $D^{\star}$) and for each of the three original models independently. We fix the number of PGD iterations during fine-tuning to $N = 200$ and optimize the weighting factor γ and the learning rate η. The best parameters are used during adversarial training.

For each of the three original models θ'_i and each adversarial goal, we execute all explanation-aware AT approaches, including RAR. Each hyperparameter search consists of 30 trials for 3 epochs for X-PGD-AT and RAR, and is implemented using the framework `optuna` with the `GridSampler`. To account for the additional hyperparameters in X-TRADES and X-MART, we optimize these methods for 60 trials instead.

From these initial trials, we select the configurations on the Pareto front of the relevant metrics and fine-tune the models for additional 3 epochs. This two-step process reduces computational costs (cf. Section 5) while still allowing for thorough hyperparameter optimization. The best hyperparameters are then selected according to a weighted sum of all relevant metrics (cf. Table 11).

In addition, we fine-tune three models per vanilla AT technique (PGD-AT, TRADES, and MART) with prediction-only (PO) attacks for 10 epochs as suggested in the original papers. The PO attacks are executed with $N = 7$ and $\eta = 2/255$ as suggested by prior work [35].

The non-AT defense strategies (ATEX and CRC) are fine-tuned for 200 epochs instead of 3 epochs, using the hyperparameters for the CIFAR-10 dataset of the original publication [29]. For CRC, the hyperparameters have been reported for the ResNet20 architecture.

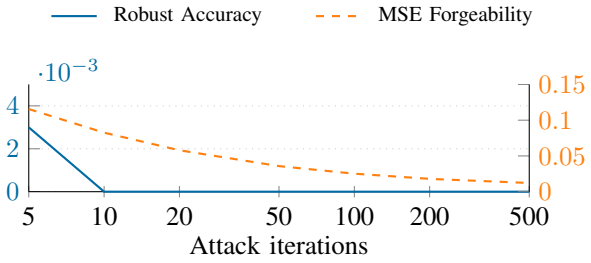


Figure 2: Convergence analysis of attack strength for the $D^{\square}$ attack scenario as a function of PGD iterations. The solid line represents the robust accuracy (left y-axis), while the dashed orange line denotes the MSE-based forgeability metric (right y-axis). The robust accuracy exhibits earlier saturation compared to the forgeability measure, which necessitates the use of more iterations for explanation-aware attacks.

For our evaluation, we optimize the hyperparameters of each explanation-aware inference-time attack again for each robustified model, i.e., we assume white-box adversaries. In other words, the hyperparameters are always optimized for the model at hand. In contrast to the fine-tuning, though, we invest more PGD iterations ($N = 500$) for the final evaluation. The number of iterations is chosen through a preliminary convergence study presented in Fig. 2. The figure shows that the robust accuracy saturates rather quickly while the explanation forgeability requires $\geq$200 iterations to converge.

Eventually, we report average values across the respective three models together with corresponding standard deviations in smaller font size. We provide further details on our experiment design and our hyperparameter optimizations in Section C.

4.4. Metrics to Compare Two Models

To quantify the adverse effects of defensive fine-tuning on explanations of clean samples, we introduce two additional metrics that compare explanations between the original model θ' and the corresponding robustified model θ^*. To motivate this approach, consider a model that produces identical explanations regardless of the input: while such a model would achieve a perfect vulnerability score, its explanations would be practically meaningless. This crucial aspect has been largely overlooked in prior work [13, 27, 46, 52].

Explanation Fidelity (Fid). The explanation fidelity measures how much the explanations of clean samples $\mathbf{x}$ differ between the original model θ' and the corresponding robustified model θ^*:

$$Fid_{\theta',\theta^*} := \mathop{\mathbb{E}}_{(\mathbf{x},\cdot)\in D} \left[d_{\mathcal{E}}(h_{\theta'}(\mathbf{x}), h_{\theta^*}(\mathbf{x})) \right] .$$

Intuitively, the fidelity assesses how much the clean explanations suffer from the applied defensive technique.

Explanation Deviation (Dev). The explanation deviation compares the clean explanations in the original model with the manipulated explanations in the robustified model:

$$Dev_{\theta',\theta^*} := \mathop{\mathbb{E}}_{(\mathbf{x},\cdot)\in D} \left[d_{\mathcal{E}}(h_{\theta'}(\mathbf{x}), h_{\theta^*}(\tilde{\mathbf{x}})) \right] .$$

TABLE 4: Summary of metrics to evaluate one model (blue) and to compare two models with each other (orange).

Metric		Formula	
For	Forgeability	For_θ	$:= \mathbb{E}_{(\mathbf{x},\cdot)\in D}[d_{\mathcal{E}}(h_\theta(\tilde{\mathbf{x}}), \mathbf{r}^t)]$
FR	Fooling Rate	FR_θ	$:= \mathbb{E}_{(\mathbf{x},\cdot)\in D}[\mathbf{1}_{d_{\mathcal{E}}(h_\theta(\tilde{\mathbf{x}}),\mathbf{r}^t)<\tau}]$
Vul	Vulnerability	Vul_θ	$:= \mathbb{E}_{(\mathbf{x},\cdot)\in D}[d_{\mathcal{E}}(h_\theta(\tilde{\mathbf{x}}), h_\theta(\mathbf{x}))]$
Fid	Fidelity	Fid_{θ',θ^*}	$:= \mathbb{E}_{(\mathbf{x},\cdot)\in D}[d_{\mathcal{E}}(h_{\theta'}(\mathbf{x}), h_{\theta^*}(\mathbf{x}))]$
Dev	Deviation	Dev_{θ',θ^*}	$:= \mathbb{E}_{(\mathbf{x},\cdot)\in D}[d_{\mathcal{E}}(h_{\theta'}(\mathbf{x}), h_{\theta^*}(\tilde{\mathbf{x}}))]$

The explanation deviation thus combines the effects of the explanation fidelity, i.e., how much the explanations change between models for clean samples, and the vulnerability/forgeability of the robustified model, i.e., how much the explanation can be manipulated.

Table 4 summarizes the metrics from Section 3.4 in the upper part and the metrics from Section 4.4 in the lower part.

4.5. Results

Our core results are presented in Fig. 3, with each of the six rows corresponding to a distinct adversarial goal. The quantitative results associated with this figure are detailed in Table 14 within Section A, which also provides results for alternative distance measures, specifically PCC and SSIM, presented in Table 15.

Prediction-Only Attacks (PO) manipulate predictions while ignoring explanations. Both vanilla AT variants and explanation-aware extensions are substantially effective against such PO attacks. The robust accuracy increases from $0\,\%$ in the non-robustified models to above $32.8\,\%$ for X-TRADES, and exceeds $40\,\%$ for the remaining AT techniques. Our empirical results for vanilla AT methods align with those reported in the original publications for both clean and robust accuracy metrics.

As anticipated, explanation-aware variants exhibit degraded performance relative to their vanilla counterparts when evaluated against PO attacks, manifesting in reduced clean and robust accuracy. Nevertheless, the relative ranking among the three approaches remains consistent across both vanilla and explanation-aware variants.

The specialized defense mechanisms ATEX and CRC prove least effective against PO attacks, yielding only marginal improvements in robust accuracy, which remains below $10\,\%$. This outcome is not surprising, as neither ATEX nor CRC explicitly protects predictions.

Explanation-Preserving Attacks (EP) alter the predictions while the explanations are preserved. Thereby, an ongoing attack against the predictions can be disguised from explanation-based analysis. For such EP attacks, we observe similar trends as already seen for PO attacks. Both vanilla AT and explanation-aware extensions achieve comparable performance across all evaluated metrics. EP attacks, hence, do not appear more challenging to defend against than PO attacks. ATEX and CRC fail to improve the robust accuracy and are also not substantially superior in the other metrics, except for ATEX, which maintains a high clean accuracy.

Prediction-Preserving Attacks (PP) manipulate only the explanations while keeping the predictions unchanged. The most pronounced performance differences emerge in the context of these PP attacks. For the $PP^{\square}$ attack scenario, X-PGD-AT, X-MART, RAR, ATEX, and CRC prove ineffective, exhibiting high fooling rates and low forgeability. The remaining methods (PGD-AT, TRADES, MART, X-TRADES) demonstrate superior performance, with PGD-AT, TRADES, and MART achieving a fooling rate of $0\,\%$. These three methods induce only minimal alterations to the clean explanations, as evidenced by their low fidelity scores (≤ 0.047) and low vulnerability values (≤ 0.013). Similar trends are observed for $PP^{\star}$ attacks. While X-PGD-AT, X-MART, RAR, ATEX, and CRC fail to prevent manipulation of explanations, PGD-AT, TRADES, and MART consistently achieve strong results with low vulnerability (≤ 0.023) and low fidelity (≤ 0.047). As expected, the robust accuracy of all methods approaches $100\,\%$ for PP attacks, since the adversary optimizes toward the ground-truth label.

Dual Attacks (D) aim to simultaneously manipulate both predictions and explanations. In both D attack scenarios, the robust accuracy of explanation-aware variants is consistently lower than that of their vanilla counterparts. The explanation-aware variants also exhibit slightly worse forgeability compared to vanilla AT methods. The specialized methods ATEX and CRC demonstrate the weakest performance against D attacks.

Overall Assessment. Across all scenarios, ATEX consistently outperforms CRC. Notably, the AT techniques surpass both ATEX and CRC in nearly all metrics and attacks. Surprisingly, vanilla AT techniques frequently even outperform their explanation-aware counterparts. Additionally, vanilla methods offer lower complexity and are more extensively studied in the literature. Our explanation-aware extensions achieve superior performance only in terms of clean accuracy, regularly exceeding $80\,\%$ (see Table 5 and Table 6). Based on these results, we state our core finding:

> **Core Finding #1:** Vanilla AT techniques are superior in defending against explanation-aware adversarial attacks compared to specialized defense mechanisms.

5. Cost Analysis

Explanation-aware adversarial training incurs substantially higher computational costs compared to vanilla adversarial training. The complete replication of our experimental evaluation requires approximately 16,320 GPU hours on NVIDIA A100 GPUs.

