Are Classification Robustness and Explanation Robustness Really Strongly Correlated? An Analysis Through Input Loss Landscape

Tiejin Chen
Arizona State University
tchen169@asu.edu

Wenwang Huang
Independent Researcher
Weistrasshww@gmail.com

Linsey Pang
Paypal AI
panglinsey@gmail.com

Dongsheng Luo
Florida International University
dluo@fiu.edu

Hua Wei
Arizona State University
hua.wei@asu.edu

ABSTRACT

This paper looks into the critical area of deep learning robustness and challenges the common belief that classification robustness and explanation robustness in image classification systems are inherently correlated. Through a novel evaluation approach leveraging clustering for efficient assessment of explanation robustness, we demonstrate that enhancing explanation robustness does not necessarily flatten the input loss landscape with respect to explanation loss - contrary to flattened loss landscapes indicating better classification robustness. To further investigate this contradiction, a training method designed to adjust the loss landscape with respect to explanation loss is proposed. Through the new training method, we uncover that although such adjustments can impact the robustness of explanations, they do not have an influence on the robustness of classification. These findings not only challenge the previous assumption of a strong correlation between the two forms of robustness but also pave new pathways for understanding the relationship between loss landscape and explanation loss. Codes are provided in the supplement.

1. INTRODUCTION

Understanding the relationship between classification robustness and explanation robustness is critical for deploying reliable machine learning systems in real-world applications. In medical diagnostics, autonomous vehicles, and financial fraud detection, models must not only maintain accurate predictions under adversarial attacks (known as classification robustness [34]) but also preserve consistent, interpretable rationales (e.g., highlighted image regions) for their decisions during adversarial attacks (known as explanation robustness). Prior work has widely assumed these two properties are positively correlated [5, 22], leading to potential security gaps if they are not directly linked.

Our investigation reveals a fundamental challenge to this paradigm: *improving classification robustness provides no guarantee of enhanced explanation robustness*, challenging a key assumption in adversarial learning. This occurs because explanations, such as saliency maps, are themselves susceptible to adversarial perturbations [14, 15]. As illustrated in Figure 1, small adversarial perturbations can significantly alter explanation maps while leaving the model's classification unchanged. This raises a critical question about whether adversarial robustness in classification translates to robustness in model interpretability.

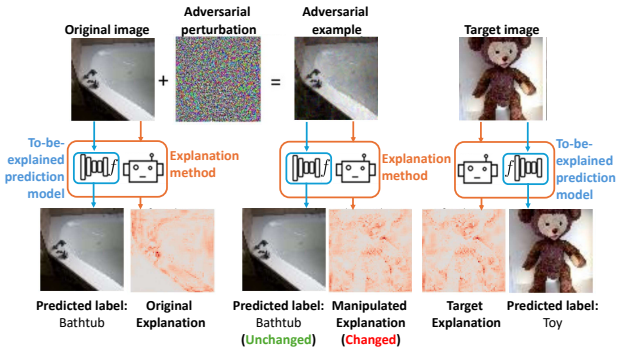


Figure 1: Example of an adversarial explanation attack. While the to-be-explained model f gives the same predicted label after adding adversarial noise, the explanation saliency map can be manipulated to the target saliency map, leading to wrong explanations.

To better understand robustness, one key approach is analyzing the input loss landscape [29]. Prior work has shown that a flatter input loss landscape with respect to classification loss indicates stronger classification robustness [57, 29]. As visualized in Figure 2, adversarially trained models, which show increased classification robustness, exhibit a flatter input loss landscape compared to normally trained models. Since classification robustness is linked to a flatter loss landscape, a natural question arises: *Does increasing explanation robustness lead to a flatter input loss landscape for explanation loss?*

Surprisingly, this paper shows that increasing explanation robustness does not flatten the input loss landscape with respect to explanation loss. To systematically analyze the relationship between explanation robustness and input loss landscapes, we generate models with varying levels of explanation robustness using adversarial training methods [60]

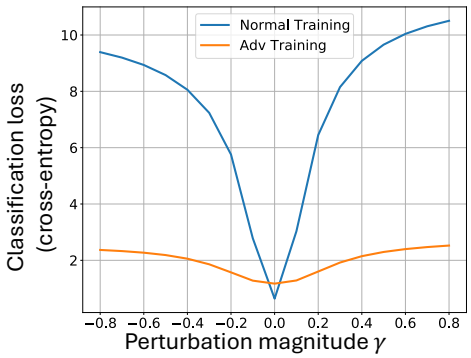


Figure 2: Comparison of input loss landscapes between normal and adversarially trained models on CIFAR-10. Adversarially trained models show flatter loss landscapes and increased classification robustness.

that allow fine-grained control over classification robustness. Our findings (detailed in Figure 4) contradict previous assumptions that classification and explanation robustness are inherently linked.

