Empowering a Robust Model with Stable and Object-Aligned Explanations

Sowrya Gali Anindya Sarkar Vineeth N Balasubramanian
Indian Institute of Technology Hyderabad
Hyderabad, Telangana, India
cs18btech11012@iith.ac.in, anindyasarkar.ece@gmail.com, vineethnb@cse.iith.ac.in

Abstract

The current state-of-the-art adversarially robust models have more object-aligned attributions, and attributionally robust models have stabler attributions than adversarially trained models. However, these robust models' attributions suffer from another stability-alignment tradeoff problem, i.e., these models' attributions are either stable against explanation-based attacks like IFIA at the cost of relevance or have well object-aligned attributions but unstable against the explanation-based attacks. We propose a training strategy that enforces attribution alignment through teacher saliency within the robust attribution training framework to curb this tradeoff. Moreover, we also note that the current evaluation metrics for measuring the stability of the attribution maps do not consider the object alignment of the generated attribution map and propose new metrics that capture both facets of the attribution maps, i.e., stability and alignment.

1. Introduction

Deep Neural Networks (DNNs) have emerged as the de facto technique for vision and have myriad applications in various domains, including autonomous navigation, surveillance, medical diagnosis, etc. However, there has been a significant impedance in deploying these models in mission-critical applications due to their vulnerability to adversarial – and more recently, attributional – attacks like PGD [18], AutoAttack [9], C&W [3], IFIA [12], etc. Moreover, it has been shown that explanations generated by these networks are fragile and can be disturbed easily by explanation-based attacks like IFIA [12]. From a viewpoint of *adversarial robustness*, Adversarial Training (AT) [18] is the most widely used method, and most improvements on AT have since been developed by adding regularizers without changing the basic min-max formulation [2, 15, 18, 20, 27, 29, 31, 35]. Prior research works have also demonstrated that attribution (or saliency) maps generated by adversarially trained networks are more interpretable and object-aligned [11]. From another perspective of robustness, researchers have recently shown that both natural and adversarially trained models generate different attribution maps for two similar-looking input images with minor perturbations [12] (while maintaining the predicted class). To handle this issue, training strategies such as RAR [7] and ERAR [21] have been proposed, which try to make DNNs capable of producing similar attribution maps for visually indistinguishable images with perturbations. A DNN trained with such methods has been termed to be *attributionally robust*. Our studies in this work reveal that these strategies (adversarial and attributional robustness) bear a tradeoff. On one hand, an adversarially trained model suffers from instability to explanation attacks and produces dissimilar attributions. The inability to provide a stable attribution map against l_∞- norm based explanation attack is conspicuous in Fig 1 (*right half*). On the other hand, although an attributionally robust model provides stable attributions, they are often not object-aligned, as illustrated in Fig 1 (*left half*). We also confirmed this inference using a user study with ten subjects, all of whom unanimously agreed that the attribution maps obtained using the RAR model, such as on Fig 1 (left), were not object-aligned. A primary reason that the above issues have not been highlighted in previous efforts is the metrics used to evaluate attributional robustness, viz., Top-K intersection, Spearman and Kendall's correlation between attribution maps before and after the attack. These metrics measure the stability of attribution maps before and after explanation-based attacks but do not consider the object-alignment of the generated attributions. Besides, the regularizers used for attributional training implicitly force the model towards generating sparser attributions since they try to constrain the norms between gradients (integrated gradient, IG [25], in particular) of natural and attacked images. Such a loss function can move the model towards degenerate solutions and affect its attribution. While one can argue that sparsity removes spurious correlations and hence can be desirable, sparsity should not come at the expense of object-alignment, which is the focus of this work. Such sparsity should ideally be enabled through pruning of background pixels to eliminate spurious correlations and not in the object's pixels. Moreover, it has been shown that as a model's attributions become more aligned with the object, its robustness to adversarial examples is also improved [11]. In this work, we aim to address the issue of the stability-alignment tradeoff between adversarial robustness and attributional robustness. In particular, we leverage saliency matching with off-the-shelf pre-trained teacher networks that generate well-aligned and interpretable attributions to guide the training of a given network. We also propose newer metrics for evaluating the robustness of attribution maps against an attributional attack that consider the stability and object-alignment of the attribution maps.

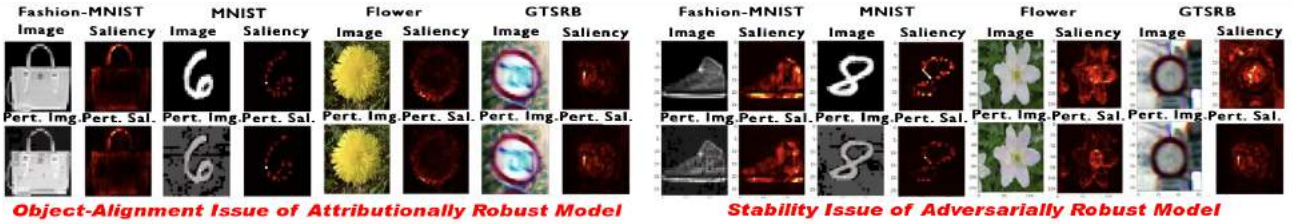

Figure 1. Original image *(top row)* and IFIA [12] attacked image with same predicted class *(bottom row)* and their corresponding saliency maps obtained using RAR model [7] (***left half***) and AT model [18] (***right half***). Note that the RAR model generates similar saliency maps before and after an attack, which are not object-aligned. In contrast, the AT model generates object-aligned saliency maps but is vulnerable to an attribution attack, IFIA [12].

2. Proposed Methodology

Training : We aim to enhance the attribution map of a robust model without disturbing the model's actual objectives. In other words, for an adversarially robust network, the training strategy should not damage its robustness to adversarial attacks and, at the same time, must improve the robustness to explanation-based attacks. Similarly, for attributionally robust networks, the training strategy should not hamper its robustness to the IFIA attack and, at the same time, must enhance the object-alignment of attribution maps. A meaningful way to achieve this is to employ a saliency matching regularizer to ensure object alignment along with a robust attribution regularizer to ensure stability. It is possible to optimize saliency matching regularizer and robust attribution regularizer jointly within a single training objective due to their compatible nature. Our experiments have also revealed that a robust attribution regularizer helps attain stability or robustness to explanation attacks for an adversarially trained model, whereas the saliency matching regularizer enables the model to retain its capability to generate interpretable attribution maps. On the other hand, for an attributionally robust model, the robust attribution regularizer or IG-regularizer helps retain its robustness to explanation attacks, and the saliency matching regularizer allows the model to generate more interpretable attributions without giving up the attributional robustness. However, such a joint training strategy requires a teacher network with interpretable attributions. Prior research has shown that an adversarially trained network is an ideal candidate for such a teacher. We explain this below.

Robustness and Alignment : For an n-class classifier $F(x) = \underset{i}{\operatorname{argmax}}\ \Psi^i(x)$ where $\Psi = (\Psi^1, ..., \Psi^n) : X \longrightarrow \mathbb{R}^n$ be differentiable in x. Then we call $\nabla\Psi^{F(x)}$ the saliency map of F and the alignment with respect to Ψ in x is represented by:

$$\alpha(x) := \frac{|\langle x, \nabla\Psi^{F(x)}(x)\rangle|}{\|\nabla\Psi^{F(x)}(x)\|} \tag{1}$$

Connection of robustness with alignment was studied in [11] (cf.Thm 2). This states that a network's linearized robustness ($\hat{\rho}$) around an input x is upper-bounded by the binarized alignment term α^+ as:

$$\hat{\rho}(x) \leq \alpha^+(x) + \frac{C}{\|g\|} \tag{2}$$

where C is a constant, and linearized robustness $\hat{\rho}(x)$ is given by:

$$\hat{\rho}(x) := \min_{j \neq i^*} \frac{\Psi^{i^*}(x) - \Psi^j(x)}{\|\nabla\Psi^{i^*}(x) - \nabla\Psi^j(x)\|} \tag{3}$$

Also, g is the Jacobian of the top two logits i.e., $g = \nabla(\Psi^{i^*}(x) - \Psi^{j^*}(x))$ and binarized alignment i.e., α^+ is given by

$$\alpha^+(x) = \frac{|\langle x, \nabla(\Psi^{i^*} - \Psi^{j^*})(x)\rangle|}{\|\nabla(\Psi^{i^*} - \Psi^{j^*})(x)\|} \tag{4}$$

Here j^* is the minima of Eqn. 3. We also have $\alpha(x) = \alpha^+(x)$ for linear model and binary classifier. Eqn. 2 explains the deviation of different terms for linearized robustness in the case of a neural network. Also, a small error term in Eqn. 2 implies that robust networks yield better alignment, i.e., more interpretable saliency maps. The results are reported for the MNIST, F-MNIST, Flower, and GTSRB, for which the ground-truth attributions are not available. The "true" and "interpretable" attributions should highlight the distinctive features of the object, i.e., the parts that humans use to distinguish that object from the others. The reason for using an adversarially trained model as a reference is that as the adversarial robustness of a model increases, attributions become more aligned to the object and give out a more interpretable attribution that captures essential features [11]. Motivated by such findings and following the similar arguments as in [5], [22], we have also used an AT teacher. This explains why the saliency matching gives good object alignment and why adversarially trained networks are candidate choices for teacher networks.

Training with this methodology has improved both adversarial robustness, i.e., accuracy on adversarial examples, and attributional robustness, i.e., in terms of metrics used in literature like Top-K intersection, Kendall and Spearman correlations, and their corresponding attribution aware metrics, which we discuss in the coming sections.