We now present a comparative analysis of the computational costs across the investigated defense methods. Fig. 4 illustrates the wall-clock time requirements for each approach, measured on a dedicated AMD Ryzen 9 5900X 12-Core processor equipped with two NVIDIA RTX 3090Ti GPUs. The total execution time comprises three components: adversarial example generation (▨), model training time (▩), and auxiliary operations (□). For both vanilla AT and explanation-aware AT methods, adversarial sample generation is parallelized across both GPUs. The

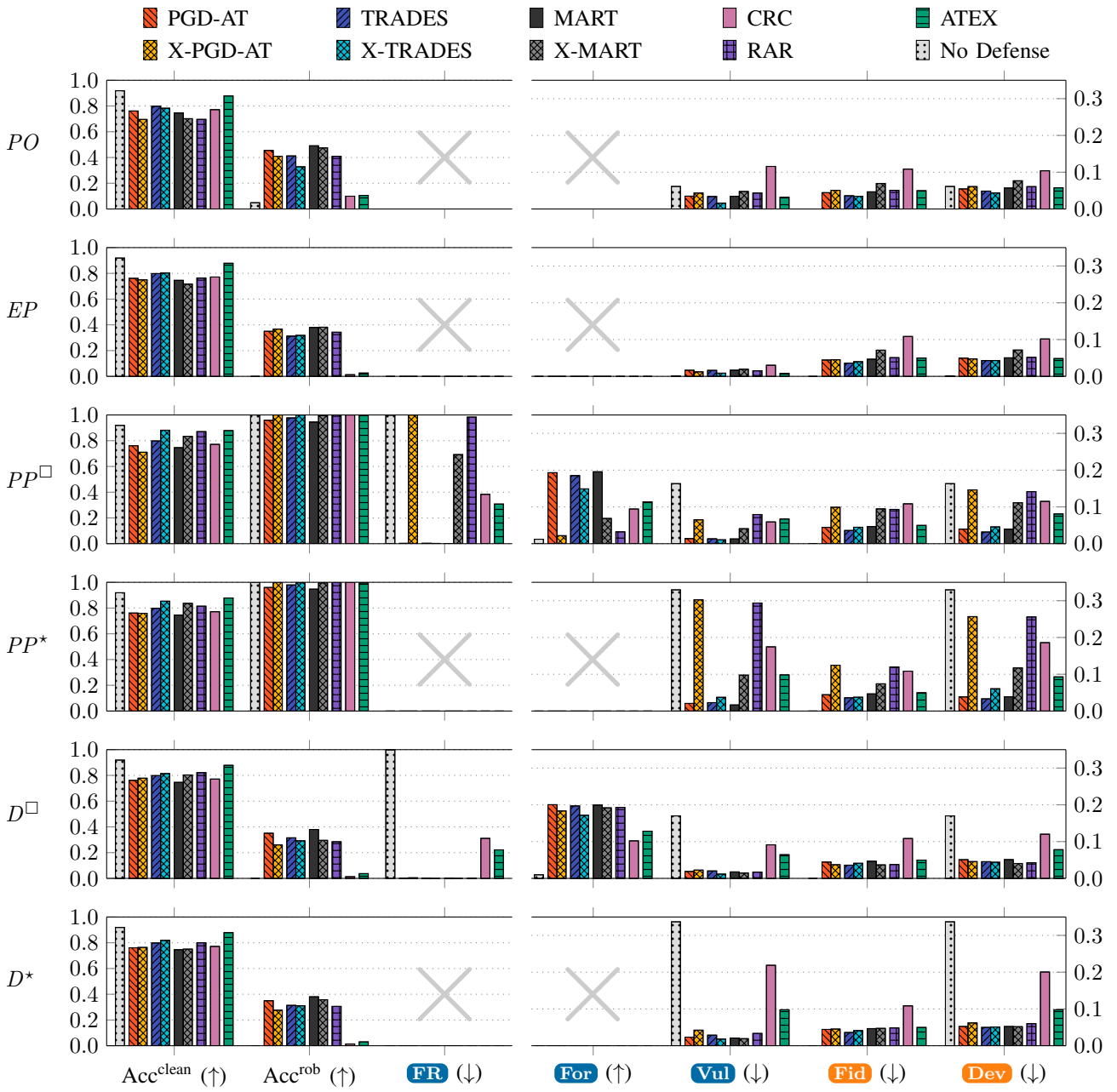


Figure 3: The results of the five existing defenses and the explanation-aware extensions. Clean accuracy ($\text{Acc}^{\text{clean}}$), robust accuracy ($\text{Acc}^{\text{rob}}$), and fooling rate (FR) are with respect to the left y-axis while forgeability (For), vulnerability (Vul), fidelity (Fid), and deviation (Dev) are with respect to the right y-axis. Table 14 contains the quantitative results to this figure.

explanation-aware approaches exhibit approximately 6-fold computational overhead compared to vanilla AT when utilizing 200 PGD iterations, with this disparity further increasing for 500 iterations. This performance degradation is attributable to two primary factors: (1) the increased complexity of generating explanation-aware adversarial examples necessitates additional iterations, and (2) each iteration requires explicit computation of the explanation method and back-propagation of the explanation loss.

Results. The vanilla AT techniques demonstrate the lowest computational overhead. CRC exhibits only moderately elevated computational costs relative to vanilla AT. The explanation-aware AT methods incur substantially higher computational expenses, with wall-clock times up to 6 times greater than their vanilla counterparts. This overhead is predominantly concentrated in the generation of explanation-aware adversarial samples. Also, ATEX incurs particularly high computational costs, exceeding 750 seconds per epoch. The substantial computational burden of ATEX is primarily attributable to its sample generation procedure, which precludes parallelization across GPUs.

Core Finding #2: Vanilla AT techniques are computationally substantially less expensive than explanation-aware adversarial training and its variants.

TABLE 5: The model utility results of the explanation-aware AT methods.

Defense	$\text{Acc}^{\text{clean}}$	Fid	$\text{Acc}^{\text{clean}}$	Fid	$\text{Acc}^{\text{clean}}$	Fid	$\text{Acc}^{\text{clean}}$	Fid
PO	$0.696_{0.01}$	$0.051_{0.04}$	$0.783_{0.01}$	$0.035_{0.03}$	$0.702_{0.01}$	$0.069_{0.04}$	$0.696_{0.01}$	$0.051_{0.04}$
EP	$0.750_{0.01}$	$0.045_{0.04}$	$0.803_{0.01}$	$0.040_{0.03}$	$0.716_{0.02}$	$0.071_{0.05}$	$0.763_{0.01}$	$0.051_{0.04}$
$PP^{\square}$	$0.710_{0.05}$	$0.099_{0.04}$	$0.881_{0.01}$	$0.045_{0.04}$	$0.835_{0.01}$	$0.095_{0.04}$	$0.871_{0.01}$	$0.093_{0.04}$
$PP^{\star}$	$0.759_{0.08}$	$0.125_{0.05}$	$0.853_{0.01}$	$0.038_{0.02}$	$0.836_{0.01}$	$0.074_{0.04}$	$0.815_{0.03}$	$0.120_{0.05}$
$D^{\square}$	$0.778_{0.02}$	$0.037_{0.03}$	$0.814_{0.01}$	$0.041_{0.03}$	$0.804_{0.01}$	$0.037_{0.03}$	$0.821_{0.01}$	$0.038_{0.03}$
$D^{\star}$	$0.764_{0.02}$	$0.045_{0.03}$	$0.818_{0.01}$	$0.041_{0.03}$	$0.751_{0.02}$	$0.047_{0.03}$	$0.800_{0.01}$	$0.048_{0.03}$
	(a) X-PGD-AT		(b) X-TRADES		(c) X-MART		(d) RAR	

TABLE 6: The model utility results of the vanilla AT methods, CRC, and ATEX.

Defense	$\text{Acc}^{\text{clean}}$	Fid
PGD-AT	$0.761_{0.01}$	$0.044_{0.04}$
TRADES	$0.798_{0.01}$	$0.036_{0.03}$
MART	$0.746_{0.01}$	$0.047_{0.04}$
CRC	$0.772_{0.03}$	$0.108_{0.05}$
ATEX	$0.878_{0.01}$	$0.050_{0.03}$

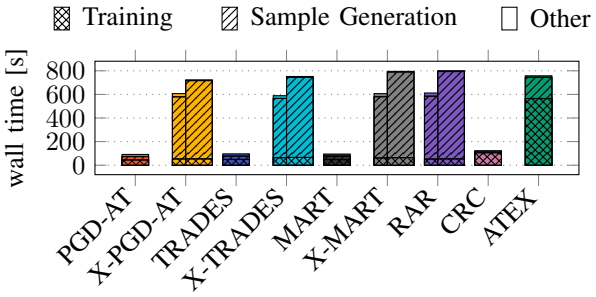


Figure 4: Computational costs, in wall time in seconds. For each evaluated approach the total time is split in the amount of time required for generating attack samples, the actual training, and other functionality. For explanation-aware AT methods, the displayed data is for the $D^{\square}$ attacks with 200 (left bars) and 500 PGD iterations (right bars). Other attack scenarios yield similar results, while higher PGD iterations increase the computational costs proportionally.

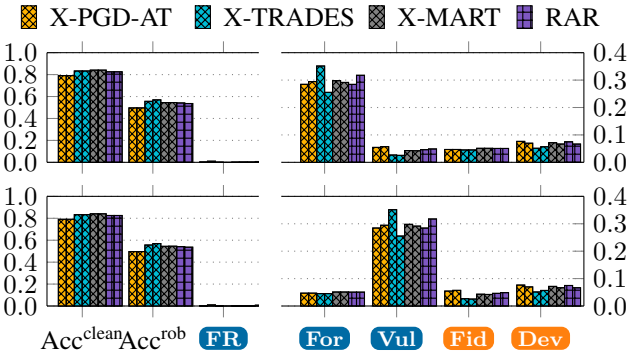


Figure 5: Transferability analysis of our three explanation-aware extensions for different target explanations. The clean accuracy ($\text{Acc}^{\text{clean}}$), robust accuracy ($\text{Acc}^{\text{rob}}$), and the fooling rate are with respect to the left y-axis, while the rest is with respect to the right y-axis. The corresponding quantitative results are presented in Table 13 and 16.

6. Transferability Studies

Explanation-aware adversarial training requires the defender to anticipate the attacker's goal and their target explanation. In this section, we investigate what happens if one of those two assumptions is broken, i.e., we investigate if a certain transferability can be observed with respect to the adversarial goal or the target explanation.

We first investigate the transferability between adversarial goals in Section 6.1. Thereafter, in Section 6.2, we analyze the transferability between two maximally different target explanations.

6.1. Goal Transferability

Tables 7 and 8 show how training against a specific adversarial goal affects the robustness to other adversarial goals. Notably, the clean accuracy ($\text{Acc}^{\text{clean}}$) and the fidelity remain constant across all attack scenarios, independent of the investigated real attack scenario and are thus reported in the Tables 5 and 6. When training on PO or EP attacks, the resulting models are robust against both, their anticipated attack scenario and other scenarios. While training on PO attacks yields slightly better results for explanations, note that the accuracy of models trained on these attacks is significantly lower. The same holds true for training with EP attacks; however the differences for explanations and accuracy are smaller. For PP attacks ($PP^{\square}$ and $PP^{\star}$), the fine-tuning does not improve the robust accuracy for not-anticipated attack scenarios. This result is not surprising since PP attacks do not manipulate the predictions. Our hyperparameter search for the explanation-aware variants (except for X-TRADES) does not yield satisfying results for PP attacks. This finding is again reflected in the transferability results. However, the explanation robustness of models trained with X-TRADES on PP attacks does not transfer to other attack scenarios. In particular, $PP^{\square}$ models are not robust against $D^{\star}$ attacks. We observe that training on either $D^{\square}$ or $D^{\star}$ yields similar robust accuracies against both D attacks.

Overall, the explanation robustness transfers well across different attacks, indicating that, regardless of the attack anticipated, the explanations are generally robust to manipulation. This finding is also in line with our observations for vanilla AT in Fig. 3, where standard AT methods perform well or even superior for explanation-aware attacks.

TABLE 7: Transferability analysis of explanation-aware defenses across distinct adversarial goals. The best performance is emphasized in bold font, evaluated from the defender's perspective relative to the actual attack scenario (column one). This table presents all attack scenarios without target explanations; consequently, the explanation forgeability metric and fooling rates are omitted. Explanation-targeted scenarios are detailed in Table 8. For prediction-only attacks (PO), the adversary disregards explanations; therefore, the Vul and Dev columns contain no bold emphasis.