To further investigate this phenomenon, we reverse the question: *Does a flatter input loss landscape for explanation loss improve explanation robustness?* This paper shows that the answer is no. Flattening the input loss landscape for explanation loss actually decreases explanation robustness. To confirm this, we propose **S**eparate **E**xplanation Robustness via **P**GD (`SEP`) to explicitly control the input loss landscape w.r.t explanation robustness. By regularizing on the input loss landscape w.r.t explanation robustness, `SEP` reduces explanation robustness while leaving classification robustness unchanged. Extensive results on different model architectures, datasets, and explanation methods demonstrate the effectiveness of `SEP` in decoupling explanation robustness and classification robustness. This demonstration challenges the widely held belief that classification and explanation robustness are positively correlated. In summary, the key contributions of this paper are as follows:

- We introduce a clustering-based sampling method to efficiently evaluate explanation robustness.
- We use TRADES [60] to control classification robustness and visualize the input loss landscape with respect to explanation loss, revealing that increasing explanation robustness does not flatten the input loss landscape, which contradicts with classification robustness.
- We develop a novel training method that flattens the input loss landscape for explanation loss and show that, contrary to prior assumptions, this reduces explanation robustness, suggesting that explanation and classification robustness are not strongly correlated.

2. RELATED WORK

Adversarial Attack and Adversarial Training (AT). It has been proven that deep learning models are vulnerable to adversarial examples [48, 17, 8], where noise that is imperceptible to humans, when added to the original inputs, can lead to the misclassification of models. Projected Gradient Descent (PGD) [34] is one of the most popular methods that generate such a noise or evaluate models' classification robustness by calculating accuracy under its attack. Many methods [37, 58, 30, 45, 7] have been introduced to defend against adversarial attacks, while they do not involve a training process and may be vulnerable to adaptive attack [2]. Goodfellow et al. [17] first introduced adversarial training (AT), which trains a model from scratch with adversarial samples and proves its performance, including adversarial competitions [40, 32, 34, 6]. In this paper, we focus on classification robustness increased by AT methods like Madry adversarial training [34] and TRADES [60]. Many works tend to increase the performance of AT through external datasets [19, 9, 53], metric learning [35], self-supervised learning [11], ensemble learning [51], label smoothing [12], and Taylor Expansion [23]. Wu et al. [55] found that obtaining a flat loss landscape can help increase classification robustness, which inspired the ideas in this paper.

Explanation Robustness. Saliency maps [42, 41, 3, 39] are widely used to explain image-related tasks in deep learning, and our focus is on the robustness of these explanations. However, similar to an adversarial attack, it is possible to find an adversarial noise on original images so that it can easily manipulate the saliency maps without changing classification results in both white-box [14, 15, 20, 44] and black-box settings [49]. Zhang et al. [61] further introduced a new method that can attack both saliency maps and classification results. In order to evaluate the explanation robustness, Wicker et al. [54] introduced the max-sensitivity and average-sensitivity of saliency maps. Alvarez et al. [1] estimated explanation robustness by the Local Lipschitz of interpretation, while Tamam et al. [49] directly used attack loss to evaluate explanation robustness. In this paper, we use attack loss based on the proposed cluster method to evaluate explanation robustness.

Several works have also aimed to improve explanation robustness. Chen et al. [10] introduced a regularization term during training to make the explanation more robust. Boopathy et al. [5] improved the performance by training with noisy labels. Tang et al. [50] proposed a first-order gradient-based approach to reduce computational training costs. Huang et al. [22] explored genetic algorithms to optimize for stronger explanation robustness.

Relationship between Classification Robustness and Explanation. It has been suggested that improved interpretability contributes to classification robustness [43, 52], implying a correlation between high-quality saliency maps and classification performance. Studies have demonstrated that models robust to explanation attacks tend to be more resistant to classification attacks as well, supporting the idea that explanation robustness benefits classification robustness [5, 50, 22]. However, our work challenges this assumption. We show that adversarial training with TRADES [60] can improve explanation robustness while sometimes enhancing classification robustness. Yet, further analysis reveals that these two aspects of robustness are not inherently linked—classification robustness does not necessarily ensure explanation robustness. This finding suggests a fundamental disconnect between the two, which we investigate in depth.