Teacher-guided Attribution Enhancement : Let us say we have a pre-trained Teacher network (represented as f_T), and we have a student network represented by a neural network f_S, parameterized by θ. Given an input image x, we obtain the saliency map from a pre-trained Teacher Network, which is denoted as J_T^{TCI} (TCI represents True Class

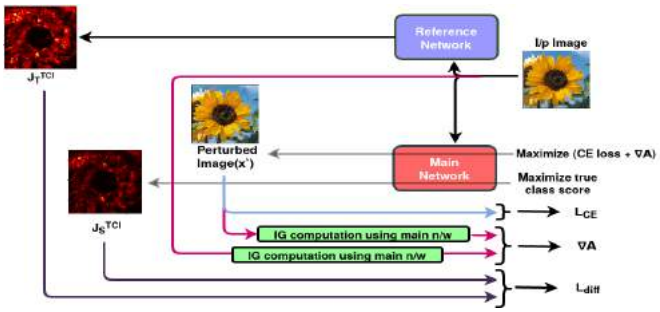

Figure 2. Our proposed saliency enhancer

Index). Now, we maximize the true class prediction score of the student network w.r.t input pixels and measure the net change in input pixels, which is represented as J_S^{TCI}. Now for an image of dimension $h \times w$ with c channels and $d = h \times w \times c$, J_T^{TCI} can be considered as per-pixel gradient and represented as:

$$J_T^{TCI}(x) = \nabla\Psi^{f_T(x)} = [\nabla\Psi^{f_T(x_1)} ... \nabla\Psi^{f_T(x_d)}] \quad (5)$$

Similarly, J_S^{TCI} is represented as:

$$J_S^{TCI}(x) = \nabla\Psi^{f_S(x)} = [\nabla\Psi^{f_S(x_1)} ... \nabla\Psi^{f_S(x_d)}] \quad (6)$$

Our sole purpose is to influence the model to generate better saliency that matches the teacher saliency. This is enforced by imposing similarity between the two saliency maps J_T^{TCI} and J_S^{TCI}. We add a loss term at outer minimization to minimize the l_2 distance between J_T^{TCI} and J_S^{TCI}, which is represented as $\mathcal{L}_{\text{diff}}$ and defined as below:

$$\mathcal{L}_{\text{diff}} = \|J_S^{TCI} - J_T^{TCI}\|_2^2 \quad (7)$$

Overall Optimization: It can be framed as a two-step process: (i) *Inner maximization* (ii) *Outer minimization*. The inner maximization is typically used to identify a suitable perturbation that achieves the objective of an attribution attack. On the other hand, the outer minimization seeks to use the above regularizer to counter the attack and match the saliency with the teacher. We describe each of them below.

Inner Maximization: In order to obtain the perturbed image $\mathbf{x}'$ through attributional attack, we use the following objective function:

$$\max_{\mathbf{x}' \in N(\mathbf{x},\epsilon)} \mathcal{L}_{CE}(\mathbf{x}', \mathbf{y}; \theta) + S(\nabla\tilde{\mathcal{A}}) \quad (8)$$
$$\text{where } \nabla\tilde{\mathcal{A}} = IG_{\mathbf{x}}^{\mathcal{L}_{CE}}(\mathbf{x}, \mathbf{x}')$$

Here, $\tilde{\mathcal{A}}$ denotes the computation of IG w.r.t the loss value, and our objective is to maximize loss. We use $\mathcal{L}_{CE}$ to denote the cross-entropy loss for the true class, and L_1-norm as $S(\cdot)$. Since the inner maximization is iterative by itself (and solved before the outer minimization), we randomly initialize each pixel of $\mathbf{x}'$ within an l_∞-norm ball of $\mathbf{x}$ and then iteratively maximize the objective function in Eqn. 8.

Outer Minimization: Our overall objective function for the outer minimization step is given by:

$$\min_\theta [\mathcal{L}_{CE}(\mathbf{x}', \mathbf{y}; \theta) + \underbrace{S(\nabla\tilde{\mathcal{A}})}_{\textit{robust attribution reg.}} + \lambda \underbrace{\mathcal{L}_{\text{diff}}}_{\textit{saliency matching reg.}}] \quad (9)$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss used for the multi-class classification setting. The term $\nabla\tilde{\mathcal{A}}$ represents difference in IG terms w.r.t the input image $\mathbf{x}$ and the perturbed image $\mathbf{x}'$ i.e., $IG_i^{\mathcal{L}_{CE}}(\mathbf{x}_0, \mathbf{x}) - IG_i^{\mathcal{L}_{CE}}(\mathbf{x}_0, \mathbf{x}')$ considering the CE loss for IG computations. We use λ as a weighting coefficient for $\mathcal{L}_{\text{diff}}$. We show the effects of considering different λ values on our proposed method in ablation studies. Refer to Fig 2 for better understanding.

Attribution-Aware Attributional Robustness Metrics : The metrics used in attributional robustness literature like the Top-K intersection and Kendall and Spearman's correlations quantify the similarity between attribution maps before and after the attacks. However, they do not take into account the goodness of the attribution maps. Such a regularizer restricts the change in attribution maps with imperceptible changes in the input image. In effect, a robust attribution regularizer drives the derivative of the attribution map towards zero. Moreover, it is a well-known fact that the attribution methods signify the gradient of the target class concerning the input image. It, in turn, implies that robust attribution regularizer pushes the 2nd order derivative towards zero. In Fig 1, we show that attributionally robust models suffer from object-alignment issues and generate very sparse and blackish attributions. This observation signifies that the network eventually settles in a degenerate state where 1st and 2nd order derivatives are close to zero. Hence we have ill-attributed but robust attribution maps which may not be trustworthy enough to be deployed in mission-critical applications like navigation, medical diagnosis, etc. Keeping in mind these shortcomings of prior metrics, we design new metrics that consider both the quality and stability of attribution maps. We call these metrics as Attribution-Aware (AA) metrics and are defined as follows:
Attribution Aware *Attributional Robustness* **= (*Goodness*** of attribution map) $\times$ **(*Similarity*** of attribution maps for original and attacked images).

We note that the essence of the current evaluation metrics are maintained and the new metrics, namely ***AA-Top-K, AA-Kendall*** and ***AA-Spearman*** are defined respectively as the following:

$$\underbrace{[TopK(A) \cap TopK(B)]}_{\textit{Goodness of attr. map}} \times \underbrace{[TopK(B) \cap TopK(C)]}_{\textit{Similarity of attr. map}} \quad (10)$$

$$\underbrace{[Kendall\rho(A, B)]}_{\textit{Goodness of attr. map}} \times \underbrace{[Kendall\rho(B, C)]}_{\textit{Similarity of attr. map}} \quad (11)$$

$$\underbrace{[Spearman\rho(A, B)]}_{\textit{Goodness of attr. map}} \times \underbrace{[Spearman\rho(B, C)]}_{\textit{Similarity of attr. map}} \quad (12)$$

Here, Kendall corr. and Spearman corr. are denoted as $Kendall\rho$ and $Spearman\rho$ respectively. B and C represent attribution maps before and after the attack. A represents a "true" and "interpretable" attribution map. Inspired

Datasets	Methods	Clean	Adv. Acc.	Top-K	Kendall	AA-Top-K	AA-Kendall
F-MNIST	Nat	**90.86**	0.01	39.01	0.4610	5.41	0.1946
	AT [18]	85.73	73.01	46.12	0.6251	46.12	0.6251
	AT-start + **Regularizer**	86.21	**76.52**	72.33	0.6624	54.96	0.6354
	RAR [7]	85.44	70.26	72.08	0.6747	51.48	0.5754
	RAR-start + **Regularizer**	86.81	73.24	73.49	0.6920	55.39	0.6145
	ERAR [21]	85.45	71.61	81.50	0.7216	59.21	0.6154
	ERAR-start + **Regularizer**	87.01	74.16	**82.31**	**0.7368**	**62.34**	**0.65**
MNIST	Nat	**99.17**	0.00	46.61	0.1758	4.12	0.0021
	AT [18]	98.40	92.47	62.56	0.2422	62.56	0.2422
	AT-start + **Regularizer**	98.20	**93.5**	71.84	0.3465	64.78	0.2652
	RAR [7]	98.34	88.17	72.45	0.3111	58.42	0.2851
	RAR+**Regularizer**	98.62	90.79	73.48	0.3317	61.12	0.2961
	ERAR [21]	98.41	89.53	81.00	0.3494	66.45	0.2821
	ERAR+**Regularizer**	98.72	92.66	**82.89**	**0.3625**	**73.28**	**0.3019**

Table 1. Comparative results of clean, adversarial, and attributional robustness achieved by models following our training method (as in eq. 8 and eq. 9) and another baseline attributional robustness training methods. Here, by "**Regularizer**" we mean *robust attribution regularizer + saliency alignment regularizer* (as in eq. 9). Each gray row indicates a model is trained following our proposed joint training strategy. Among the colored rows, the difference is the starting point of the model, i.e., e.g., "AT-start" indicates that we initialize the training starting from an adversarially trained model. The Cyan row indicates the best result achieved on the corresponding dataset.

by the findings of [11] and the reasons briefly discussed in section 2, we choose an attribution map of an adversarially trained model as a potential candidate for A. We note that any model capable of generating interpretable or object-aligned attribution maps could also be a good choice for A. We notice the attributional robustness of a model trained using the baseline attributionally robust methods, such as RAR [7] and ERAR [21] drops under these Attribution Aware evaluation schemes. Owing to space constraints, we have moved results on Flower and GTSRB datasets and ablation studies to the supplementary section. Table 1 also show a similar trend. We also observe that a model's attributional robustness when trained using our method improves significantly under these schemes. Such findings indicate the efficacy of a joint training strategy for providing a robust model with stable and object-aligned explanation maps.

APPENDIX: Empowering a Robust Model with Stable and Interpretable Explanations

3. Related Work

Attributional Robustness: Despite being an important consideration for explainable AI, attributional robustness has mostly been overlooked by researchers in the community. Chen *et al*. [7] first proposed a training methodology to improve the attributional robustness of a model following the adversarial robustness framework proposed by Madry *et al*. [18]. This consists of an iterative inner maximization and an outer minimization step. At the inner maximization step, an input image is perturbed iteratively, which changes the model's attribution map maximally. In contrast, at the outer minimization step, the model is trained with the objective of minimizing the change in the attribution map due to an indistinguishable change in the image. Sarkar *et al*. [21] further improved the attributional robustness of a model by proposing a training strategy based on two regularizers. The first one, i.e., *class attribution based contrastive regularizer* forces the true class attribution to assume a skewed shape distribution and the negative class attribution to behave uniformly. Another regularize, i.e., *weighted attribution based regularizer* was introduced to weight the change in the attribution of each pixel due to an indistinguishable change in the image. Following a similar training framework, Singh et al. [23], proposed a robust attribution training method, which effectively tries to maximize the cosine similarity between the saliency map of true class and the actual input image. At the same time, this training method minimizes the cosine similarity between the negative class saliency map and the actual image. This method fails specifically for images where the ground truth class contains darker object parts compared to the rest of the image or if there exist bright pixels anywhere in the image outside the object of interest. Wang *et al*. [28] proposed smooth surface regularization to minimize the difference between saliency maps for nearby points and showed that the model trained with this regularizer helps improve attributional robustness compared to the model trained by adversarial training [18]. However, there is another line of work that deals with the sanity checks of explanation maps like [1], [4], [14], though they do not use terms like attributional robustness, which have similar goals.