Att.	Def.	Acc^{rob}	Vul	Dev
PO	PO	**0.410** 0.02	0.043 0.05	0.061 0.04
	EP	0.398 0.02	0.040 0.05	0.058 0.04
	$PP^{\square}$	0.162 0.02	0.079 0.04	0.123 0.03
	$PP^{\star}$	0.129 0.03	0.125 0.06	0.165 0.05
	$D^{\square}$	0.326 0.01	0.033 0.03	0.045 0.03
	$D^{\star}$	0.340 0.02	0.042 0.04	0.053 0.04
EP	PO	0.279 0.02	**0.018** 0.03	0.054 0.04
	EP	**0.367** 0.01	0.012 0.02	0.047 0.03
	$PP^{\square}$	0.002 0.0	0.006 0.01	0.101 0.04
	$PP^{\star}$	0.000 0.0	0.004 0.01	0.125 0.05
	$D^{\square}$	0.257 0.02	0.012 0.02	**0.041** 0.03
	$D^{\star}$	0.274 0.02	0.015 0.02	0.048 0.03
$PP^{\star}$	PO	0.946 0.01	0.024 0.04	0.043 0.04
	EP	0.959 0.01	**0.019** 0.03	**0.037** 0.03
	$PP^{\square}$	**1.000** 0.0	0.202 0.05	0.169 0.05
	$PP^{\star}$	**1.000** 0.0	0.303 0.08	0.257 0.07
	$D^{\square}$	0.987 0.0	0.042 0.05	0.051 0.04
	$D^{\star}$	0.980 0.0	0.039 0.05	0.051 0.04
$D^{\star}$	PO	0.280 0.02	0.025 0.04	0.056 0.04
	EP	**0.367** 0.01	**0.017** 0.03	**0.050** 0.04
	$PP^{\square}$	0.003 0.0	0.216 0.05	0.195 0.05
	$PP^{\star}$	0.001 0.0	0.319 0.07	0.276 0.07
	$D^{\square}$	0.259 0.02	0.046 0.05	0.061 0.05
	$D^{\star}$	0.276 0.02	0.042 0.05	0.062 0.05

(a) X-PGD-AT

Att.	Def.	Acc^{rob}	Vul	Dev
PO	PO	0.328 0.02	0.016 0.02	0.044 0.03
	EP	0.365 0.02	0.023 0.03	0.048 0.03
	$PP^{\square}$	0.096 0.01	0.038 0.04	0.054 0.04
	$PP^{\star}$	0.208 0.02	0.025 0.02	0.045 0.03
	$D^{\square}$	0.362 0.02	0.023 0.03	0.048 0.03
	$D^{\star}$	**0.373** 0.02	0.020 0.03	0.047 0.03
EP	PO	0.220 0.02	0.007 0.01	**0.039** 0.03
	EP	**0.318** 0.02	0.009 0.01	0.043 0.03
	$PP^{\square}$	0.029 0.01	**0.017** 0.03	0.053 0.04
	$PP^{\star}$	0.147 0.02	0.011 0.01	**0.039** 0.02
	$D^{\square}$	0.290 0.02	0.009 0.01	0.044 0.03
	$D^{\star}$	0.310 0.02	0.009 0.01	0.043 0.03
$PP^{\star}$	PO	0.983 0.0	**0.013** 0.02	**0.033** 0.03
	EP	0.983 0.0	0.020 0.03	0.051 0.03
	$PP^{\square}$	**0.999** 0.0	0.040 0.03	0.060 0.04
	$PP^{\star}$	0.997 0.0	0.037 0.04	0.061 0.04
	$D^{\square}$	0.989 0.0	0.020 0.03	0.051 0.03
	$D^{\star}$	0.988 0.0	0.021 0.03	0.051 0.03
$D^{\star}$	PO	0.221 0.02	**0.012** 0.02	**0.043** 0.03
	EP	**0.319** 0.02	0.016 0.02	0.049 0.04
	$PP^{\square}$	0.030 0.01	0.068 0.05	0.087 0.05
	$PP^{\star}$	0.150 0.02	0.051 0.05	0.071 0.05
	$D^{\square}$	0.291 0.02	0.020 0.03	0.053 0.04
	$D^{\star}$	0.310 0.02	0.018 0.03	0.051 0.04

(b) X-TRADES

Att.	Def.	Acc^{rob}	Vul	Dev
PO	PO	**0.475** 0.02	0.048 0.05	0.077 0.04
	EP	0.458 0.02	0.054 0.06	0.083 0.05
	$PP^{\square}$	0.217 0.02	0.087 0.05	0.107 0.04
	$PP^{\star}$	0.216 0.02	0.076 0.05	0.086 0.04
	$D^{\square}$	0.380 0.02	0.028 0.03	0.047 0.03
	$D^{\star}$	0.446 0.02	0.038 0.05	0.058 0.04
EP	PO	0.356 0.02	0.024 0.04	0.069 0.04
	EP	**0.382** 0.01	0.019 0.04	0.071 0.05
	$PP^{\square}$	0.134 0.01	**0.036** 0.04	0.097 0.04
	$PP^{\star}$	0.138 0.01	0.031 0.03	0.076 0.04
	$D^{\square}$	0.295 0.02	0.013 0.02	**0.040** 0.03
	$D^{\star}$	0.358 0.02	0.015 0.03	0.049 0.03
$PP^{\star}$	PO	0.916 0.01	0.021 0.04	0.063 0.04
	EP	0.936 0.01	0.023 0.04	0.061 0.04
	$PP^{\square}$	**0.996** 0.0	0.104 0.07	0.140 0.06
	$PP^{\star}$	0.995 0.0	0.098 0.07	0.117 0.06
	$D^{\square}$	0.987 0.0	0.017 0.03	**0.035** 0.03
	$D^{\star}$	0.962 0.01	**0.016** 0.03	0.041 0.03
$D^{\star}$	PO	0.355 0.02	0.027 0.04	0.070 0.04
	EP	**0.380** 0.01	0.025 0.04	0.075 0.05
	$PP^{\square}$	0.134 0.01	0.138 0.08	0.155 0.07
	$PP^{\star}$	0.140 0.01	0.127 0.08	0.137 0.08
	$D^{\square}$	0.297 0.02	0.021 0.03	**0.045** 0.03
	$D^{\star}$	0.357 0.02	**0.019** 0.03	0.052 0.04

(c) X-MART

Att.	Def.	Acc^{rob}	Vul	Dev
PO	PO	**0.409** 0.02	0.043 0.05	0.061 0.04
	EP	0.390 0.02	0.046 0.05	0.063 0.04
	$PP^{\square}$	0.085 0.01	0.080 0.04	0.122 0.05
	$PP^{\star}$	0.121 0.02	0.123 0.06	0.158 0.05
	$D^{\square}$	0.343 0.02	0.028 0.03	0.046 0.03
	$D^{\star}$	0.372 0.02	0.040 0.04	0.057 0.04
EP	PO	0.278 0.02	**0.019** 0.03	0.054 0.04
	EP	**0.343** 0.02	0.015 0.02	0.052 0.04
	$PP^{\square}$	0.003 0.0	0.008 0.01	0.094 0.04
	$PP^{\star}$	0.001 0.0	0.005 0.01	0.121 0.05
	$D^{\square}$	0.283 0.02	0.013 0.02	**0.041** 0.03
	$D^{\star}$	0.305 0.02	0.017 0.03	0.051 0.03
$PP^{\star}$	PO	0.945 0.01	0.026 0.04	**0.043** 0.04
	EP	0.968 0.0	**0.023** 0.04	0.046 0.03
	$PP^{\square}$	0.998 0.0	0.235 0.07	0.212 0.07
	$PP^{\star}$	**1.000** 0.0	0.294 0.08	0.256 0.07
	$D^{\square}$	0.991 0.0	0.028 0.04	0.044 0.03
	$D^{\star}$	0.986 0.0	0.031 0.04	0.052 0.04
$D^{\star}$	PO	0.278 0.02	0.025 0.04	0.056 0.04
	EP	**0.344** 0.02	**0.023** 0.03	0.056 0.04
	$PP^{\square}$	0.005 0.0	0.271 0.06	0.245 0.07
	$PP^{\star}$	0.002 0.0	0.316 0.07	0.281 0.07
	$D^{\square}$	0.285 0.02	0.030 0.04	**0.052** 0.04
	$D^{\star}$	0.306 0.02	0.033 0.05	0.060 0.04

(d) RAR

6.2. Target Explanation Transferability

We investigate how our explanation-aware extensions depend on the specific target explanation anticipated during fine-tuning. To this end, we conduct a transferability study using two maximally different target explanations: one where the left half is activated and one where the right half is activated. For both target explanations, we train three models per defensive approach and visualize the averaged results in Fig. 5 and report the quantitative results in Tables 13 and 16 in Section A.

In the top half of the figure, models are trained on dual attacks with the left-half target explanation; for each method, the two bars show evaluation against attacks with the left-half target (first bar) and the right-half target (second bar). In the bottom half, models are trained on the right-half target and evaluated on both right-half (first bar) and left-half (second bar) attacks. If both bars for a method are similar, this indicates a high transferability between target explanations.

Across all methods and metrics, we observe that the transferability is very high: models trained on one target explanation are also robust against attacks using the opposite target. This trend holds for all four explanation-aware adversarial training methods and is consistent across all evaluated metrics.

> **Core Finding #3:** Robustness gained from adversarial training with one specific target explanation generalizes to other—even maximally different—target explanations.

7. Conclusion

We present the first extensive evaluation of adversarial training (AT) techniques to counter explanation-aware attacks across all established attack scenarios targeting the popular GradCAM explanation method. We also evaluate vanilla AT methods (PGD-AT, TRADES, and MART) as a simple baseline. Moreover, we introduce new explanation-aware extensions for TRADES and MART, and show that FAR is equivalent to the extension of PGD-AT. Further, we find that CRC and ATEX are ineffective against explanation-aware attacks that target predictions *and* explanations, i.e., dual attacks (D). Note that both methods have been proposed for attacks against explanations only (EO). RAR, in turn, performs similarly to our explanation-aware extension of PGD-AT. In particular, both fail to prevent prediction-preserving (PP) attacks.

Our work reveals that vanilla adversarial training already provides strong robustness against explanation-aware attacks, and that the additional effort for explanation-aware adversarial training yields little to no benefit right now. With this result, we answer a long-standing question in the field regarding the necessity of explanation-aware adversarial training. However, the underlying reasons for this phenomenon remain unclear and warrant further investigation. We encourage the community to explore the relationship between explanations and predictions, and to analyze which heuristics are exploited by attacks on explanation methods and why these cannot be leveraged in adversarially trained models.

Ethical Considerations

The explanation-aware attacks discussed in this work are already known and explored. Our work does not improve attacks. Instead, we propose and evaluate defenses to counter those attacks. Our investigation can make AI systems more robust and trustworthy. However, we emphasize that the attack surface for manipulating explanations is complex and not yet fully understood. Overall, we do not expect our research to have negative effects on third parties.

Acknowledgement

The authors gratefully acknowledge funding from the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems". This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. The authors further acknowledge support by the state of Baden-Württemberg through bwHPC. Further, the authors thank their colleagues Qi Zhao, Nicolai Neuer, Shayari Bhattacharjee, and Zhiyang Cheng for their support in determining the fooling-rate thresholds.

References

[1] E. Abdukhamidov, M. Abuhamad, S. S. Woo, E. Chan-Tin, and T. Abuhmed. Interpretations cannot be trusted: Stealthy and effective adversarial perturbations against interpretable deep learning. *CoRR*, abs/2211.15926, 2022.

[2] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3] D. Alvarez-Melis and T. S. Jaakkola. On the Robustness of Interpretability Methods. *Proc. of the ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, page 46, 2015.

[5] E. Bagdasaryan and V. Shmatikov. Blind backdoors in deep learning models. In *Proc. of the USENIX Security Symposium*, pages 1505–1521, 2021.

[6] H. Baniecki and P. Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. In *Proc. of the IJCAI Workshop of explainable AI (XAI)*, 2023.