3. ANALYSIS AND METHODOLOGY

This section establishes a systematic framework to investi-

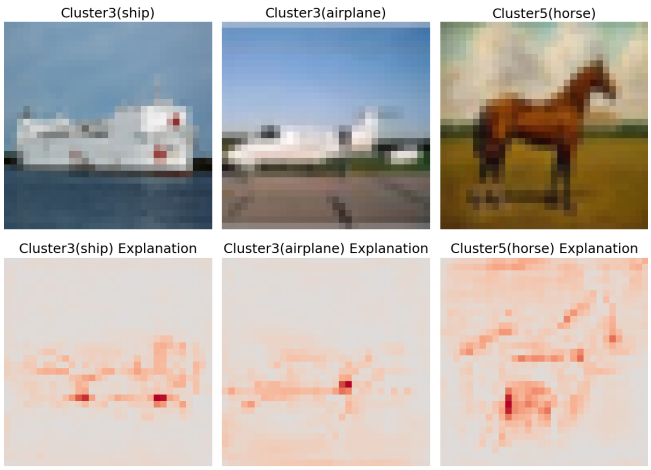


Figure 3: Explanation consistency within clustered groups: (a) Cluster 3 exhibits focused attention on the shape of machines though they have different classification labels, while (b) Cluster 5 highlights the shape of horses.

gate the relationship between classification robustness and explanation robustness, ultimately developing our proposed solution (SEP) to address observed contradictions. We create a controlled level of explanation robustness using TRADES, then quantify explanation robustness through clustered evaluation, analyze loss landscape paradoxes, and finally introduce SEP for targeted landscape engineering.

3.1 Classification and Explanation Robustness

3.1.1 Classification Robustness

Our systematic investigation begins by establishing precise control over model robustness characteristics through the TRADES framework [60]. TRADES provides fine-grained control of classification robustness via its parameterized loss function and the parameter α:

$$\mathcal{L}_{\text{TRADES}} = \underbrace{\mathcal{L}_{\text{sc}}(f(x), y)}_{\text{Standard Classification Loss}} + \alpha \underbrace{\mathcal{L}_{\text{adv}}(f(x), f(x_{\text{adv}}))}_{\text{Adversarial Regularization}} \tag{1}$$

where $f(x) \in \mathbb{R}^C$ denotes model outputs for input x with C classes, x_{adv} is generated via projected gradient descent (PGD) attack [34]: $x_{adv} = x + \epsilon_{adv}$ with $|\epsilon_{adv}|_\infty \leq \xi$, where the ξ is a pre-defined constraint. The adversarial regularization term $\mathcal{L}_{\text{adv}}$ measures the KL divergence between original and adversarial predictions. The α parameter explicitly controls the tradeoff between clean accuracy and adversarial robustness, enabling controlled robustness levels.

Measuring Classification Robustness. Following existing work [34, 60], classification robustness is commonly measured by adversarial accuracy (Acc_{adv}), which is the accuracy under adversarial attack on classification:

$$Acc_{\text{adv}} := \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(f(x_{\text{adv},i}) = y_i\right), \tag{2}$$

Where $\mathbb{I}$ is the indicator function. Compared with the clean accuracy ($Acc_{\text{clean}} := \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(f(x_i) = y_i)$) which measures classification performance, adversarial accuracy Acc_{adv} measures how good a model is against adversarial attacks on classification. Therefore, a higher Acc_{adv} indicates a better classification robustness.

3.1.2 Explanation Robustness

Previous works on adversarial attacks have extended the framework to explanation through finding a perturbation δ^* with constrained optimization [49]:

$$\delta^* = \arg\min_{|\delta|_\infty \leq \xi} \|I_f(x_v + \delta) - I_f(x_t)\|_2 \tag{3}$$

where x_v represents the victim (original) image, x_t is the target image with desired explanation pattern, f is the to-be-explained model, and $I_f(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ denotes an explanation method (e.g., Grad-CAM [39] or Saliency Maps [42]) that explains the behavior of model f on an input image.
Explanation robustness is commonly measured by the following definition of attack loss:

Definition 1 (Explanation Attack Loss)**.** *Given a victim image x_v, a target image x_t with desired explanation pattern, a to-be-explained model f and an explanation method $I_f(\cdot)$, the explanation attack loss is $\mathcal{L}_e(x_v, x_t) = \|I_f(x_v + \delta) - I_f(x_t)\|_2^2$, where δ is the optimization target.*

Since the ultimate goal of adversarial attacks on explanation is to manipulate the explanation of the victim image to resemble that of the target image, a higher explanation attack loss indicates a bigger difference between two explanations $I_f(x_v + \delta)$ and $I_f(x_t)$, which means a less successful attack.