Adversarial Robustness: Unlike attributional robustness, adversarial robustness is a well-explored research area. Starting from [12, 26], there are numerous efforts that show the vulnerability of neural networks against carefully crafted human-imperceptible perturbations in an image. Goodfellow *et al*. [13] introduced Fast Gradient Sign Method (FGSM) attack method which was followed by more effective iterative adversarial attacks such as proposed by Kurakin *et al*. [16], C&W [3] attack, PGD [18], momentum iterative attack [10], diverse input iterative attack [32]. On the other hand, a parallel line of work also became very popular such as [2,8,15,18,22,27,31,34], which aim to find training strategy to defend against stronger adversarial attacks. Mis-classification Aware Adversarial Training [27], Geometry Aware Adversarial Training [35], TRADES [34], Feature Denoising Training [31], Adversarial Logit Pairing [15], Parseval's Network [8], Curriculum Adversarial Training [2], etc. Also, there exist saliency-based methods to improve adversarial robustness, such as Jacobian Adversarially Regularized Network (JARN) [6] which improves model robustness by matching the input gradient w.r.t. loss to the actual image. Chen *et al*. [5] proposed a method that leverages a discriminator to compare the jacobian of the image and the image saliency. This work [7] is very similar to JARN [6] as JARN compares the image to the transformed version of the Jacobian through an adaptive network. Our work is based on improving existing attributional robustness training methods by attribution alignment through a teacher network that is shown to improve the adversarial robustness of the model.

4. Preliminaries

With the purpose of enforcing the restriction on the saliency maps, produced under attributional robustness training, to match to a *good* saliency, we propose to include alignment of saliency maps in the training method. With an adversarially trained teacher, the alignment is analogous to the amalgamation of adversarial features to the model, leading to improvement of performance against adversarial attacks. Hence, our proposed training method requires evaluation against two types of attacks, i.e., adversarial as well as attributional attacks. We introduce each of them below.

Adversarial Attack: The goal of an adversarial attack is to find out the minimum perturbation δ in the input space of $\mathbf{x}$ (i.e., input pixels for an image) that results in a maximal change in classifier(f)'s output. In this work, to test the adversarial robustness of a model, we use one of the strongest adversarial attacks, Projected Gradient Descent (PGD) [18], which is considered a benchmark for adversarial accuracy in other recent state-of-the-art attributional robustness methods [7, 21, 23]. PGD is an iterative variant of the Fast Gradient Sign Method (FGSM) [13]. PGD adversarial examples are constructed by iteratively applying FGSM and projecting the perturbed output to a valid constrained space S. PGD attack is formulated as follows:

$$\mathbf{x}^{i+1} = Proj_{\mathbf{x}+S}\left(\mathbf{x}^{i} + \alpha(\nabla_{\mathbf{x}}\mathcal{L}(\theta, \mathbf{x}^{i}, \mathbf{y}))\right) \quad (13)$$

Here, θ denotes the classifier parameters; input and output are represented as $\mathbf{x}$ and $\mathbf{y}$ respectively; and the classification loss function as $\mathcal{L}(\theta, \mathbf{x}, \mathbf{y})$. Usually, the magnitude of adversarial perturbation is constrained in a L_p-norm ball ($p \in \{0, 2, \infty\}$) to ensure that the adversarially perturbed

example is perceptually similar to the original sample. Note that $\mathbf{x}^{i+1}$ denotes the perturbed sample at $(i+1)^{th}$ iteration.

Naturally trained models are fooled by such images generated by adversarial attacks, which are human-imperceptible [13]. This gives rise to a different training method that focuses on persisting in the model decisions for both original and attacked images. Consider an image classifier $f(x;\theta) : x \longrightarrow \mathbb{R}^c$ with parameters θ, which maps input image x to a c-dimensional output. The network f is called adversarially robust if for an attacked image $\hat{x}$, we have:

$$\underset{i\in c}{\operatorname{argmax}} f_i(x;\theta) = \underset{i\in c}{\operatorname{argmax}} f_i(\hat{x};\theta) \tag{14}$$

Attributional Attack: The goal of an attributional attack is to devise visually imperceptible perturbations that change the attribution map of the test input maximally while preserving the predicted label. To test the attributional robustness of a model, we use the Iterative Feature Importance Attack (IFIA) in this work. As [12] convincingly demonstrated, IFIA helps generate minimal perturbations that substantially change model interpretations while keeping their predictions intact. The IFIA method is formally defined as below:

$$\begin{gathered}\arg\max_{\delta}\ D(I(\mathbf{x};f), I(\mathbf{x}+\delta;f)) \\ \text{subject to: } ||\delta||_\infty \leq \epsilon \\ \text{such that: } \operatorname{argmax} f(\mathbf{x};\theta) = \operatorname{argmax} f(\mathbf{x}+\delta;\theta)\end{gathered} \tag{15}$$

Here, $I(\mathbf{x}, f)$ is a vector of attribution scores over all input pixels when an input image $\mathbf{x}$ is presented to a classifier network f parameterized by θ. $D(I(\mathbf{x};f), I(\mathbf{x}+\delta;f))$ measures the dissimilarity between attribution vectors $I(\mathbf{x};f)$ and $I(\mathbf{x}+\delta;f)$. In our work, we choose D as Kendall's correlation computed on top-k pixels as in [12].

Similar to adversarial robustness, a model has to train differently to acquire attributional robustness. RAR [7] proposed a method that consists of inner maximization and outer minimization framework as shown below:

$$\min_{\theta}[\max_{\mathbf{x}'\in N(\mathbf{x},\epsilon)}\{l_{CE}(\mathbf{x}',\mathbf{y};\theta) + S(\nabla\tilde{\mathcal{A}})\}] \tag{16}$$

where, l_{CE} is standard cross-entropy loss, $\tilde{\mathcal{A}}$ denotes comparison of saliency maps w.r.t. some attribution method and L_1-norm is used as $S(\cdot)$.

Specifically RAR [7] used Integrated Gradient (IG) method [25] which was followed by other attributionally robustness training methods such as [21, 23]. It functions as a technique to provide axiomatic attribution to different input features proportional to their influence on the output. IG obeys axioms of attribution (IG is candidate attribution, among others) and can be extended to any other combination without significant modification. So we adopted IG. Moreover, all the prior works use PGD and IG combination; hence, for a fair comparison, we followed the same combination. Computation of IG is mathematically approximated by constructing a sequence of images interpolating from a baseline to the actual image and then averaging the gradients of neural network output across these images, as shown below:

$$IG_i^f(\mathbf{x}_0,\mathbf{x}) = (\mathbf{x}^i - \mathbf{x}_0^i) \times \sum_{k=1}^{m} \frac{\partial f(\mathbf{x}_0^i + \frac{k}{m}\times(\mathbf{x}^i - \mathbf{x}_0^i))}{\partial \mathbf{x}^i} \times \frac{1}{m} \tag{17}$$

Here $f : \mathbb{R}^n \rightarrow \mathcal{C}$ represents a deep network with $\mathcal{C}$ as the set of class labels, $\mathbf{x}_0$ is a baseline image with all black pixels (zero intensity value), and i is the pixel location on input image $\mathbf{x}$ for which IG is being computed.

The term $\nabla\tilde{\mathcal{A}}$ in Eqn. 16 is represented by $IG_i^f(\mathbf{x},\mathbf{x}')$. This is similar to the difference in IG terms w.r.t the input image x and perturbed image x' i.e. $IG_i^f(\mathbf{x}_0,\mathbf{x}) - IG_i^f(\mathbf{x}_0,\mathbf{x}')$, where $\mathbf{x}_0$ is the baseline image for IG computation.

5. Experiments and Results

We conduct a comprehensive suite of experiments to show the effectiveness of our proposed training strategy for improving the adversarial robustness of a model when compared with a model trained with any baseline attributional robustness training technique. Our training strategy improves the quality of attribution maps in previous and proposed metrics and improves clean and adversarial accuracies. We report our results on four benchmark datasets viz., MNIST [17], Fashion-MNIST [30], Flower [19], and GTSRB [24]. Note that the attack methods used for training and evaluations are **not the same**. For training, we attacked the IG-regularizer using an iterative approach similar to the PGD. For evaluations, we used IFIA, which attacks the Top-K salient features in the attribution map.

Architecture Details: We follow similar architectures used in RAR and ERAR for our experiments for a fair comparison. For Fashion-MNIST and MNIST datasets, we use a network comprising two CNN layers with 32 and 64 filters, each followed by 2×2 max-pooling and a fully connected layer with 1024 neurons. For the others, we use a Resnet model consisting of 5 residual units, each with (16,16,32,64) filters. We compare our method with [23] using WRN 28-10 [33] architecture. For the teacher network, we use the same architectures of respective datasets and trained using the [18]'s framework

Results: Table 2 reports comparisons of natural, adversarial accuracies along with attributional robustness performance (top-k intersection, Kendall's and Spearman correlation) with other baselines of attributionally and adversarially robust models on Flower and GTSRB datasets. Our experimental finding suggests including a saliency alignment regularizer in a baseline attributional robustness train-

ing framework, such as RAR [7] or ERAR [21], not only improves adversarial accuracy significantly but it also helps the model retain its attributional robustness. Thus, our proposed training strategy allows an attributionally robust model to attain both-way robustness. Similarly, a saliency alignment regularizer in a baseline adversarial robustness training framework, such as AT [18] allows the model to retain its adversarial robustness and a robust attribution regularizer boosts the model's attributional robustness substantially. Such findings also justify the joint training strategy's potency in achieving a robust model against adversarial and attributional attacks. Our proposed training strategy also significantly improves attributional robustness under our proposed attributional robustness evaluation setup. We present comparative results with [23] following the similar architectural details and different attack configurations they used for all the results.

6. Ablation Studies

Qualitative Analysis of Object-alignment in Attribution Maps. In Fig 3, we show comparative visualizations of attribution maps generated from our model and a baseline attributionally robust model [7] with sample test images from MNIST, F-MNIST, GTSRB, and Flower datasets. These results show that our model improves attributional robustness and produces more interpretable saliency maps compared to the baseline attributionally robust model.

Quantitative Analysis of Object-alignment in Attribution Maps. We compare attribution maps generated by our model as well as from another baseline attributionally robust models [7, 21] through a quantitative measure. In order to evaluate the quality of the attribution map quantitatively, we measure the similarity between attribution maps generated by a reference model and the main model under consideration using Top-k intersection, Kendall's, and Spearman correlation. Note that any model capable of generating an interpretable attribution map can be considered the reference network for evaluation. We provide results with the adversarially trained model as a reference network in table [3] which shows that attribution maps generated from our model are of better quality and more interpretable compared to the attribution maps generated from the baseline attributionally robust models.