[7] H. Baniecki, M. Chrabaszcz, A. Holzinger, B. Pfeifer, A. Saranti, and P. Biecek. Be careful when evaluating explanations regarding ground truth. *CoRR*, abs/2311.04813, 2023.

[8] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, volume 8190, pages 387–402, 2013.

[9] N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[10] N. Carlini and D. Wagner. Towards Evaluating the Robustness of Neural Networks. *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017.

[11] N. Carlini and D. A. Wagner. Defensive distillation is not robust to adversarial examples. *CoRR*, abs/1607.04311, 2016.

[12] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[13] J. Chen, X. Wu, V. Rastogi, Y. Liang, and S. Jha. Robust attribution regularization. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 14300–14310, 2019.

[14] J. Cortellazzi, F. Pendlebury, D. Arp, E. Quiring, F. Pierazzi, and L. Cavallaro. Intriguing properties of adversarial ML attacks in the problem space. *CoRR*, abs/1911.02142v3, 2025.

[15] N. N. Dalvi, P. M. Domingos, Mausam, S. K. Sanghai, and D. Verma. Adversarial classification. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 99–108, 2004.

[16] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 13567–13578, 2019.

[17] A.-K. Dombrowski, C. J. Anders, K.-R. Müller, and P. Kessel. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022.

[18] W. Fan, H. Xu, W. Jin, X. Liu, X. Tang, S. Wang, Q. Li, J. Tang, J. Wang, and C. C. Aggarwal. Jointly attacking graph neural network and its explanations. In *Proc. of the IEEE International Conference on Data Engineering (ICDE)*, 2023.

[19] S. Fang and A. Choromanska. Backdoor attacks on the dnn interpretation system. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 561–570, 2022.

[20] A. Ghorbani, A. Abid, and J. Y. Zou. Interpretation of neural networks is fragile. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pages 3681–3688, 2019.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

[22] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:43, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[24] W. He, J. Wei, X. Chen, N. Carlini, and D. Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *Proc. of the USENIX Workshop on Offensive Technologies (WOOT)*, 2017.

[25] A. Hegde, M. Noppel, and C. Wressnegger. Model-manipulation attacks against black-box explanations. In *Proc. of the Annual Computer Security Applications Conference (ACSAC)*, pages 974–987, 2024.

[26] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári. Learning with a strong adversary. *CoRR*, abs/1511.03034, 2015.

[27] A. Ivankay, I. Girardi, C. Marchiori, and P. Frossard. FAR: A general framework for attributional robustness. In *Proc. of the British Machine Vision Conference (BMVC)*, page 24, 2021.

[28] A. Ivankay, I. Girardi, P. Frossard, and C. Marchiori. Fooling Explanations in Text Classifiers. *Proc. of the International Conference on Learning Representations (ICLR)*, page 13, 2022.

[29] S. Joo, S. Jeong, J. Heo, A. Weller, and T. Moon. Towards more robust interpretation via local gradient alignment. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pages 8168–8176, 2023.

[30] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700, pages 267–280. 2019.

[31] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[32] A. Kuppa and N.-A. Le-Khac. Black box attacks on explainable artificial Intelligence(XAI) methods in cyber security. In *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[33] J. R. Lee, S. Kim, I. Park, T. Eo, and D. Hwang. Relevance-CAM: Your model already knows where to look. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[34] C. Lyu, K. Huang, and H.-N. Liang. A unified gradient regularization family for adversarial examples. In *Proc. of the International Conference on Data Mining (ICDM)*, pages 301–309, 2015.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

[36] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In

Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2574–2582, 2016.

[37] A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[38] M. Noppel and C. Wressnegger. SoK: Explainable machine learning in adversarial environments. In *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2024.

[39] M. Noppel and C. Wressnegger. Composite explanation-aware attacks. In *Proc. of the Deep Learning Security and Privacy Workshop (DLSP)*, 2025.

[40] M. Noppel, L. Peter, and C. Wressnegger. Disguising attacks with explanation-aware backdoors. In *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2023.

[41] E. Pachl, F. Langer, T. Markert, and J. M. Lorenz. A view on vulnerabilites: The security challenges of XAI (academic track). In *Proc. of the Symposium on Scaling AI Assessments (SAIA)*, volume 126, pages 12:1–12:23, 2024.

[42] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro. Intriguing properties of adversarial ML attacks in the problem space. In *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2020.

[43] G. Plumb, D. Molitor, and A. S. Talwalkar. Model agnostic supervised local explanations. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2520–2529, 2018.

[44] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[45] A. S. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pages 1660–1669, 2018.

[46] A. Sarkar, A. Sarkar, and V. N. Balasubramanian. Enhanced regularizers for attributional robustness. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pages 2532–2540, 2021.

[47] 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. *International Journal of Computer Vision*, 128(2):336–359, 2020.

[48] U. Shaham, Y. Yamada, and S. Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. 307:195–204, 2018.

[49] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proc. of the ACM Conference on Computer and Communications Security (CCS)*, pages 1528–1540, 2016.

[50] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 3145–3153, 2017.

[51] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Proc. of the International Conference on Learning Representations (ICLR) Workshop Track Proceedings*, 2014.

[52] M. Singh, N. Kumari, P. Mangla, A. Sinha, V. N. Balasubramanian, and B. Krishnamurthy. Attributional robustness training using input-gradient spatial alignment. In *Proc. of the European Conference on Computer Vision (ECCV)*, volume 12372 of *Lecture Notes in Computer Science*, pages 515–533, 2020.

[53] S. Sinha, H. Chen, A. Sekhon, Y. Ji, and Y. Qi. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing. *CoRR*, abs/2108.04990, 2021.

[54] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In *Proc. of the International Conference on Learning Representations (ICLR) Workshop Track Proceedings*, 2015.

[55] A. Subramanya, V. Pillai, and H. Pirsiavash. Fooling network interpretation in image classification. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2020–2029, 2019.

[56] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. *CoRR*, abs/1703.01365, 2017.

[57] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2014.

[58] S. V. Tamam, R. Lapid, and M. Sipper. Foiling explanations in deep neural networks. *CoRR*, abs/2211.14860, 2022.

[59] R. Tang, N. Liu, F. Yang, N. Zou, and X. Hu. Defense against explanation manipulation. *Frontiers Big Data*, 5:704203, 2022.

[60] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

[61] J. Vadillo, R. Santana, and J. A. Lozano. Adversarial attacks in explainable machine learning: A survey of threats against models and humans. *WIREs Data Mining and Knowledge Discovery*, page e1567, 2024.

[62] F. Wang and A. W.-K. Kong. A practical upper bound for the worst-case attribution deviations. *CoRR*, abs/2303.00340, 2023.

[63] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8565–8574, 2021.

[64] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu. Improving adversarial robustness requires revisiting misclassified examples. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

[65] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[66] M. Wu, S. Parbhoo, M. C. Hughes, R. Kindle, L. A. Celi, M. Zazzi, V. Roth, and F. Doshi-Velez. Regional tree regularization for interpretability in deep neural networks. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pages 6413–6421, 2020.

[67] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482, 2019.

[68] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *Proc. of the USENIX Security Symposium*, pages 1659–1676, 2020.

[69] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

Appendix A. Additional Results

In the following section, we present further results. Aside from the MSE metric presented in the main paper, we also provide results for the Pearson Correlation Coefficient (PCC) and the Structural Similarity Index Measure (SSIM) [65] as additional distance metrics. In particular, Tables 14 and 15 contain a summary of the quantitative results for all methods and attack scenarios, complementing Fig. 3 in the main paper. Table 13 shows quantitative results for the transferability across different target explanations, complementing Fig. 5 in the main paper. In Fig. 6, we show additional examples of adversarial samples and their explanations for the different attack scenarios and the investigated defense methods.

We explore the differences between the three investigated distance metrics used in our paper. Fig. 7 shows that similarity rankings of explanations can differ between the metrics and that a combination of metrics should be investigated to obtain a more holistic view of the similarity between explanations.

TABLE 8: Transferability study of explanation-aware defenses across different adversarial goals. The best results are, again, highlighted in bold according to the defender's perspective on the actual attack scenario (column one). In this table, we report the remaining two explanation-targeted scenarios; hence, we report the explanation forgeability metric and the fooling rates alongside all the other metrics. The non-explanation-targeted scenarios are presented in Table 7.

Att.	Def.	Acc^{rob}	FR	For	Vul	Dev
$PP^{\square}$	PO	0.943 0.01	0.002 0.0	0.191 0.02	0.018 0.03	0.045 0.03
	EP	0.956 0.0	**0.000** 0.0	**0.200** 0.03	**0.012** 0.03	0.038 0.03
	$PP^{\square}$	**1.000** 0.0	0.999 0.0	0.022 0.01	0.065 0.03	0.146 0.03
	$PP^{\star}$	**1.000** 0.0	0.992 0.0	0.039 0.02	0.082 0.03	0.136 0.03
	$D^{\square}$	0.987 0.0	0.013 0.01	0.171 0.03	0.017 0.03	**0.036** 0.03
	$D^{\star}$	0.977 0.0	0.016 0.01	0.157 0.03	0.019 0.03	0.044 0.03
$D^{\square}$	PO	0.281 0.02	0.001 0.0	0.196 0.02	0.020 0.03	0.055 0.04
	EP	**0.368** 0.01	**0.000** 0.0	**0.209** 0.02	**0.014** 0.03	0.048 0.03
	$PP^{\square}$	0.002 0.0	0.997 0.0	0.020 0.01	0.074 0.03	0.155 0.03
	$PP^{\star}$	0.001 0.0	0.973 0.01	0.042 0.02	0.092 0.03	0.143 0.03
	$D^{\square}$	0.260 0.02	0.005 0.0	0.183 0.03	0.022 0.03	**0.046** 0.03
	$D^{\star}$	0.276 0.02	0.005 0.0	0.170 0.03	0.023 0.03	0.051 0.03

(a) X-PGD-AT

Att.	Def.	Acc^{rob}	FR	For	Vul	Dev
$PP^{\square}$	PO	0.981 0.0	0.004 0.0	**0.198** 0.02	**0.008** 0.02	**0.031** 0.02
	EP	0.983 0.0	**0.000** 0.0	0.162 0.02	**0.008** 0.01	0.037 0.02
	$PP^{\square}$	**1.000** 0.0	0.002 0.0	0.149 0.02	0.011 0.02	0.046 0.04
	$PP^{\star}$	0.998 0.0	0.029 0.0	0.134 0.02	0.012 0.02	0.039 0.02
	$D^{\square}$	0.987 0.0	**0.000** 0.0	0.157 0.02	**0.008** 0.02	0.038 0.02
	$D^{\star}$	0.987 0.0	0.001 0.0	0.153 0.02	**0.008** 0.01	0.038 0.02
$D^{\square}$	PO	0.222 0.02	0.001 0.0	**0.210** 0.02	**0.008** 0.01	**0.039** 0.03
	EP	**0.318** 0.02	**0.000** 0.0	0.178 0.02	0.010 0.02	0.043 0.03
	$PP^{\square}$	0.030 0.01	0.001 0.0	0.156 0.02	0.032 0.04	0.062 0.04
	$PP^{\star}$	0.151 0.02	0.014 0.0	0.148 0.03	0.021 0.02	0.044 0.02
	$D^{\square}$	0.292 0.02	**0.000** 0.0	0.171 0.02	0.012 0.02	0.044 0.03
	$D^{\star}$	0.311 0.02	**0.000** 0.0	0.170 0.02	0.011 0.02	0.043 0.03