Generating $\mathcal{D}_e$ with cluster-based sampling. In order to estimate the explanation robustness of a model f, existing methods use evaluation set $\mathcal{D}_e = \{(x_v, x_t) | x_v, x_t \in \mathcal{D}_{\text{origin}}$, where $\mathcal{D}_{\text{origin}}$ is the images in the original classification dataset. As the number of samples in $\mathcal{D}_{\text{origin}}$ increases, the size of $\mathcal{D}_e$ grows quadratically, which would be computationally infeasible.
To ensure comprehensive coverage of the explanation space while maintaining computational traceability, we introduce a cluster-based sampling protocol grounded in explanation consistency. Specifically, the protocol operates through three systematic steps:

1. **Feature Extraction**: Compute high-level representations using ResNet18's penultimate layer [18], capturing semantically meaningful features for explanation consistency

2. **Clustering**: Apply clustering algorithm. In our experiment, we use K-means [33] with $k = 10$ on CIFAR10.

3. **Representative Sampling**: Select N prototypes per cluster and forms evaluation set $\mathcal{D}_e$ for each cluster. In our experiment we set $N = 15$, forming evaluation set $\mathcal{D}_e$ with 150 images for 10 clusters and $150 \times 149 = 22,350$ unique (x_v, x_t) pairs in total.

As shown in Figure 3, images from the same cluster have similar explanations. We also report the explanation loss for intra-cluster and inter-cluster pairs to show that our clustering method indeed makes intra-cluster pairs share similar

Table 2: Comparison of classification robustness and explanation robustness of models trained with TRADES and different α on CIFAR10. Within a certain range, using the TRADES training method and increasing the value of α can not only improve the classification robustness but also improve the explanation robustness.

α	Acc_{adv} (%)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)
0	0.00	6.206	10.375
0.5	23.57	10.640	16.635
1.0	28.31	10.946	17.271
2.0	31.77	10.965	17.290
4.0	33.28	11.293	18.004
5.0	33.98	11.469	18.278
10.0	34.87	11.592	18.643

explanations quantitatively in Table 1, which shows intra-cluster pairs do have a smaller loss.

ResNet18	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)
Intra-cluster	13.726
Inter-cluster	15.437

Table 1: Explanation loss at start of intra and inter clusters. The smaller explanation loss in the intra-cluster shows that images in the same cluster have similar explanations.

Measuring Explanation Robustness. Specifically, we quantify the explanation robustness through the final explanation loss after explanation attack converges: $\mathcal{L}_e^{\text{end}} = \mathbb{E}_{(x_v,x_t)\sim\mathcal{D}_e}[\mathcal{L}_e(x_v+\delta^{\text{end}}, x_t)]$, where δ^{end} is the optimal perturbation found by adversarial attacks on explanation. A higher $\mathcal{L}_e^{\text{end}}$ indicates that the attack struggles to manipulate the model's explanations, suggesting stronger explanation robustness. We can also measure the explanation loss before the explanation attack starts:

$$\mathcal{L}_e^{\text{start}} = \mathbb{E}_{(x_v,x_t)\sim\mathcal{D}_e}[\mathcal{L}_e(x_v+\delta^{\text{start}}, x_t)], \tag{4}$$

here δ^{start} is a random initial perturbation. It should be noted that $\mathcal{L}_e^{\text{start}}$ does not indicate the explanation robustness of a model since it is calculated before the explanation attack applies. However, providing $\mathcal{L}_e^{\text{start}}$ could let us know the difficulty of the attack before the beginning of the attack. Besides, we could also use the difference between $\mathcal{L}_e^{\text{start}}$ and $\mathcal{L}_e^{\text{end}}$ to roughly estimate the flatness of the loss landscape during training, while the definition for loss landscape at one point is introduced in the next section. After defining $\mathcal{L}_e^{\text{start}}$, we could obtain the quantitative results of explanation loss at start for the intra-cluster and inter-cluster for the clusters we obtain in the last section. In detail, Table 1 represents the intra-cluster and inter-cluster results for $\mathcal{L}_e^{\text{start}}$ on ResNet18 and CIFAR10 dataset. From the results, we can see that $\mathcal{L}_e^{\text{start}}$ for intra-cluster is indeed smaller than inter-cluster results, indicating our cluster-based method can obtain similar explanations in the cluster.

3.1.3 Input Loss Landscape

An input loss landscape is a visualization of how a model's loss changes across different possible input values [1], mapping the terrain of the loss function with respect to the input space, allowing researchers to understand how sensitive the model is to variations in the input data and identify potential robustness issues. Prior work has shown that a flatter input loss landscape for classification loss $\mathcal{L}_{\text{sc}}$ indicates stronger classification robustness [57, 29].