Generalization of proposed metrics and training strategy.

Since any adversarially robust model relies on robust input features, such models can produce object-aligned attribution maps that capture all crucial features. Hence, one could evaluate our method with other adversarial training strategies like TRADES. We show our results with TRADES as the reference in Table 4 and show that our training strategy shows superior performance here too. Apart from using a different adversarially trained reference network, we also studied the alignment of the attribution with the object itself, i.e., we take the input itself as reference. We can study this for datasets like MNIST and F-MNIST by segmenting the images using simple thresholding. We report these results in Table 5, supporting our claim.

Effect of λ (Coefficient of Saliency Alignment Regularizer). We conduct experiments to understand the effect of different λ values, i.e., regularizer coefficient of saliency alignment regularizer in Eqn 9 on our training method. λ values are selected through a binary search between 0.01 and 0, with the best value in the vicinity of 0.001 for all datasets. Our experimental findings indicate that a model's adversarial and attributional robustness improves with increasing values of λ up to a certain point and then drops gradually. Fig 5 shows the trend of a model's natural, adversarial, and attributional robustness performance when trained with different values of λ. We notice similar trends for all datasets considered for our experiments.

Qualitative Analysis of Stability of Attribution Map. In Fig 4, we depict relative visualizations of attribution maps generated using a model trained using our proposed joint training strategy and using an adversarially trained model [18]. It is highly indicative from these visualizations that our proposed coordinated training strategy boosts the robustness of attribution maps against attributional attacks significantly without penalizing the object alignment or the quality of the attribution maps. Such visualizations also align with our experimental findings in tables 1 and 2, as the attributional robustness attained by the AT-start**+Regularizer** model is significantly higher than the AT model.

7. Studying Object-Alignment through Segmentation Masks

The Flower dataset is provided with segmentation masks, and we used that to see how well the attribution maps are localized and aligned to the object. We provide quantitative and qualitative analysis to show how our training strategy helps object alignment. For quantitative analysis, we used the segmentation map provided with the dataset to isolate the object pixels in the attribution map and compared it with the original attribution map. We present the results of our quantitative analysis in Tab. 6. We also give the qualitative results in Fig. 6, 7, 8, 9. From these results, it can be inferred that our training strategy produces well-aligned and localized attribution maps.

Datasets	Methods	Clean	Adv. Acc.	Top-K	Kendall	AA-Top-K	AA-Kendall
Flower	Nat	**86.76**	0.00	8.12	0.4978	3.9	0.071
	AT [18]	83.82	41.91	55.87	0.7784	55.87	0.7784
	AT-start + **Regularizer**	83.36	45.72	65.24	0.7958	57.71	0.7647
	RAR [7]	82.35	47.06	66.33	0.7974	33.67	**0.8124**
	RAR-start + **Regularizer**	83.47	48.35	67.29	0.8091	59.64	0.7521
	ERAR [21]	83.09	51.47	69.50	0.8121	57.21	0.7314
	ERAR-start + **Regularizer**	84.06	**52.18**	**71.07**	**0.8356**	**62.45**	0.8041
GTSRB	Nat	**98.57**	21.05	54.16	0.6790	4.21	0.041
	AT [18]	97.59	83.24	68.85	0.7520	62.56	0.2422
	AT-start + **Regularizer**	97.92	**98.17**	74.43	0.7633	65.89	0.2718
	RAR [7]	95.68	77.12	74.04	0.7684	43.0	0.2631
	RAR-start + **Regularizer**	96.81	80.07	75.39	0.7843	63.16	0.3012
	ERAR [21]	97.61	82.33	77.15	0.7889	67.41	0.3014
	ERAR-start + **Regularizer**	98.23	84.86	**78.64**	**0.8017**	**73.45**	**0.3541**

Table 2. Comparative results of clean, adversarial, and attributional robustness achieved by models following our training method (as in eq. 8 and eq. 9) and another baseline attributional robustness training methods. Here, by "**Regularizer**" we mean *robust attribution regularizer + saliency alignment regularizer* (as in eq. 9). Each gray row indicates a model is trained following our proposed joint training strategy. Among the colored rows, the difference is the model's starting point, i.e., e.g., "AT-start" indicates that we initialize the training starting from an adversarially trained model. The Cyan row indicates the best result achieved on the corresponding dataset.

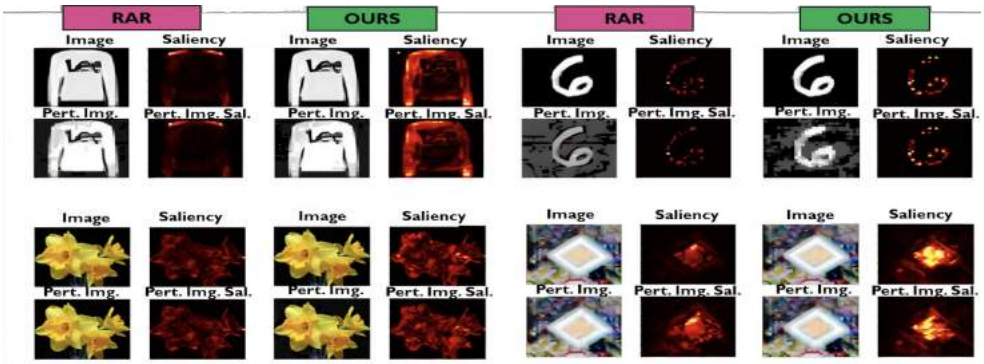

Figure 3. Comparative visualizations of original and human-imperceptible attributionally attacked images (same predicted class) and corresponding attribution maps obtained using RAR [7] and OUR model (RAR-start**+Regularizer**) for test samples from FMNIST, MNIST, Flower & GTSRB datasets. The model trained using RAR [7] generates similar attribution maps before and after the attack, but neither captures the salient part of objects effectively. Our model not only generates similar attribution maps before and after an attack but also produces better quality attribution maps compared to RAR. These visualizations illustrate the importance of a saliency alignment regularizer for improving the quality of attribution maps of attributionally robust models.

Datasets	Methods	Top-K	Kendall	Spearman
F-MNIST	AT-start + **Regularizer**	0.8134	0.8892	0.9733
	RAR [7]	0.741	0.8626	0.9619
	RAR-start +**Regularizer**	0.7941	0.9145	0.9741
	ERAR [21]	0.8124	0.9452	0.9745
	ERAR-start+**Regularizer**	**0.8214**	**0.9752**	**0.9847**
MNIST	AT-start + **Regularizer**	**0.9352**	**0.9764**	**0.9989**
	RAR [7]	0.8232	0.9336	0.9922
	RAR-start + **Regularizer**	0.8283	0.9338	0.9923
	ERAR [21]	0.8124	0.8974	0.9145
	ERAR-start + **Regularizer**	0.8541	0.9451	0.9974
Flower	AT-start + **Regularizer**	**0.6354**	**0.8695**	**0.9774**
	RAR [7]	0.565	0.7374	0.903
	RAR-start + **Regularizer**	0.5785	0.7587	0.927
	ERAR [21]	0.5012	0.6945	0.8974
	ERAR-start + **Regularizer**	0.5978	0.7841	0.9452
GTSRB	AT-start + **Regularizer**	**0.6381**	0.6154	0.7483
	RAR [7]	0.5687	**0.7568**	0.7558
	RAR-start + **Regularizer**	0.5854	0.5804	0.764
	ERAR [21]	0.5541	0.7398	0.7427
	ERAR-start + **Regularizer**	0.6174	0.7541	**0.7793**

Table 3. Comparative results of *goodness* of attribution map obtained using OUR model and baseline attributionally robust models such as - RAR [7] and ERAR [21], and baseline adversarial robust models like AT [18]. The results clearly show the effectiveness of a saliency-matching regularizer for improving the quality of attribution maps.

Dataset	Model	AA-Top-K	AA-Kendall
MNIST	AT [18]+**Regularizer**	**62.84**	0.2763
	RAR [7]+**Regularizer**	61.69	**0.3024**
	RAR [7]	56.48	0.2651
	AT [18]	60.76	0.2214
F-MNIST	AT [18]+**Regularizer**	**55.21**	0.6139
	RAR [7]+**Regularizer**	54.74	0.5631
	RAR [7]	50.38	0.5321
	AT [18]	45.23	**0.6241**

Table 4. TRADES as reference

Dataset	Model	AA-Top-K	AA-Kendall
MNIST	AT [18]+**Regularizer**	**61.47**	0.2458
	RAR [7]+**Regularizer**	59.65	**0.2894**
	RAR [7]	54.89	0.2517
	AT [18]	58.69	0.2178
F-MNIST	AT [18]+**Regularizer**	**50.47**	0.5148
	RAR [7]+**Regularizer**	48.57	0.4747
	RAR [7]	40.25	0.4152
	AT [18]	42.85	**0.4875**

Table 5. Object as reference

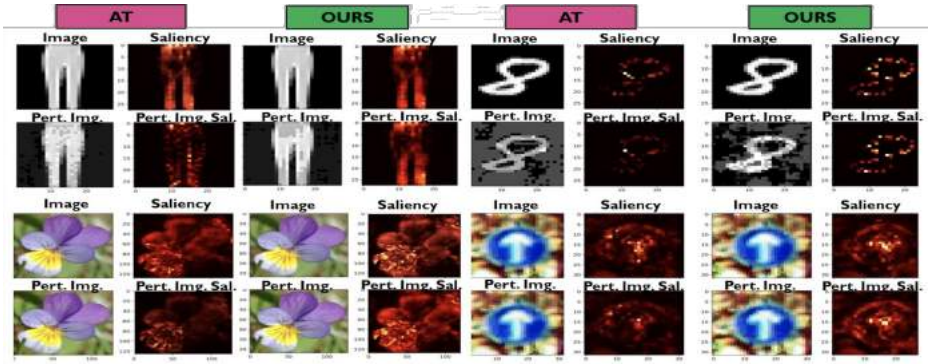

Figure 4. Comparative visualizations of original and human-imperceptible attributionally attacked images (same predicted class) and their corresponding attribution maps obtained using AT [18] and OUR model (AT-start+**Regularizer**) for test samples from FMNIST, MNIST, Flower & GTSRB datasets. The model, trained using AT [18], generates dissimilar attribution maps before and after an attack and is not robust/stable against an attributional attack. OUR model can generate similar attribution maps before and after an attack without harming object alignment or quality of the attribution maps. These visualizations illustrate the importance of a joint training strategy for empowering adversarially robust models with stable explanations.