(b) X-TRADES

Att.	Def.	Acc^{rob}	FR	For	Vul	Dev
$PP^{\square}$	PO	0.915 0.01	0.002 0.0	0.172 0.03	0.018 0.04	0.063 0.04
	EP	0.935 0.01	0.002 0.0	0.173 0.03	0.018 0.04	0.059 0.04
	$PP^{\square}$	**0.997** 0.0	0.692 0.04	0.069 0.03	0.041 0.03	0.111 0.04
	$PP^{\star}$	0.994 0.0	0.325 0.02	0.101 0.03	0.035 0.03	0.085 0.04
	$D^{\square}$	0.986 0.0	**0.000** 0.0	**0.181** 0.02	**0.008** 0.02	**0.031** 0.02
	$D^{\star}$	0.960 0.01	**0.000** 0.0	0.180 0.02	0.012 0.03	0.040 0.03
$D^{\square}$	PO	0.356 0.02	0.001 0.0	0.175 0.03	0.024 0.04	0.069 0.04
	EP	**0.382** 0.01	0.002 0.0	0.173 0.03	0.021 0.04	0.072 0.04
	$PP^{\square}$	0.135 0.01	0.530 0.03	0.084 0.04	0.068 0.04	0.116 0.04
	$PP^{\star}$	0.140 0.01	0.178 0.02	0.120 0.04	0.060 0.04	0.091 0.04
	$D^{\square}$	0.295 0.02	**0.000** 0.0	**0.192** 0.02	**0.015** 0.02	**0.041** 0.03
	$D^{\star}$	0.358 0.02	**0.000** 0.0	0.183 0.02	0.016 0.03	0.050 0.03

(c) X-MART

Acc^{rob}	FR	For	Vul	Dev
0.944 0.01	**0.002** 0.0	**0.192** 0.02	0.018 0.03	0.044 0.03
0.967 0.0	**0.002** 0.0	0.176 0.03	0.013 0.03	0.044 0.03
1.000 0.0	0.983 0.0	0.032 0.02	0.079 0.03	0.142 0.03
0.999 0.0	0.986 0.0	0.040 0.02	0.080 0.03	0.137 0.03
0.991 0.0	0.004 0.0	0.178 0.02	**0.011** 0.02	**0.035** 0.02
0.985 0.0	0.005 0.0	0.160 0.02	0.013 0.03	0.044 0.03
0.280 0.02	0.001 0.0	**0.197** 0.02	0.020 0.03	0.055 0.04
0.343 0.02	**0.000** 0.0	0.185 0.02	**0.017** 0.03	0.052 0.04
0.003 0.0	0.962 0.0	0.034 0.02	0.093 0.03	0.146 0.03
0.003 0.0	0.964 0.01	0.041 0.02	0.092 0.03	0.144 0.03
0.285 0.02	**0.000** 0.0	0.193 0.02	**0.017** 0.02	**0.043** 0.03
0.306 0.02	**0.000** 0.0	0.174 0.02	0.021 0.03	0.052 0.03

(d) RAR

Appendix B. Attack Algorithm Selection

In this section, we provide additional details on the selection of the attack strategy for explanation-aware prediction-untargeted attacks (cf. Section 4.2). We conduct a hyperparameter search for each prediction-untargeted attack scenario ($D^{\square}$, $D^{\star}$, and EP) to determine the optimal attack strategy. In particular, we investigate three attack strategies: *trivial*, *random*, and *highest-wrong*. For each attack strategy, we search for an optimal combination of learning rate η and weight γ using a grid search with `optuna` [2] for 1,000 trials. We select the best hyperparameters based on the robust accuracy (Acc^{rob}) and the forgeability. Specifically, we use an equally weighted sum of the respective min-max normalized MSE values. To measure these metrics, we attack a pre-trained model on the first 2,000 samples of the test dataset. We use the best hyperparameters for each attack strategy and scenario to attack the pre-trained model on the entire test dataset. Based on these three hyperparameter searches, we select the highest-wrong strategy as the best strategy.

TABLE 9: Hyperparameters found for the explanation-aware extensions for each attack scenario.

	Attack	LR	λ_{p}	λ_{e}^1	λ_{e}^2
X-PGD-AT	PO	0.0200	–	–	4.0
	$D^{\square}$	0.0500	–	–	1.0
	$D^{\star}$	0.1000	–	–	8.0
	EP	0.1000	–	–	0.5
	$PP^{\square}$	0.0100	–	–	1.0
	$PP^{\star}$	0.0100	–	–	1.0
RAR	PO	0.0200	–	–	1.0
	$D^{\square}$	0.0100	–	–	1.0
	$D^{\star}$	0.0100	–	–	0.5
	EP	0.0200	–	–	1.0
	$PP^{\square}$	0.0010	–	–	4.0
	$PP^{\star}$	0.0050	–	–	4.0

	Attack	LR	λ_{p}	λ_{e}^1	λ_{e}^2
X-TRADES	PO	0.0500	0.5	1.0	2.0
	$D^{\square}$	0.0100	2.0	8.0	1.0
	$D^{\star}$	0.0200	1.0	4.0	0.5
	EP	0.0200	1.0	4.0	0.5
	$PP^{\square}$	0.0050	0.5	0.5	8.0
	$PP^{\star}$	0.0500	1.0	2.0	4.0
X-MART	PO	0.0020	1.0	4.0	0.5
	$D^{\square}$	0.0005	0.5	2.0	0.5
	$D^{\star}$	0.0010	0.5	8.0	2.0
	EP	0.0010	0.5	8.0	2.0
	$PP^{\square}$	0.0005	8.0	2.0	2.0
	$PP^{\star}$	0.0020	4.0	1.0	0.5

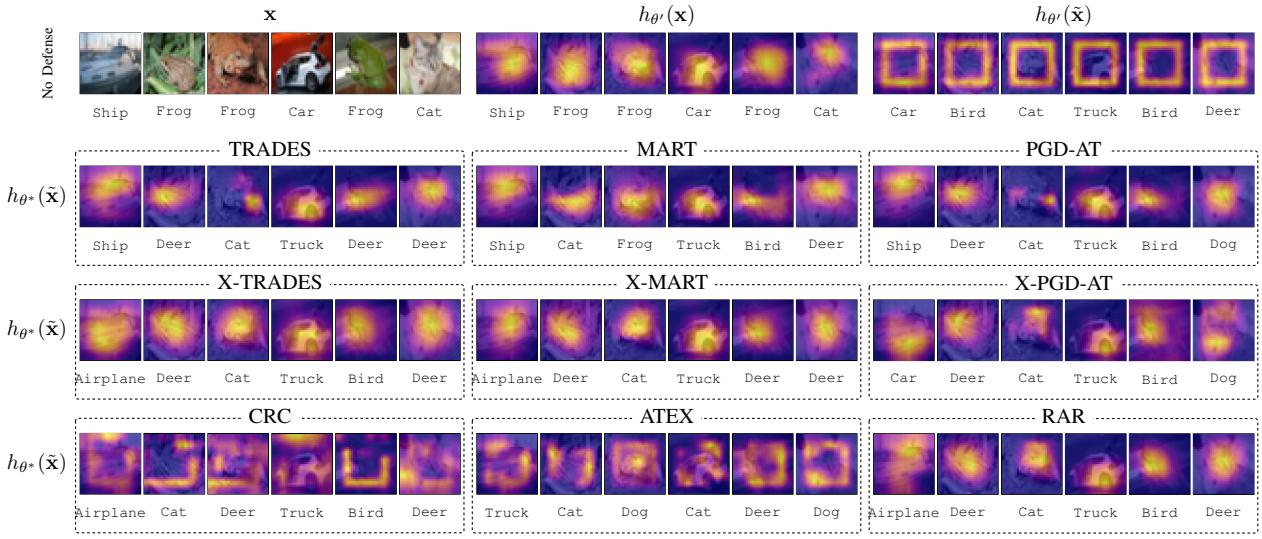


Figure 6: Results of the different defense methods against the $D^{\square}$ attack scenario for five different images from the CIFAR-10 test dataset (first column) and their explanations (second column). Complementary to Fig. 1 in the main paper, which only shows three examples.

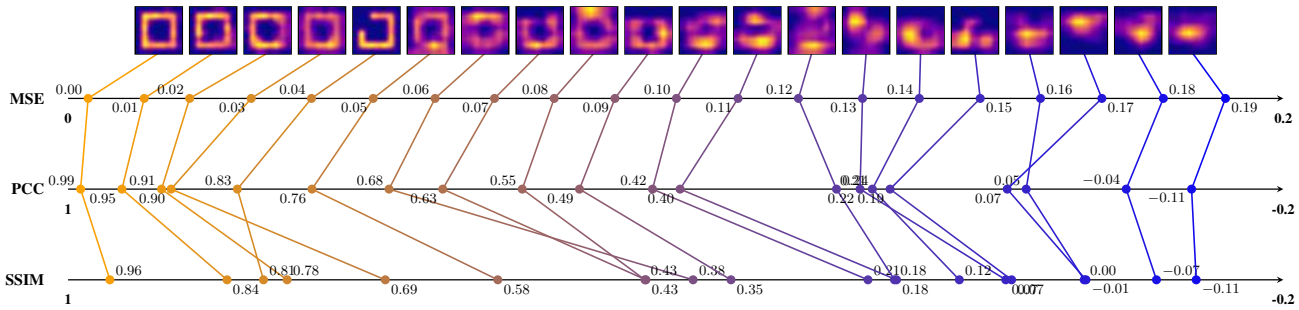


Figure 7: Visual comparison of the three metrics used in this work: MSE, PCC, and SSIM using sample explanations. Values for each metric are computed between the shown explanations and the square target explanation, with the leftmost explanation being the closest to the target. The scales for PCC and SSIM are inverted to allow for better comparison.

Appendix C. Hyperparameters

The following section provides an overview of the hyperparameters for the clean training (Section C.1), the attacks during the fine-tuning with adversarial attacks (Section C.2), the other hyperparameters of the fine-tuning processes (Section C.3), and the hyperparameters of the attacks during the evaluation (Section C.4). Again, we use `optuna` [2] in all settings to find the best hyperparameters.

C.1. Clean Pre-Training

To train the clean ResNet20 models [23] on the CIFAR-10 dataset [31], we use an SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001 for 200 epochs. The starting learning rate is set to 0.1 and is reduced by a factor of 10 at epochs 100 and 150. Additionally, we employ a batch size of 128 and the standard preprocessing transforms provided by PyTorch for ResNet models. During training, we evaluate the model after every epoch and save the best model according to the clean accuracy ($\text{Acc}^{\text{clean}}$) on the validation split. In total, we train three clean models, which we use as the baseline for the evaluation and as starting points for defensive fine-tuning approaches.

C.2. Attack Hyperparameters during Fine-Tuning

For each explanation-aware fine-tuning scenario, we perform a hyperparameter search to identify the optimal adversaries to fine-tune against. While the loss terms L_{pred} and L_{expl} differ for different attack modes, γ and the learning rate η used to minimize this loss remain the only two hyperparameters relevant for searching across all attacks. The number of iterations N used in the

TABLE 10: Hyperparameters of the attacks used during fine-tuning. All attacks use $\epsilon = 0.031$.

Attack	N	η	γ
PO	7	0.0078	-
EP	200	0.0011	0.9918
$PP^{\square}$	200	0.0018	0.9912
$PP^{\star}$	200	0.0019	0.9668
$D^{\square}$	200	0.0010	0.9892
$D^{\star}$	200	0.0015	0.9852

optimization process is manually selected to minimize compute time while preserving attack effectiveness (cf. Fig. 2). For all explanation-aware attack modes, we run 1000 trials and select the best hyperparameters based on the equally weighted sum of the robust accuracy (Acc^{rob}) and forgeability according to MSE. To measure these metrics while saving computational costs, we attack the first 2000 samples in the test dataset. For the PO attack, we employ the parameters proposed by [35]. The hyperparameters for all attacks used during fine-tuning are shown in Table 10. The weighting is done according to Table 12.