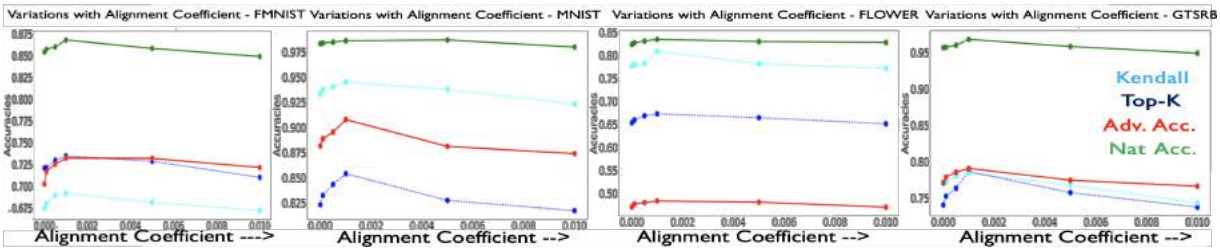

Figure 5. Effect of λ(i.e., regularizer coefficient of alignment loss) on different datasets' natural, adversarial, and attributional accuracies.

Model	Top-K	Kendall	Spearmann
RAR [7]	98.45	0.9424	0.9752
Adv-Trained [18]	98.96	0.9574	0.9635
RAR-start	100.0	0.9852	0.9952
Adv-start	100.0	0.9985	0.9995

Table 6. Flower Dataset Segmentation

8. Comparative Results with Singh *et al.* [23]

We provide a detailed comparison in the main paper, which shows improvement in the adversarial robustness of a model when compared with a model trained with any baseline attributional robustness methods [7, 21]. We also report improvement in clean accuracy and attributional robustness performance with our training method. A similar trend is visible when we consider another baseline attributional robustness technique proposed by Singh et al. [23]. As this work used WRN28-10 [33], which is different from the network used by [7, 21], we separately add comparative results with [23] here using a similar architecture in Tab.7 for fair comparison. These experimental findings support the potency of our proposed joint training strategy for improving the stability and object-alignment of attribution maps of a robust model, such as [23].

9. Quantitative Analysis of Attribution Map with Naturally Trained Teacher

To show that our method works for any teacher, we quantitatively evaluate the performance of a model trained using a naturally trained teacher. The values corresponding to the metrics show that our joint training strategy works with any teacher. However, the quality and stability of attribution maps and the model's performance depend on the teacher's quality of attribution maps. One can validate this observation by analyzing the trends in Table 8. With a naturally trained teacher, the stability of the attribution map is worse than the attributionally trained models like [7, 21]. Though the saliency alignment regularizer improves the proposed attribution-aware metrics, the values are far below compared to the state-of-the-art numbers. This discrepancy occurs due to the conflicting goals of robust attribution regularizer and saliency alignment regularizer with a naturally trained teacher. The robust attribution regularizer tries to maintain sparsity while the saliency alignment regularizer pushes it toward a more distributed naturally trained model's attribution. Hence, our model works with any teacher, but it is crucial to consider a teacher that produces reasonably accurate attribution maps, like an adversarially trained model.

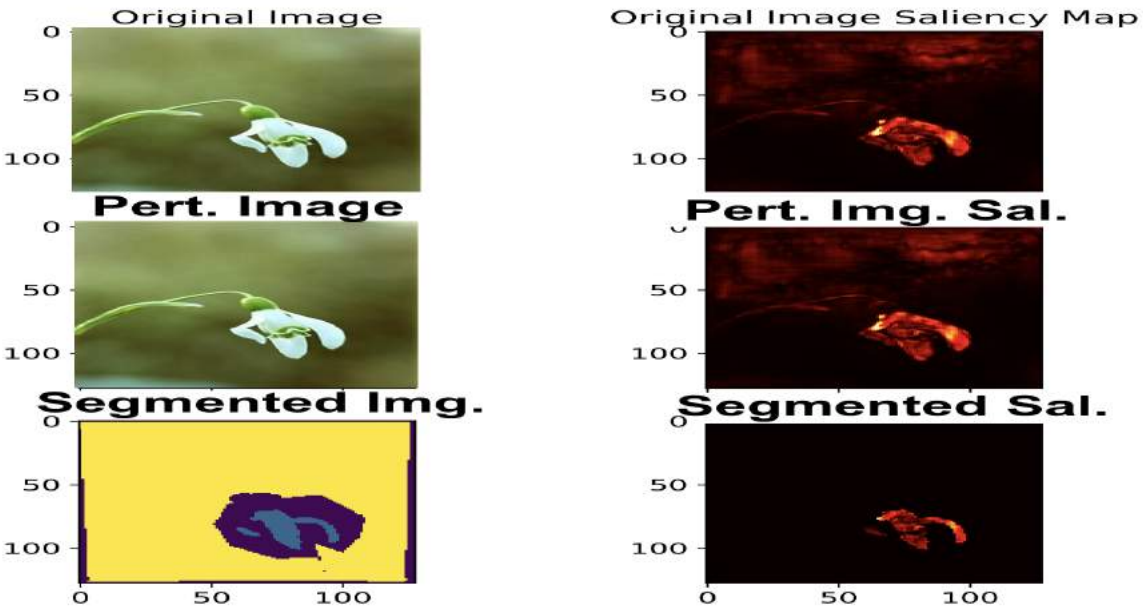

Figure 6. Saliency Comparison with ground truth Segmentation map

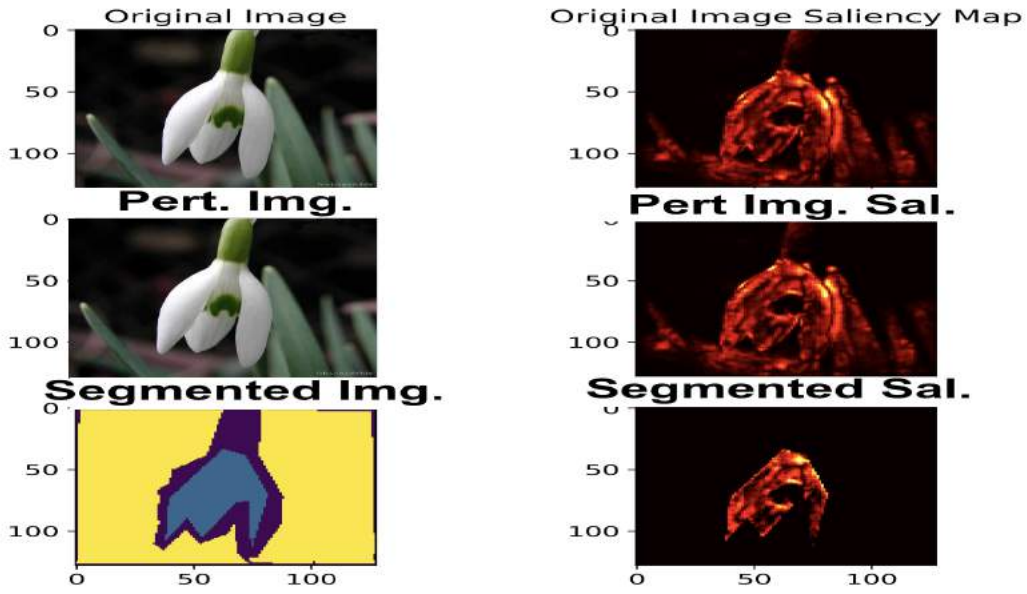

Figure 7. Additional Visualizations: Saliency comparison with ground truth segmentation map

10. Hyperparameters and Attack Configurations

Herein, we present details of training hyperparameters as well as attack configuration for all our experiments. For attributional attack, we use Iterative Feature Importance Attacks (IFIA) proposed by [12] (specific settings for each dataset described below). We set the feature importance function as Integrated Gradients (IG) and dissimilarity function D as Kendall's rank order correlation across all datasets. Also, we kept adversarial and attributional attack configurations fixed while comparing the result with other baseline methods for fairness.

10.1. Flower Dataset:

Training Hyperparameters: We use a momentum optimizer with weight decay, momentum rate 0.9, weight decay rate 0.0002, batch size 16, and training steps 90,000. We use

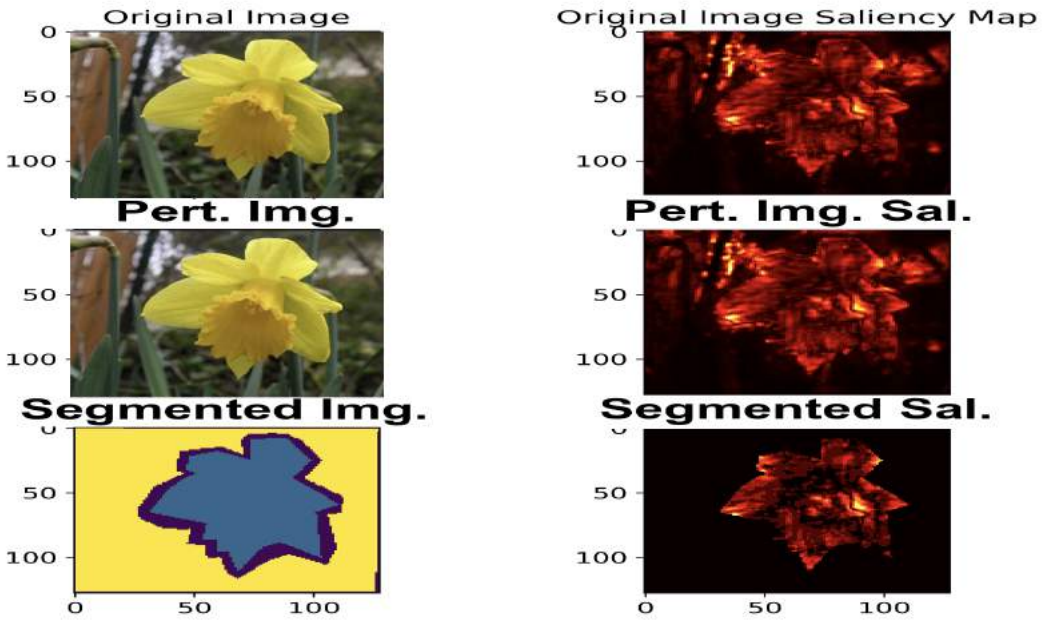

Figure 8. Additional Visualizations: Saliency comparison with ground truth segmentation map

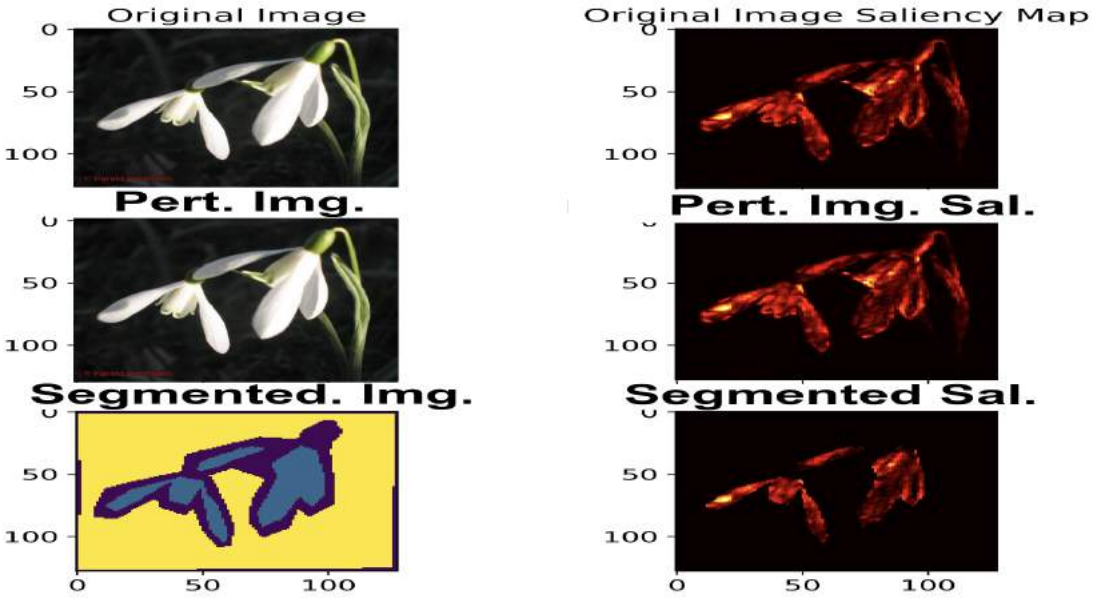

Figure 9. Additional Visualizations: Saliency comparison with ground truth segmentation map

a learning rate schedule as follows: the first 1500 steps have a learning rate of 10^{-4}; after 1500 steps and until 70,000 steps have a learning rate of 10^{-3}; after 70,000 steps have a learning rate of 10^{-4}. We use a PGD attack as an adversary with a random start, the number of steps of 7, a step size of 2, $m = 5$ as the number of steps for approximating IG computation in the attack step, and adversarial budget ϵ of 8.

Attack Configuration for Evaluation: For evaluating adversarial robustness, we use a PGD attack with the number of steps of 40, adversarial budget ϵ of 8, and step size of 2. For attributional attack, we use IFIA's top-k attack with $k = 1000$, adversarial budget $\epsilon = 8$, step size $\alpha = 1$ and number of iterations $P = 100$.

10.2. Fashion-MNIST Dataset:

Training Hyperparameters: We use the learning rate as 10^{-4}, batch size as 32, training steps as 100,000, and

Datasets	Methods	Clean	Adv. Acc.	Top-K	Kendall	AA-Top-K	AA-Kendall
Flower	Nat	93.91	0.00	38.22	0.5643	7.2	0.3124
	AT [18]	92.64	69.85	80.84	0.8414	80.84	0.8414
	[23]	93.21	33.08	79.84	0.8487	60.24	0.6295
	[23]+**Alignment**	**93.98**	**71.64**	**86.95**	**0.9271**	**82.95**	**0.8654**
GTSRB	Nat	**99.43**	19.9	68.74	0.7648	5.3	0.12
	AT [18]	98.36	87.49	86.13	0.8842	86.13	0.8842
	[23]	98.47	84.66	91.96	0.8934	83.37	0.8124
	[23]+**Alignment**	99.18	**88.79**	**93.56**	**0.9124**	**87.65**	**0.8541**

Table 7. Comparative results of clean, adversarial, and attributional robustness achieved by a model obtained using our training method and a baseline attributional robustness training method [23] using WRN28-10 Architecture.

Datasets	Methods	Clean	Adv. Acc.	Top-K	Kendall	AA-Top-K	AA-Kendall
F-MNIST	Nat	**90.86**	0.01	39.01	0.4610	39.01	0.4610
	AT [18]	85.73	**73.01**	46.12	0.6251	5.41	0.1946
	RAR [7]	85.44	70.26	72.08	0.6747	45.63	0.4215
	RAR+**Alignment**	84.81	50.29	64.49	0.6837	**51.21**	**0.7185**
	ERAR [21]	85.45	71.61	81.50	0.7216	34.56	0.4154
	ERAR+**Alignment**	85.01	71.16	**82.31**	**0.7368**	45.84	0.5741
MNIST	Nat	**99.17**	0.00	46.61	0.1758	46.61	0.1758
	AT [18]	98.40	**92.47**	62.56	0.2422	4.12	0.0021
	RAR [7]	98.34	88.17	72.45	0.3111	47.69	**0.5741**
	RAR+**Alignment**	88.62	80.79	67.78	0.3317	61.52	0.2761
	ERAR [21]	98.41	89.53	**81.00**	**0.3494**	45.45	0.2121
	ERAR+**Alignment**	85.72	77.66	63.89	0.2625	**65.46**	0.2459
Flower	Nat	**86.76**	0.00	8.12	0.4978	8.12	0.4978
	AT [18]	83.82	41.91	55.87	0.7784	3.9	0.071
	RAR [7]	82.35	47.06	66.33	0.7974	27.67	0.7224
	RAR+**Alignment**	79.47	42.82	65.23	0.7691	57.64	0.6921
	ERAR [21]	83.09	**51.47**	**69.50**	**0.8121**	53.21	0.6614
	ERAR+**Alignment**	81.06	49.18	65.07	0.7856	**59.45**	**0.7541**
GTSRB	Nat	98.57	21.05	54.16	0.6790	54.16	**0.6790**
	AT [18]	97.59	**83.24**	68.85	0.7520	4.21	0.041
	RAR [7]	95.68	77.12	74.04	0.7684	37.0	0.2471
	RAR+**Alignment**	94.26	75.71	69.39	0.7249	59.16	0.6212
	ERAR [21]	**97.61**	82.33	**77.15**	**0.7889**	51.53	0.6317
	ERAR+**Alignment**	95.51	79.75	73.54	0.7541	**68.27**	0.6321

Table 8. Comparative results of clean, adversarial, and attributional robustness achieved by a model obtained using our training method (trained with the naturally trained teacher) and another baseline attributional robustness training methods [7] [21].

Adam optimizer. We use PGD attack as the adversary with a random start, the number of steps of 20, step size of 0.01, $m = 10$ as the number of steps for approximating IG computation in the attack step, and adversarial budget $\epsilon = 0.1$.

Attack Configuration for Evaluation: For evaluating adversarial robustness, we use a PGD attack with a random start, number of steps of 100, adversarial budget ϵ of 0.1, and step size of 0.01. For attributional attack, we use IFIA's top-k attack with $k = 100$, adversarial budget $\epsilon = 0.1$, step size $\alpha = 0.01$ and number of iterations $P = 100$.

10.3. MNIST Dataset:

Training Hyperparameters: We use the learning rate as 10^{-4}, batch size as 50, training steps as 90,000, and Adam optimizer. We use a PGD attack as the adversary with a random start, the number of steps of 40, step size of 0.01, $m = 10$ as the number of steps for approximating IG computation in the attack step, and adversarial budget $\epsilon = 0.3$.

Evaluation Attacks Configuration: For evaluating adversarial robustness, we use a PGD attack with a random start, number of steps of 100, adversarial budget ϵ of 0.3, and step size of 0.01. For attributional attack, we use IFIA's top-k attack with $k = 200$, adversarial budget $\epsilon = 0.3$, step size $\alpha = 0.01$ and number of iterations $P = 100$.

10.4. GTSRB Dataset:

Training Hyperparameters: We use momentum with a weight decay rate of 0.0002, momentum rate of 0.9, batch size 32, and training steps 100,000. We use the learning rate schedule as follows: the first 5000 steps have a learning rate of 10^{-5}; after 5000 steps and until 70,000 steps have a learning rate of 10^{-4}; after 70,000 steps have a learning rate of 10^{-5}. We use PGD attack as the adversary with a random start, the number of steps of 7, step size of 2, $m = 5$ as the number of steps for approximating IG computation in the attack step, and adversarial budget $\epsilon = 8$.

Evaluation Attacks Configuration: For evaluating adversarial robustness, we use PGD attack with the number of steps as 40, adversarial budget ϵ of 8, and step size of 2. For evaluating attributional robustness, we use IFIA's top-k attack with $k = 100$, adversarial budget $\epsilon = 8$, step size $\alpha = 1$ and number of iterations $P = 50$.