C.3. Hyperparameters of the Fine-Tuning

The following paragraphs describe both pre-defined and searched hyperparameters for the methods used in this work. These pre-defined hyperparameters for existing methods are chosen according to the respective original works. For the explanation-aware extensions, we choose the same pre-defined hyperparameters as for their vanilla counterparts. Specifically, these pre-defined hyperparameters include the batch size, the beta smoothing factor, the number of fine-tuning epochs, and the optimizer. The batch size is set to 1,000 for all explanation-aware extensions and to 128 for all other methods. Beta smoothing as introduced by [16] is set to 3 for ATEX and CRC, and to 8 for all other methods. The number of fine-tuning epochs is set to 10 for all methods, except for ATEX and CRC, which are fine-tuned for 200 epochs. The optimizer used during fine-tuning is SGD for all methods, except for ATEX and CRC, which use Adam. These differences between ATEX and CRC and the other methods stem from the different choice of hyperparameters in the original works, which we follow to ensure a fair comparison. To determine the remaining hyperparameters for the different explanation-aware extensions used in this work, we run grid searches for each. Specifically, this includes the learning rate and the different loss weights λ_p, λ_e^1, and λ_e^2. The grid search employs the `GridSampler` from `optuna` and runs every combination of predefined hyperparameters to be searched. The search space is defined logarithmically, spanning from 10^{-5} to 5×10^{-2} for the learning rate and from 0.5 to 8 for λ_p, λ_e^1, and λ_e^2. For explanation-aware extensions, we conduct one grid search for each attack mode. This is necessary due to the different natures of attacks used during fine-tuning: e.g., while some attack modes maximize the difference in predictions (e.g., dual attacks), others minimize this difference (prediction-preserving attacks). Therefore, the corresponding loss terms of the explanation-aware extension may have to be weighted differently for the different attack modes. The model is fine-tuned with attacks on the training dataset and evaluated with the same attack scenario on the validation dataset. For evaluation, the same attack hyperparameters as for fine-tuning are used, except that the number of attack iterations N is increased from 200 during fine-tuning to 500 during evaluation. This is done to evaluate against stronger attacks and to obtain a more accurate measure of the model's robustness. To save computational costs, we first fine-tune models for 3 epochs for each hyperparameter trial. For the best trials, we extend fine-tuning for another 3 epochs to determine the best hyperparameters. The best trials to extend are selected by `optuna`, which calculates the Pareto front of the trials by the given metrics to optimize. To determine the best hyperparameter combination, we use a weighted sum of different metrics specified in Table 11. For X-PGD-AT and RAR, we use 30 trials. Since X-TRADES and X-MART have more hyperparameters to search, we use 60 trials. The hyperparameters for each method and attack scenario are listed in Table 9. Note that we do not conduct hyperparameter searches for ATEX and CRC, and instead employ the hyperparameters from the original works.

C.4. Attack Hyperparameters during Evaluation

An adversary is generally expected to be able to adapt themselves to the specific environment they attacks, such as the specific model. To evaluate the models that were fine-tuned with the different methods presented in this work, we conduct a hyperparameter search for each robustified model. These hyperparameter searches use the same methodology as described in Section C.2. However, we only run 100 trials for each model instead of 1,000, but use 500 adversary iterations, as we expect the adversary to be able to use more iterations than the defender. Note that this hyperparameter search is conducted for each combination of defense method and attack scenario, representing the majority of the total computational cost to evaluate our work. The best learning rate η and weight γ are selected as described in Section C.2. These hyperparameters are then used to evaluate the model on the test dataset. Since in the whole evaluation pipeline we train 435 models and therefore have to conduct 435 hyperparameter searches, we do not list the individual hyperparameters here.

TABLE 11: Weights for the hyperparameter searches of the explanation-aware extension. The arrows indicate whether a metric should be minimized or maximized.

Attack	$\text{Acc}^{\text{clean}}$	Acc^{rob}	For	Vul	Fid	Dev
PO	$\uparrow \frac{1}{2}$	$\uparrow \frac{1}{2}$	–	–	–	–
EP	$\uparrow \frac{1}{6}$	$\uparrow \frac{1}{6}$	–	$\downarrow \frac{1}{6}$	$\downarrow \frac{1}{6}$	$\downarrow \frac{2}{6}$
$PP^{\square}$	$\uparrow \frac{1}{8}$	$\uparrow \frac{1}{8}$	$\uparrow \frac{2}{8}$	$\downarrow \frac{1}{8}$	$\downarrow \frac{1}{8}$	$\downarrow \frac{2}{8}$
$PP^{\star}$	$\uparrow \frac{1}{6}$	$\uparrow \frac{1}{6}$	–	$\downarrow \frac{1}{6}$	$\downarrow \frac{1}{6}$	$\downarrow \frac{2}{6}$
$D^{\square}$	$\uparrow \frac{1}{8}$	$\uparrow \frac{1}{8}$	$\uparrow \frac{2}{8}$	$\downarrow \frac{1}{8}$	$\downarrow \frac{1}{8}$	$\downarrow \frac{2}{8}$
$D^{\star}$	$\uparrow \frac{1}{6}$	$\uparrow \frac{1}{6}$	–	$\downarrow \frac{1}{6}$	$\downarrow \frac{1}{6}$	$\downarrow \frac{2}{6}$

TABLE 12: Weights for the hyperparameter searches of the attacks. The arrows indicate whether a metric should be minimized or maximized.

Attack	Acc^{rob}	For	Vul
PO	$\downarrow 1$	–	–
EP	$\downarrow \frac{1}{2}$	–	$\downarrow \frac{1}{2}$
$PP^{\square}$	$\uparrow \frac{1}{2}$	$\downarrow \frac{1}{2}$	–
$PP^{\star}$	$\uparrow \frac{1}{2}$	–	$\uparrow \frac{1}{2}$
$D^{\square}$	$\downarrow \frac{1}{2}$	$\downarrow \frac{1}{2}$	–
$D^{\star}$	$\downarrow \frac{1}{2}$	–	$\uparrow \frac{1}{2}$

TABLE 13: Quantitative results for the target transferability of explanation-aware attacks for the MSE metric. Numerical counterpart to Fig. 5 in the main paper, which only shows bars.

	Defense	Attack	$\text{Acc}^{\text{clean}}$	Acc^{rob}	FR	For	Vul	Fid	Dev
X-PGD-AT	◧	◧	0.791 0.01	0.496 0.02	0.002 0.00	0.323 0.10	0.057 0.06	0.047 0.03	0.071 0.05
		◨	0.791 0.01	0.496 0.02	0.010 0.00	0.266 0.09	0.053 0.05	0.047 0.03	0.074 0.05
	◨	◧	0.810 0.01	0.487 0.03	0.006 0.00	0.284 0.09	0.054 0.06	0.047 0.03	0.076 0.05
		◨	0.810 0.01	0.492 0.03	0.007 0.00	0.295 0.10	0.057 0.06	0.047 0.03	0.070 0.05
X-TRADES	◧	◧	0.833 0.01	0.556 0.02	0.001 0.00	0.270 0.07	0.027 0.04	0.047 0.03	0.061 0.04
		◨	0.833 0.01	0.570 0.02	0.000 0.00	0.333 0.07	0.027 0.03	0.047 0.03	0.052 0.03
	◨	◧	0.835 0.01	0.572 0.02	0.000 0.00	0.351 0.07	0.027 0.03	0.045 0.03	0.051 0.03
		◨	0.835 0.01	0.567 0.01	0.002 0.00	0.255 0.07	0.025 0.03	0.045 0.03	0.057 0.04
X-MART	◧	◧	0.841 0.01	0.544 0.02	0.001 0.00	0.304 0.09	0.045 0.05	0.053 0.03	0.071 0.04
		◨	0.841 0.01	0.545 0.02	0.002 0.00	0.279 0.08	0.043 0.05	0.053 0.03	0.070 0.04
	◨	◧	0.843 0.01	0.540 0.02	0.001 0.00	0.298 0.09	0.043 0.05	0.051 0.03	0.072 0.04
		◨	0.843 0.01	0.537 0.02	0.001 0.00	0.291 0.08	0.042 0.05	0.051 0.03	0.067 0.04
RAR	◧	◧	0.826 0.01	0.541 0.02	0.002 0.00	0.332 0.10	0.052 0.06	0.051 0.03	0.069 0.04
		◨	0.826 0.01	0.536 0.01	0.007 0.00	0.268 0.08	0.047 0.05	0.051 0.03	0.073 0.04
	◨	◧	0.831 0.01	0.534 0.02	0.002 0.00	0.284 0.09	0.046 0.05	0.051 0.03	0.075 0.05
		◨	0.831 0.01	0.534 0.02	0.002 0.00	0.317 0.09	0.049 0.05	0.051 0.03	0.067 0.04

TABLE 14: The effectiveness of our investigated defense techniques against explanation-aware attacks for each attack scenario for the MSE metric. For the PO attack, no bold highlights are used for the vulnerability and the deviation as PO attackers ignore explanations.