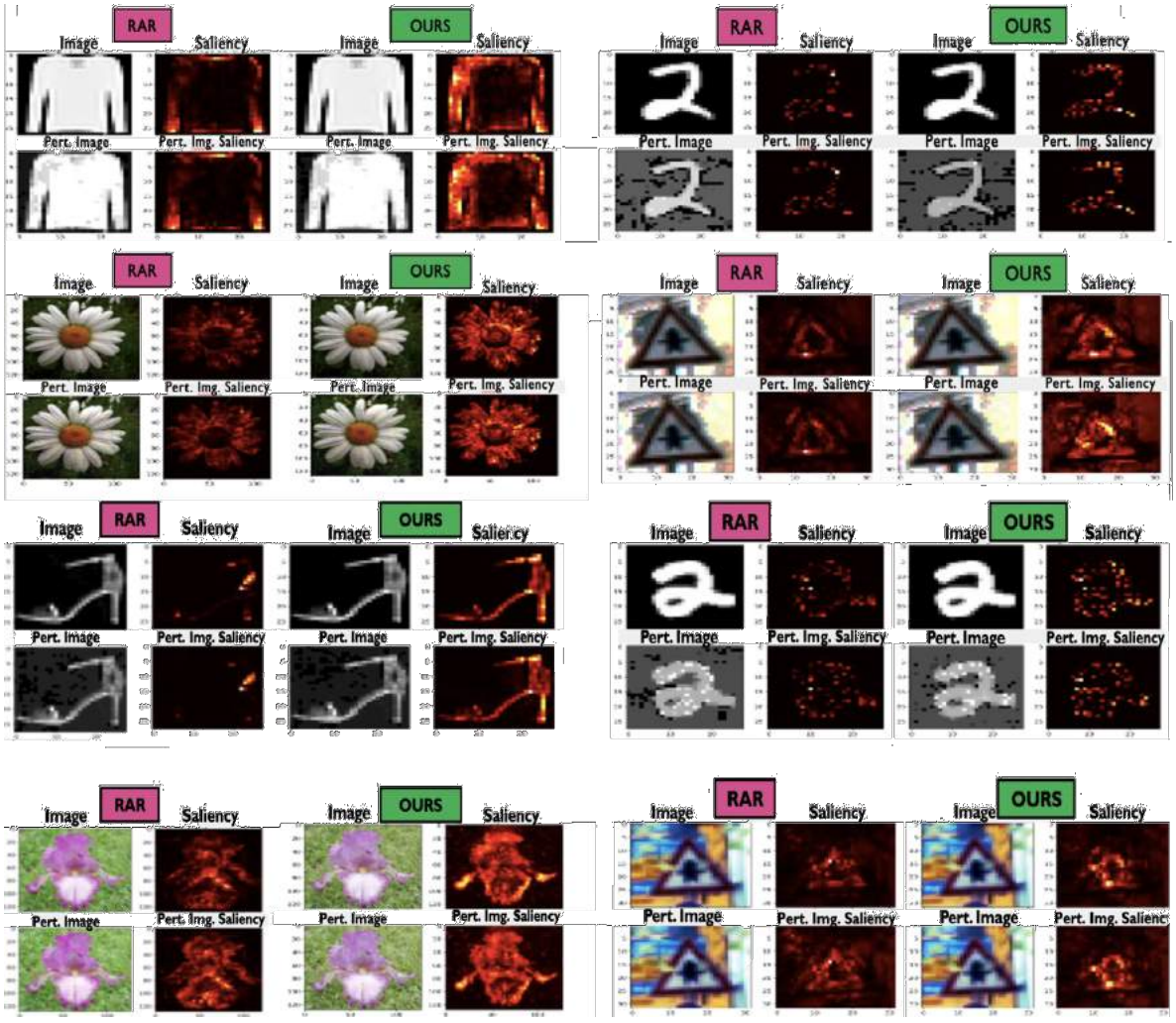

Figure 10. More comparative visualizations of the original image and human-imperceptible attributional attacked images (same predicted class) and their corresponding attribution maps obtained using RAR [7] and OUR model for test samples each from Fashion-MNIST, MNIST, Flower, and GTSRB datasets. Note that the model trained using RAR [7] generates a very similar attribution map before and after the attack, but none of them captures the salient part of the objects properly. On the other hand, OUR model generates similar attribution maps before and after an attack and can produce better quality attribution maps compared to RAR. These visualizations illustrate the importance of an alignment regularizer for improving the quality of the attribution map.

11. More Visualizations of Object-Alignment

We provide more visualizations in this section, consisting of generated attribution maps with images before and after attack by our method as well as [7], which we couldn't show in the main paper due to space constraints. These examples in Fig. 10 show that our joint training method not only helps to generate more interpretable attribution maps compared to [7] but also maintains the stability of attribution maps before and after the attributional attack.

12. Limitations & Broader Impacts

The limitations of the proposed method are confined to limitations in computational resources. Optimizing the objective function requires computing 2^{nd} order derivatives; since our method uses first-order derivatives, for IG, in the loss function, updating requires Jacobians' calculation, which can involve additional computation and memory. While our model needs additional offline training time (up to 1.3x times the baseline on average across the datasets, which is not prohibitive), its inference time, which matters in practice, is the same as earlier baseline models.

Future Work. As an extension of our current study, we like to study the efficacy of optimization problems in dealing with loss functions that involve first-order derivatives because optimizing them requires regulating higher-order derivatives. Moreover, regulating higher-order derivatives boils down to regulating the curvature of the loss

surface, another exciting direction to study the stability-interpretability tradeoff. Due to the reach of ML models to common people nowadays, the importance of the trustworthiness of AI/ML systems has increased multi-fold in recent years. Our method aims to enhance stability and object-alignment of attributions that help boost a model's applicability in the real world, especially in safety-critical applications such as healthcare and autonomous navigation, where interpretable and robust explanations are critical for end users.

13. Rationale behind Inner Maximization Step in Our Training Strategy

One of the natural questions that arise while trying to mitigate the tradeoff between stability and object-alignment is what happens if the network is trained with images that affect both facets. In other words, the dataset is augmented with images that are attacked adversarially on cross-entropy loss which affects the object-alignment of the attribution, and images that are attacked on their attributions which destabilize the attribution map. To explore this direction, we have conducted experiments that leverage these ideas in the inner maximization step. In all of these experiments, we use the terms X_{Adv} and X_{Attr}, which we explain below:

- The X_{Adv} is an image generated after maximizing the cross-entropy loss, i.e., vanilla PGD-attack:

$$\begin{aligned} X_{\text{Adv}} &= X + \delta \\ \delta &= \underset{\delta \in B(\epsilon)}{\arg\max}\, \mathcal{L}_{\text{CE}}(X+\delta) \end{aligned} \tag{18}$$

- The X_{Attr} is an image generated after maximizing the $S(IG(\cdot,\cdot))$ term of RAR-loss using the PGD-like framework:

$$\begin{aligned} X_{\text{Attr}} &= X + \delta \\ \delta &= \underset{\delta \in B(\epsilon)}{\arg\max}\, \|IG^{l_y}(X, X+\delta)\|_1 \end{aligned} \tag{19}$$

13.1. Separate-Images (SI-1)

In order to examine the validity of the iterative inner maximization step as in Eqn. 8, we perform an experiment as described below:

- Apply adversarial attack, such as PGD, on an input image X. We denote the adversarial image as X_{Adv}.
- For the same input image X, apply an attributional attack, such as IFIA. We denote the attribution attacked image as X_{Attr}.
- Use the following objective function in outer minimization to train a neural network:

$$\begin{aligned} \mathcal{L} &= L_{\text{CE}}(x) + [L_{\text{CE}}(X_{\text{Adv}}) + L_{\text{CE}}(X_{\text{Attr}})] \\ &+ [\|IG^{l_y}(X, X_{\text{Attr}})\|_1] + \lambda \mathcal{L}_{\text{diff}} \end{aligned} \tag{20}$$

We report the result of this experiment in Table 9. The experimental findings indicate that generating an adversarial image (X_{Adv}) and attributional image (X_{Attr}) separately through iterative inner maximization step in our proposed joint training framework do not help in practice. This highly suggests that even though the adversarial (X_{Adv}) and attribution sample (X_{Attr}) are nearby, training a model with these images is not suitable for fulfilling our objective of empowering a robust model with interpretable and stable explanations.

13.2. Separate-Images-2 (SI-2)

This experiment is a slight modification of the previous (SI-1) setup. Here, additionally, we apply a robust attribution regularizer between the adversarial sample (X_{Adv}) and the natural sample (X). The experiment can be done as follows:

- Apply adversarial attack, such as PGD, on an input image X. We denote the adversarial image as X_{Adv}.
- For the same input image X, apply an attributional attack, such as IFIA. We denote such attribution attacked image as X_{Attr}.
- Use the following objective function to train the neural network:

$$\begin{aligned} \mathcal{L} &= L_{\text{CE}}(x) + [L_{\text{CE}}(X_{\text{Adv}}) + L_{\text{CE}}(X_{\text{Attr}})] \\ &+ [\|IG^{l_y}(X, X_{\text{Attr}})\|_1 + \|IG^{l_y}(X, X_{\text{Adv}})\|_1] \\ &+ \lambda \mathcal{L}_{\text{diff}} \end{aligned} \tag{21}$$

The results for these experiments are provided in Table 9. As we see from Table 9, the outcome of this experiment (SI-2) is very similar to that of SI-1. Such results also reinforce the fact that generating (X_{Adv}) and (X_{Attr}) separately through iterative inner maximization step is not desirable. Hence, the iterative inner maximization step as in Eqn. 8 is a right fit in the proposed overall optimization for our joint training strategy.

14. More Visualizations of Stability

We also provide more visualizations in this section, consisting of generated attribution maps with images before and after attack by our method and [18], which we couldn't show in the main paper due to space constraints—these examples in Figs. 11 and 12 also reinforce the fact that our joint training method helps to generate more stable attribution maps compared to [18].

15. Conclusions and Future Work

we proposed a training framework to enhance robust models' saliencies and improve the stability and object-

Datasets	Methods	Clean	Adv. Acc.	Top-K	Kendall	AA-Top-K	AA-Kendall
F-MNIST	Nat	90.86	0.01	39.01	0.4610	5.41	0.1946
	AT [18]	85.73	**73.01**	46.12	0.6251	46.12	**0.6251**
	SI-1	**90.99**	55.07	62.34	0.5624	33.48	0.4588
	SI-2	89.71	60.9	62.15	0.5620	34.35	0.4708
	RAR [7]	85.44	70.26	72.08	0.6747	51.48	0.5754
	ERAR [21]	85.45	71.61	**81.50**	**0.7216**	**59.21**	0.6154
MNIST	Nat	**99.17**	0.00	46.61	0.1758	4.12	0.0021
	AT [18]	98.40	**92.47**	62.56	0.2422	62.56	0.2422
	SI-1	97.43	71.76	61.84	0.2354	32.58	0.1085
	SI-2	96.20	76.95	62.51	0.2397	34.42	0.1109
	RAR [7]	98.34	88.17	72.45	0.3111	58.42	**0.2851**
	ERAR [21]	98.41	89.53	**81.00**	**0.3494**	**66.45**	0.2821
Flower	Nat	**86.76**	0.00	8.12	0.4978	3.9	0.071
	AT [18]	83.82	41.91	55.87	0.7784	55.87	0.7784
	SI-1	84.23	36.25	52.93	0.6764	15.93	0.2119
	SI-2	85.57	37.64	51.85	0.6846	14.25	0.2354
	RAR [7]	82.35	47.06	66.33	0.7974	33.67	**0.8124**
	ERAR [21]	83.09	**51.47**	**69.50**	**0.8121**	**57.21**	0.7314
GTSRB	Nat	**98.57**	21.05	54.16	0.6790	4.21	0.041
	AT [18]	97.59	**83.24**	68.85	0.7520	62.56	0.2422
	SI-1	96.73	63.75	65.74	0.6351	39.54	0.1241
	SI-2	95.89	65.23	64.87	0.6523	38.29	0.1455
	RAR [7]	95.68	77.12	74.04	0.7684	43.0	0.2631
	ERAR [21]	97.61	82.33	**77.15**	**0.7889**	**67.41**	**0.3014**

Table 9. Comparative results of clean, adversarial, and attributional robustness achieved by models following the tradeoff experiments (as in Eqns. 20 and 21) and another baseline attributional robustness training methods. Here, by "**SI**" we mean *Separate Images* (as in section 13).