Attack	Defense	$\text{Acc}^{\text{clean}}$	Acc^{rob}	FR	For	Vul	Fid	Dev
PO	PGD-AT	0.761 0.01	0.455 0.02	–	–	0.035 0.04	0.044 0.04	0.055 0.04
	TRADES	0.798 0.01	0.412 0.02	–	–	0.034 0.04	0.036 0.03	0.049 0.04
	MART	0.746 0.01	**0.491** 0.02	–	–	0.035 0.04	0.047 0.04	0.057 0.04
	X-PGD-AT	0.696 0.01	0.410 0.02	–	–	0.043 0.05	0.051 0.04	0.061 0.04
	X-TRADES	0.783 0.01	0.328 0.02	–	–	0.016 0.02	0.035 0.03	0.044 0.03
	X-MART	0.702 0.01	0.475 0.02	–	–	0.048 0.05	0.069 0.04	0.077 0.04
	RAR	0.696 0.01	0.409 0.02	–	–	0.043 0.05	0.051 0.04	0.061 0.04
	No Defense	**0.919** 0.01	0.049 0.01	–	–	0.061 0.00	**0.000** 0.00	0.061 0.00
	ATEX	0.878 0.01	0.106 0.01	–	–	0.032 0.02	0.050 0.03	0.058 0.03
	CRC	0.772 0.03	0.099 0.04	–	–	0.116 0.06	0.108 0.05	0.104 0.05
EP	PGD-AT	0.761 0.01	0.350 0.02	–	–	0.017 0.03	0.044 0.04	0.049 0.04
	TRADES	0.798 0.01	0.312 0.02	–	–	0.016 0.02	0.036 0.03	0.043 0.03
	MART	0.746 0.01	0.380 0.02	–	–	0.017 0.03	0.047 0.04	0.050 0.04
	X-PGD-AT	0.750 0.01	0.367 0.01	–	–	0.012 0.02	0.045 0.04	0.047 0.03
	X-TRADES	0.803 0.01	0.318 0.02	–	–	0.009 0.01	0.040 0.03	0.043 0.03
	X-MART	0.716 0.02	**0.382** 0.01	–	–	0.019 0.04	0.071 0.05	0.071 0.05
	RAR	0.763 0.01	0.343 0.02	–	–	0.015 0.02	0.051 0.04	0.052 0.04
	No Defense	**0.919** 0.01	0.000 0.00	–	–	0.000 0.00	**0.000** 0.00	**0.000** 0.00
	ATEX	0.878 0.01	0.026 0.01	–	–	0.008 0.01	0.050 0.03	0.048 0.03
	CRC	0.772 0.03	0.013 0.02	–	–	**0.030** 0.03	0.108 0.05	0.102 0.04
$PP^{\square}$	PGD-AT	0.761 0.01	0.959 0.01	0.003 0.00	0.193 0.02	0.014 0.03	0.044 0.04	0.040 0.03
	TRADES	0.798 0.01	0.977 0.01	0.003 0.00	0.185 0.03	0.013 0.03	0.036 0.03	**0.032** 0.03
	MART	0.746 0.01	0.945 0.01	**0.001** 0.00	**0.196** 0.02	0.013 0.03	0.047 0.04	0.040 0.03
	X-PGD-AT	0.710 0.05	**1.000** 0.00	0.999 0.00	0.022 0.01	0.065 0.03	0.099 0.04	0.146 0.03
	X-TRADES	0.881 0.01	**1.000** 0.00	0.002 0.00	0.149 0.02	**0.011** 0.02	0.045 0.04	0.046 0.04
	X-MART	0.835 0.01	0.997 0.00	0.692 0.04	0.069 0.03	0.041 0.03	0.095 0.04	0.111 0.04
	RAR	0.871 0.01	**1.000** 0.00	0.983 0.00	0.032 0.02	0.079 0.03	0.093 0.04	0.142 0.03
	No Defense	**0.919** 0.01	**1.000** 0.00	0.998 0.00	0.012 0.00	0.163 0.00	**0.000** 0.00	0.163 0.00
	ATEX	0.878 0.01	0.996 0.01	0.308 0.04	0.114 0.05	0.067 0.04	0.050 0.03	0.081 0.04
	CRC	0.772 0.03	**1.000** 0.00	0.384 0.22	0.094 0.03	0.059 0.04	0.108 0.05	0.115 0.03
$PP^{\star}$	PGD-AT	0.761 0.01	0.960 0.00	–	–	0.021 0.03	0.044 0.04	0.038 0.03
	TRADES	0.798 0.01	0.980 0.01	–	–	0.023 0.03	0.036 0.03	**0.033** 0.03
	MART	0.746 0.01	0.947 0.01	–	–	**0.017** 0.03	0.047 0.04	0.039 0.03
	X-PGD-AT	0.759 0.08	**1.000** 0.00	–	–	0.303 0.08	0.125 0.05	0.257 0.07
	X-TRADES	0.853 0.01	0.997 0.00	–	–	0.037 0.04	0.038 0.02	0.061 0.04
	X-MART	0.836 0.01	0.995 0.00	–	–	0.098 0.07	0.074 0.04	0.117 0.06
	RAR	0.815 0.03	**1.000** 0.00	–	–	0.294 0.08	0.120 0.05	0.256 0.07
	No Defense	**0.919** 0.01	**1.000** 0.00	–	–	0.329 0.01	**0.000** 0.00	0.329 0.01
	ATEX	0.878 0.01	0.995 0.00	–	–	0.098 0.05	0.050 0.03	0.093 0.05
	CRC	0.772 0.03	**1.000** 0.00	–	–	0.175 0.09	0.108 0.05	0.186 0.08
$D^{\square}$	PGD-AT	0.761 0.01	0.351 0.02	0.001 0.00	**0.200** 0.02	0.019 0.03	0.044 0.04	0.051 0.04
	TRADES	0.798 0.01	0.315 0.02	**0.000** 0.00	0.197 0.02	0.020 0.03	0.036 0.03	0.045 0.03
	MART	0.746 0.01	**0.380** 0.02	0.001 0.00	0.199 0.02	0.018 0.03	0.047 0.04	0.051 0.04
	X-PGD-AT	0.778 0.02	0.260 0.02	0.005 0.00	0.183 0.03	0.022 0.03	0.037 0.03	0.046 0.03
	X-TRADES	0.814 0.01	0.292 0.02	**0.000** 0.00	0.171 0.02	**0.012** 0.02	0.041 0.03	0.044 0.03
	X-MART	0.804 0.01	0.295 0.02	**0.000** 0.00	0.192 0.02	0.015 0.02	0.037 0.03	**0.041** 0.03
	RAR	0.821 0.01	0.285 0.02	**0.000** 0.00	0.193 0.02	0.017 0.02	0.038 0.03	0.043 0.03
	No Defense	**0.919** 0.01	0.000 0.00	0.998 0.00	0.010 0.00	0.170 0.00	**0.000** 0.00	0.170 0.00
	ATEX	0.878 0.01	0.037 0.01	0.222 0.01	0.128 0.05	0.064 0.04	0.050 0.03	0.078 0.04
	CRC	0.772 0.03	0.013 0.02	0.311 0.20	0.102 0.03	0.091 0.05	0.108 0.05	0.120 0.03
$D^{\star}$	PGD-AT	0.761 0.01	0.350 0.02	–	–	0.023 0.03	0.044 0.04	0.052 0.04
	TRADES	0.798 0.01	0.314 0.02	–	–	0.029 0.04	0.036 0.03	**0.050** 0.04
	MART	0.746 0.01	**0.380** 0.02	–	–	0.021 0.03	0.047 0.04	0.052 0.04
	X-PGD-AT	0.764 0.02	0.276 0.02	–	–	0.042 0.05	0.045 0.03	0.062 0.05
	X-TRADES	0.818 0.01	0.310 0.02	–	–	**0.018** 0.03	0.041 0.03	0.051 0.04
	X-MART	0.751 0.02	0.357 0.02	–	–	0.019 0.03	0.047 0.03	0.052 0.04
	RAR	0.800 0.01	0.306 0.02	–	–	0.033 0.05	0.048 0.03	0.060 0.04
	No Defense	**0.919** 0.01	0.000 0.00	–	–	0.337 0.01	**0.000** 0.00	0.337 0.01
	ATEX	0.878 0.01	0.030 0.01	–	–	0.098 0.05	0.050 0.03	0.098 0.05
	CRC	0.772 0.03	0.013 0.02	–	–	0.219 0.08	0.108 0.05	0.200 0.08

TABLE 15: The effectiveness of our investigated defense techniques against explanation-aware attacks for each attack scenario for the two additional metrics PCC and SSIM. For the PO attack, no bold highlights are used for the vulnerability and the deviation as PO attackers ignore explanations.

Att.	Def.	Accuracy		FR	For		Vul		Fid		Dev	
		clean	robust		PCC	SSIM	PCC	SSIM	PCC	SSIM	PCC	SSIM
PO	PGD-AT	0.761 0.01	0.455 0.02	–	–	–	0.748 0.30	0.592 0.31	0.727 0.25	0.516 0.22	0.666 0.27	0.451 0.23
	TRADES	0.798 0.01	0.412 0.02	–	–	–	0.776 0.27	0.599 0.29	0.783 0.22	0.578 0.21	0.710 0.25	0.489 0.22
	MART	0.746 0.01	**0.491** 0.02	–	–	–	0.756 0.30	0.605 0.31	0.713 0.25	0.502 0.22	0.647 0.28	0.436 0.23
	X-PGD-AT	0.696 0.01	0.410 0.02	–	–	–	0.695 0.34	0.547 0.32	0.684 0.27	0.478 0.23	0.624 0.29	0.416 0.23
	X-TRADES	0.783 0.01	0.328 0.02	–	–	–	0.877 0.19	0.729 0.24	0.811 0.18	0.590 0.19	0.753 0.21	0.523 0.20
	X-MART	0.702 0.01	0.475 0.02	–	–	–	0.640 0.40	0.517 0.37	0.538 0.30	0.350 0.21	0.488 0.31	0.310 0.21
	RAR	0.696 0.01	0.409 0.02	–	–	–	0.695 0.34	0.548 0.32	0.685 0.27	0.479 0.23	0.624 0.29	0.417 0.23
	No Defense	**0.919** 0.01	0.049 0.01	–	–	–	0.675 0.01	0.480 0.01	**1.000** 0.00	**1.000** 0.00	0.675 0.01	0.480 0.01
	ATEX	0.878 0.01	0.106 0.01	–	–	–	0.759 0.19	0.507 0.18	0.760 0.16	0.423 0.14	0.689 0.20	0.400 0.16
	CRC	0.772 0.03	0.099 0.04	–	–	–	0.277 0.33	0.186 0.19	0.337 0.34	0.209 0.19	0.533 0.27	0.325 0.18
EP	PGD-AT	0.761 0.01	0.350 0.02	–	–	–	0.880 0.20	0.751 0.25	0.727 0.25	0.516 0.22	0.701 0.24	0.479 0.21
	TRADES	0.798 0.01	0.312 0.02	–	–	–	0.894 0.17	0.747 0.23	0.783 0.22	0.578 0.21	0.747 0.21	0.522 0.21
	MART	0.746 0.01	0.380 0.02	–	–	–	0.884 0.19	0.758 0.25	0.713 0.25	0.502 0.22	0.694 0.24	0.473 0.21
	X-PGD-AT	0.750 0.01	0.367 0.01	–	–	–	0.922 0.16	0.814 0.21	0.734 0.24	0.508 0.21	0.724 0.22	0.492 0.20
	X-TRADES	0.803 0.01	0.318 0.02	–	–	–	0.956 0.08	0.888 0.12	0.808 0.16	0.571 0.17	0.793 0.16	0.563 0.17
	X-MART	0.716 0.02	**0.382** 0.01	–	–	–	0.867 0.26	0.763 0.27	0.543 0.34	0.364 0.25	0.544 0.33	0.363 0.24
	RAR	0.763 0.01	0.343 0.02	–	–	–	0.913 0.15	0.804 0.20	0.704 0.24	0.496 0.21	0.702 0.22	0.492 0.20
	No Defense	**0.919** 0.01	0.000 0.00	–	–	–	0.997 0.00	0.984 0.00	**1.000** 0.00	**1.000** 0.00	**0.997** 0.00	**0.984** 0.00
	ATEX	0.878 0.01	0.026 0.01	–	–	–	0.931 0.10	0.770 0.15	0.760 0.16	0.423 0.14	0.758 0.17	0.449 0.15
	CRC	0.772 0.03	0.013 0.02	–	–	–	**0.754** 0.25	**0.578** 0.29	0.337 0.34	0.209 0.19	0.353 0.32	0.220 0.18
$PP^{\square}$	PGD-AT	0.761 0.01	0.959 0.01	0.003 0.00	-0.031 0.13	-0.066 0.08	0.892 0.23	0.806 0.26	0.727 0.25	0.516 0.22	0.750 0.23	0.543 0.22
	TRADES	0.798 0.01	0.977 0.01	0.003 0.00	0.007 0.14	-0.048 0.09	0.903 0.19	0.802 0.23	0.783 0.22	0.578 0.21	**0.799** 0.20	**0.597** 0.20
	MART	0.746 0.01	0.945 0.01	**0.001** 0.00	**-0.043** 0.12	**-0.074** 0.08	0.906 0.21	**0.818** 0.25	0.713 0.25	0.502 0.22	0.754 0.22	0.545 0.21
	X-PGD-AT	0.710 0.05	**1.000** 0.00	0.999 0.00	0.931 0.06	0.769 0.10	0.522 0.16	0.443 0.16	0.340 0.29	0.124 0.15	0.039 0.13	-0.068 0.10
	X-TRADES	0.881 0.01	**1.000** 0.00	0.002 0.00	0.135 0.12	0.053 0.09	**0.914** 0.18	0.783 0.20	0.702 0.28	0.507 0.24	0.678 0.27	0.458 0.21
	X-MART	0.835 0.01	0.997 0.00	0.692 0.04	0.641 0.21	0.459 0.19	0.695 0.26	0.563 0.23	0.417 0.30	0.177 0.19	0.213 0.25	0.023 0.15
	RAR	0.871 0.01	**1.000** 0.00	0.983 0.00	0.873 0.10	0.697 0.14	0.427 0.20	0.324 0.18	0.523 0.26	0.238 0.17	0.037 0.15	-0.065 0.10
	No Defense	**0.919** 0.01	**1.000** 0.00	0.998 0.00	0.948 0.01	0.877 0.02	0.023 0.02	-0.053 0.01	**1.000** 0.00	**1.000** 0.00	0.023 0.02	-0.053 0.01
	ATEX	0.878 0.01	0.996 0.01	0.308 0.04	0.419 0.28	0.292 0.23	0.499 0.29	0.296 0.25	0.760 0.16	0.423 0.14	0.496 0.24	0.227 0.18
	CRC	0.772 0.03	**1.000** 0.00	0.384 0.22	0.452 0.23	0.319 0.19	0.526 0.35	0.396 0.27	0.337 0.34	0.209 0.19	0.127 0.25	0.041 0.14
$PP^{\star}$	PGD-AT	0.761 0.01	0.960 0.00	–	–	–	0.857 0.25	0.743 0.26	0.727 0.25	0.516 0.22	0.775 0.22	0.563 0.21
	TRADES	0.798 0.01	0.980 0.01	–	–	–	0.863 0.22	0.728 0.24	0.783 0.22	0.578 0.21	**0.812** 0.19	**0.606** 0.20
	MART	0.746 0.01	0.947 0.01	–	–	–	**0.889** 0.22	**0.789** 0.25	0.713 0.25	0.502 0.22	0.769 0.22	0.558 0.21
	X-PGD-AT	0.759 0.08	**1.000** 0.00	–	–	–	-0.445 0.37	-0.054 0.12	0.246 0.34	0.127 0.19	-0.125 0.31	-0.025 0.14
	X-TRADES	0.853 0.01	0.997 0.00	–	–	–	0.799 0.21	0.679 0.22	0.784 0.16	0.528 0.17	0.692 0.21	0.453 0.18
	X-MART	0.836 0.01	0.995 0.00	–	–	–	0.518 0.39	0.428 0.27	0.565 0.29	0.341 0.21	0.446 0.32	0.247 0.20
	RAR	0.815 0.03	**1.000** 0.00	–	–	–	-0.397 0.40	-0.040 0.13	0.302 0.32	0.147 0.19	-0.117 0.31	-0.027 0.14
	No Defense	**0.919** 0.01	**1.000** 0.00	–	–	–	-0.473 0.01	-0.183 0.00	**1.000** 0.00	**1.000** 0.00	-0.473 0.01	-0.183 0.00
	ATEX	0.878 0.01	0.995 0.00	–	–	–	0.484 0.29	0.261 0.19	0.760 0.16	0.423 0.14	0.529 0.27	0.280 0.18
	CRC	0.772 0.03	**1.000** 0.00	–	–	–	0.223 0.46	0.227 0.26	0.337 0.34	0.209 0.19	0.224 0.36	0.127 0.18
$D^{\square}$	PGD-AT	0.761 0.01	0.351 0.02	0.001 0.00	**-0.062** 0.12	-0.084 0.07	0.862 0.22	0.733 0.27	0.727 0.25	0.516 0.22	0.687 0.25	0.469 0.22
	TRADES	0.798 0.01	0.315 0.02	**0.000** 0.00	-0.040 0.13	-0.074 0.08	0.864 0.20	0.714 0.25	0.783 0.22	0.578 0.21	0.725 0.23	0.504 0.22
	MART	0.746 0.01	**0.380** 0.02	0.001 0.00	-0.056 0.11	-0.080 0.07	0.877 0.20	0.751 0.26	0.713 0.25	0.502 0.22	0.687 0.24	0.467 0.21
	X-PGD-AT	0.778 0.02	0.260 0.02	0.005 0.00	0.010 0.14	-0.050 0.10	0.862 0.18	0.709 0.23	0.779 0.20	0.573 0.20	0.714 0.22	0.496 0.20
	X-TRADES	0.814 0.01	0.292 0.02	**0.000** 0.00	0.089 0.07	0.002 0.06	**0.935** 0.10	**0.857** 0.15	0.808 0.16	0.563 0.16	**0.780** 0.17	0.543 0.17
	X-MART	0.804 0.01	0.295 0.02	**0.000** 0.00	-0.039 0.07	**-0.089** 0.05	0.915 0.12	0.805 0.18	0.777 0.19	0.575 0.19	0.758 0.18	**0.553** 0.18
	RAR	0.821 0.01	0.285 0.02	**0.000** 0.00	-0.027 0.09	-0.083 0.07	0.905 0.14	0.774 0.20	0.779 0.19	0.572 0.19	0.744 0.19	0.531 0.19
	No Defense	**0.919** 0.01	0.000 0.00	0.998 0.00	0.959 0.01	0.895 0.02	0.006 0.01	-0.058 0.01	**1.000** 0.00	**1.000** 0.00	0.006 0.01	-0.058 0.01
	ATEX	0.878 0.01	0.037 0.01	0.222 0.01	0.323 0.28	0.203 0.22	0.498 0.29	0.274 0.25	0.760 0.16	0.423 0.14	0.501 0.24	0.241 0.18
	CRC	0.772 0.03	0.013 0.02	0.311 0.20	0.405 0.24	0.284 0.19	0.262 0.34	0.176 0.24	0.337 0.34	0.209 0.19	0.099 0.25	0.040 0.14
$D^{\star}$	PGD-AT	0.761 0.01	0.350 0.02	–	–	–	0.838 0.24	0.707 0.28	0.727 0.25	0.516 0.22	0.688 0.25	0.468 0.22
	TRADES	0.798 0.01	0.314 0.02	–	–	–	0.823 0.24	0.669 0.28	0.783 0.22	0.578 0.21	0.713 0.24	0.493 0.22
	MART	0.746 0.01	**0.380** 0.02	–	–	–	0.860 0.22	0.734 0.27	0.713 0.25	0.502 0.22	0.686 0.25	0.466 0.22
	X-PGD-AT	0.764 0.02	0.276 0.02	–	–	–	0.784 0.26	0.642 0.28	0.741 0.22	0.530 0.20	0.673 0.24	0.462 0.21
	X-TRADES	0.818 0.01	0.310 0.02	–	–	–	**0.913** 0.14	**0.823** 0.18	0.804 0.16	0.553 0.16	**0.764** 0.18	**0.524** 0.17
	X-MART	0.751 0.02	0.357 0.02	–	–	–	0.893 0.18	0.783 0.23	0.711 0.23	0.508 0.20	0.696 0.22	0.490 0.19
	RAR	0.800 0.01	0.306 0.02	–	–	–	0.829 0.23	0.706 0.26	0.728 0.21	0.511 0.19	0.682 0.23	0.468 0.19
	No Defense	**0.919** 0.01	0.000 0.00	–	–	–	-0.491 0.01	-0.181 0.00	**1.000** 0.00	**1.000** 0.00	-0.491 0.01	-0.181 0.00
	ATEX	0.878 0.01	0.030 0.01	–	–	–	0.453 0.27	0.231 0.20	0.760 0.16	0.423 0.14	0.491 0.28	0.254 0.17
	CRC	0.772 0.03	0.013 0.02	–	–	–	-0.040 0.41	0.055 0.22	0.337 0.34	0.209 0.19	0.129 0.36	0.088 0.17

TABLE 16: Additional results for the target transferability of explanation-aware attacks for the metrics PCC and SSIM. Complementary to Fig. 5 in the main paper, which only shows the MSE distances.

	Def	Att.	Accuracy		FR	For		Vul		Fid		Dev	
			clean	robust		PCC	SSIM	PCC	SSIM	PCC	SSIM	PCC	SSIM
X-PGD-AT	◧	◧	0.791 0.01	0.496 0.02	0.002 0.00	0.085 0.35	0.056 0.09	0.668 0.35	0.527 0.31	0.745 0.22	0.522 0.20	0.584 0.28	0.377 0.21
		◨	0.791 0.01	0.496 0.02	0.010 0.00	0.298 0.32	0.114 0.11	0.706 0.32	0.543 0.30	0.745 0.22	0.522 0.20	0.578 0.28	0.372 0.21
	◨	◧	0.810 0.01	0.487 0.03	0.006 0.00	0.230 0.33	0.099 0.10	0.689 0.33	0.534 0.30	0.745 0.22	0.523 0.20	0.553 0.29	0.359 0.21
		◨	0.810 0.01	0.492 0.03	0.007 0.00	0.189 0.33	0.079 0.09	0.670 0.34	0.524 0.31	0.745 0.22	0.523 0.20	0.596 0.28	0.383 0.21
X-TRADES	◧	◧	0.833 0.01	0.556 0.02	0.001 0.00	0.253 0.26	0.075 0.08	0.846 0.22	0.722 0.25	0.765 0.18	0.523 0.17	0.657 0.25	0.446 0.19
		◨	0.833 0.01	0.570 0.02	0.000 0.00	-0.000 0.27	0.022 0.06	0.836 0.22	0.720 0.25	0.765 0.18	0.523 0.17	0.710 0.22	0.473 0.19
	◨	◧	0.835 0.01	0.572 0.02	0.000 0.00	-0.073 0.27	0.007 0.06	0.841 0.22	0.725 0.25	0.781 0.17	0.535 0.17	0.714 0.22	0.478 0.19
		◨	0.835 0.01	0.567 0.01	0.002 0.00	0.306 0.25	0.088 0.08	0.858 0.21	0.732 0.25	0.781 0.17	0.535 0.17	0.691 0.23	0.466 0.19
X-MART	◧	◧	0.841 0.01	0.544 0.02	0.001 0.00	0.112 0.31	0.048 0.07	0.734 0.30	0.602 0.29	0.731 0.21	0.491 0.18	0.597 0.26	0.383 0.19
		◨	0.841 0.01	0.545 0.02	0.002 0.00	0.211 0.31	0.079 0.08	0.749 0.29	0.606 0.29	0.731 0.21	0.491 0.18	0.603 0.25	0.382 0.19
	◨	◧	0.843 0.01	0.540 0.02	0.001 0.00	0.143 0.32	0.064 0.08	0.747 0.28	0.608 0.28	0.736 0.20	0.499 0.18	0.586 0.26	0.379 0.19
		◨	0.843 0.01	0.537 0.02	0.001 0.00	0.166 0.30	0.058 0.08	0.753 0.28	0.616 0.28	0.736 0.20	0.499 0.18	0.622 0.25	0.400 0.19
RAR	◧	◧	0.826 0.01	0.541 0.02	0.002 0.00	0.044 0.34	0.043 0.08	0.701 0.33	0.558 0.31	0.733 0.21	0.503 0.19	0.593 0.27	0.379 0.20
		◨	0.826 0.01	0.536 0.01	0.007 0.00	0.285 0.31	0.113 0.11	0.740 0.29	0.576 0.29	0.733 0.21	0.503 0.19	0.585 0.27	0.371 0.20
	◨	◧	0.831 0.01	0.534 0.02	0.002 0.00	0.229 0.31	0.093 0.10	0.743 0.29	0.582 0.29	0.729 0.21	0.503 0.19	0.563 0.28	0.364 0.20
		◨	0.831 0.01	0.534 0.02	0.002 0.00	0.098 0.32	0.054 0.08	0.722 0.31	0.573 0.30	0.729 0.21	0.503 0.19	0.614 0.26	0.392 0.20

KIT Scientific Working Papers
ISSN 2194-1629