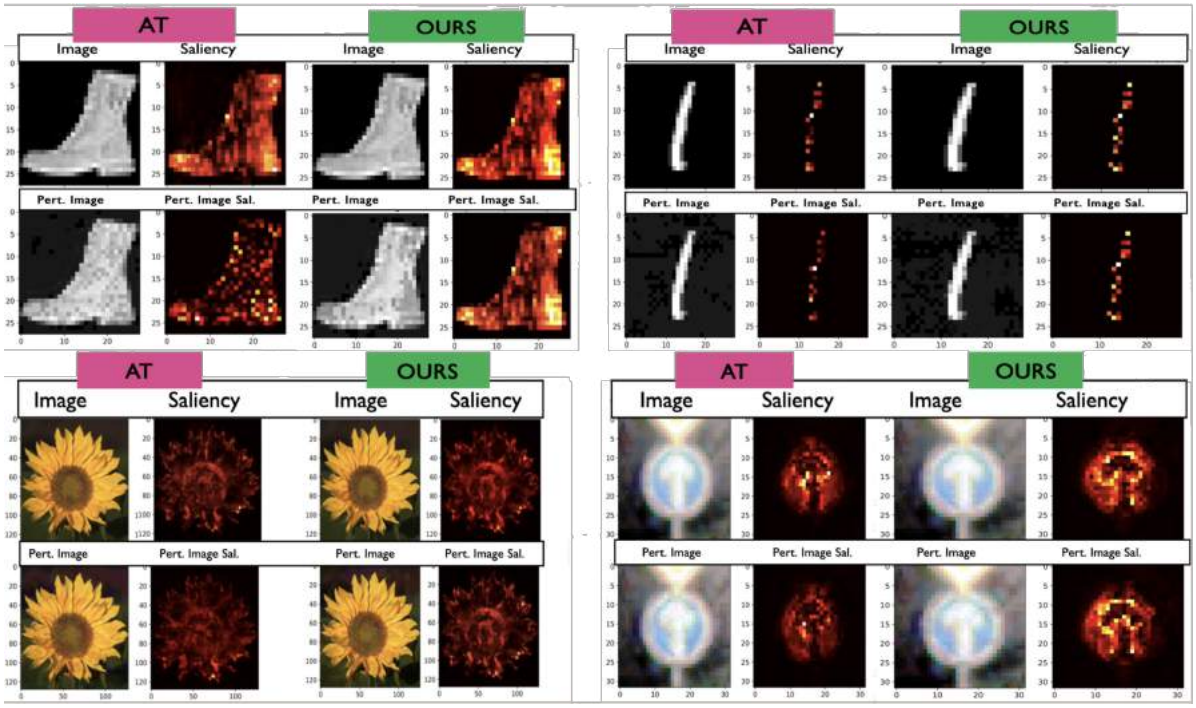

Figure 11. More comparative visualizations of Original image and human-imperceptible attributional attacked images (same predicted class) and their corresponding attribution maps obtained using AT [18] and OUR model for test samples each from Fashion-MNIST, MNIST, Flower, and GTSRB datasets. Note that the model trained using AT [18] generates a dis-similar attribution map before and after the attack. On the other hand, OUR model produces a better quality attribution map and generates similar attribution maps before and after an attack compared to AT [18]. These visualizations illustrate the importance of a joint training strategy for improving the stability of the attribution map without disturbing the object-alignment of the attribution map.

alignment of attribution maps without sacrificing the original network's goals. We also proposed new metrics to evaluate the robustness and quality of attribution maps. As an extension of our current study, we plan to study the efficacy of optimization problems in dealing with loss functions that require computing 2nd-order derivatives. Regu-

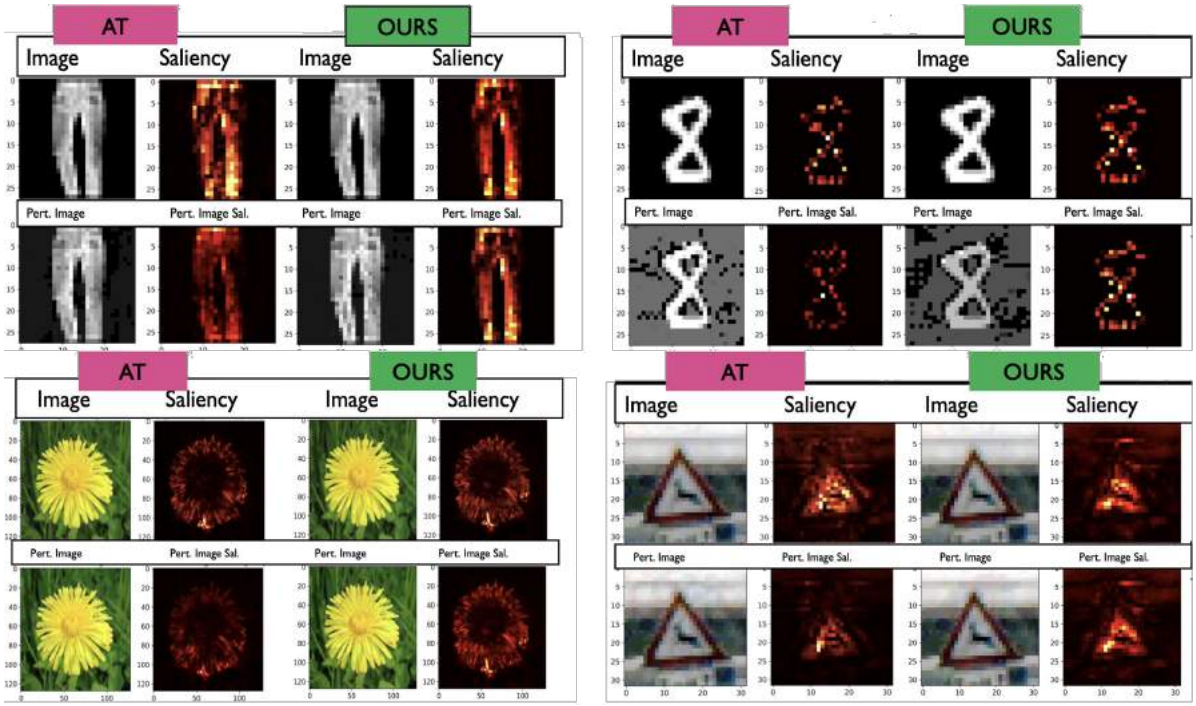

Figure 12. More comparative visualizations of Original image and human-imperceptible attributional attacked images (same predicted class) and their corresponding attribution maps obtained using AT [18] and OUR model for test samples each from Fashion-MNIST, MNIST, Flower, and GTSRB datasets. Note that the model trained using AT [18] generates a dis-similar attribution map before and after the attack. On the other hand, OUR model not only produces better quality attribution maps but also generates similar attribution maps before and after attacks compared to AT [18]. These visualizations illustrate the importance of a joint training strategy for improving the stability of the attribution map without disturbing the object-alignment of the attribution map.

lating higher-order derivatives requires analyzing the curvature of the loss surface, which is another interesting direction to study the stability-alignment relationship.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2018. 5

[2] Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Song. Curriculum adversarial training. *arXiv preprint arXiv:1805.04807*, 2018. 1, 5

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1, 5

[4] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise explanations of neural networks using adversarial training. *CoRR*, abs/1810.06583, 2018. 5

[5] Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 332–341, 2020. 2, 5

[6] Alvin Chan, Yi Tay, Yew Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. *arXiv preprint arXiv:1912.10185*, 2019. 5

[7] Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, and Somesh Jha. Robust attribution regularization. In *Advances in Neural Information Processing Systems*, pages 14300–14310, 2019. 1, 2, 4, 5, 6, 7, 8, 9, 12, 13, 15

[8] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017. 5

[9] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1

[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Discovering adversarial examples with momentum. *arXiv preprint arXiv:1710.06081*, 2017. 5

[11] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *International Conference on Machine Learning*, 2019. 1, 2, 4

[12] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019. 1, 2, 5, 6, 10

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 5, 6

[14] Jindong Gu and Volker Tresp. Saliency methods for explaining adversarial attacks. *CoRR*, abs/1908.08413, 2019. 5

[15] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018. 1, 5

[16] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 5

[17] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2010. 6

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16

[19] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006. 6

[20] Anindya Sarkar and Raghu Iyengar. Enforcing linearity in dnn succours robustness and adversarial image generation. In *International Conference on Artificial Neural Networks*, pages 52–64. Springer, 2020. 1

[21] Anindya Sarkar, Anirban Sarkar, and Vineeth N Balasubramanian. Enhanced regularizers for attributional robustness. *arXiv preprint arXiv:2012.14395*, 2020. 1, 4, 5, 6, 7, 8, 9, 12, 15

[22] Anindya Sarkar, Anirban Sarkar, Sowrya Gali, and Vineeth N Balasubramanian. Get fooled for the right reason: Improving adversarial robustness through a teacher-guided curriculum learning approach. *arXiv preprint arXiv:2111.00295*, 2021. 2, 5

[23] Mayank Singh, Nupur Kumari, Puneet Mangla, Abhishek Sinha, Vineeth N Balasubramanian, and Balaji Krishnamurthy. On the benefits of attributional robustness. *arXiv preprint arXiv:1911.13073*, 2019. 5, 6, 7, 9, 12

[24] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 6

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017. 1, 6

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 5

[27] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019. 1, 5

[28] Zifan Wang, Haofan Wang, Shakul Ramkumar, Matt Fredrikson, Piotr Mardziel, and Anupam Datta. Smoothed geometry for robust attribution. *arXiv preprint arXiv:2006.06643*, 2020. 5

[29] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *arXiv preprint arXiv:2004.05884*, 2020. 1

[30] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 6

[31] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving

adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019. 1, 5

[32] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 5

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6, 9

[34] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019. 5

[35] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736*, 2020. 1, 5