Following existing works [1], we visualize the input loss landscape w.r.t explanation loss $\mathcal{L}_e(x_v, x_t; f)$ by plotting the change of $\mathcal{L}_e$ when adding a random noise $\mathbf{d}$ to the victim image x_v with different magnitude γ:

$$l(\gamma) = ||I_f(x_v + \gamma d) - I_f(x_t)||^2, \tag{5}$$

where $\mathbf{d}$ is sampled from a standard Gaussian distribution.

3.2 Contradictory Findings on CIFAR10

Previous work has two assumptions: (1) a model with good classification robustness has a flat loss landscape w.r.t classification loss [57, 29]; (2) classification robustness and explanation robustness are positively correlated [5, 22]. A natural conclusion would follow if the previous assumptions are true: a model with good classification robustness might have a good explanation and thus a flat loss landscape w.r.t explanation loss as well.

3.2.1 Step 1: Training models with different levels of classification robustness

Our systematic investigation begins by establishing several models with different classification robustness through the TRADES framework [60] by training with different α in Equation (1). The preliminary results on the CIFAR10 dataset [26] can be found in Table 2. We can observe that with the increase of α, the adversarial accuracy Acc_{adv} increases as well. This indicates that the model adversarially trained with TRADES shows better classification robustness with the increase of α. This step provides us with models with different classification robustness.

3.2.2 Step 2: Measuring explanation attack loss for models with different classification robustness

After getting the models with different classification robustness, the next step is to measure their explanation robustness, i.e., whether they perform differently under explanation attacks. With the assumption (2) classification robustness and explanation robustness are positively correlated [5, 22], we would expect that models with better classification robustness will show better explanation robustness. That is, models with higher Acc_{adv} will have higher $\mathcal{L}_e^{\text{end}}$. The preliminary results in Table 2 shows that with the increase of α, models indeed have higher $\mathcal{L}_e^{\text{end}}$. But since $\mathcal{L}_e^{\text{start}}$ is also higher with the increase of α. Actually, calculating the $\Delta\mathcal{L} = \mathcal{L}_e^{\text{start}} - \mathcal{L}_e^{\text{end}}$, we could find that, after training, the attack is even more successful, considering the loss decrease. Therefore, we **cannot conclude** that better explanation robustness owes to better classification robustness. Next, we will explore a different way of measuring explanation robustness with the loss landscape.

3.2.3 Step 3: Connecting Robustness with Loss Landscape

To further investigate the explanation robustness of models, we visualize the input loss landscape w.r.t explanation

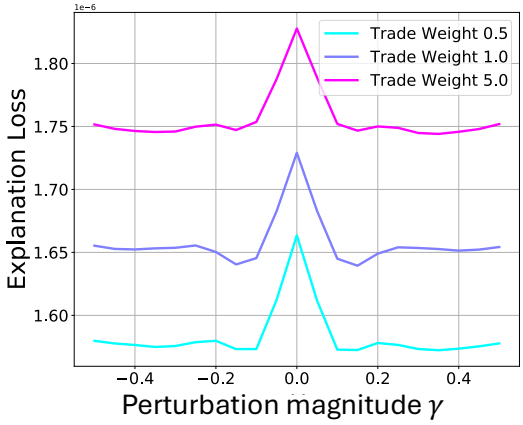


Figure 4: Input loss landscape w.r.t explanation loss for models trained with different Trade Weight α in TRADES. The loss landscape does not show a clear difference between models that vary in explanation robustness because the loss change remains the same.

loss. With the assumption (1) a model with good classification robustness has a flat loss landscape w.r.t classification loss [57, 29], we would expect a similar analogy for explanation robustness: a model with good explanation robustness has a flat loss landscape w.r.t explanation loss.

Using Equation (5), we visualize the input loss landscape by plotting the change of explanation loss, and the results are shown in Figure 4. We also visualize the input loss landscape with normal training and Madry adversarial training (MAT) in Appendix Figure 9. From Figure 4, we can find that the input loss landscape w.r.t. explanation loss **does not** show a difference in flatness, despite that their explanation loss $\mathcal{L}_e$ are different. Different from the conclusions drawn in classification robustness, models with good explanation robustness do not exhibit a flat loss landscape w.r.t explanation loss. This contradiction motivates us to further explore and propose a method to decouple classification and explanation robustness in the following section by changing explanation robustness by flattening the input loss landscape w.r.t explanation robustness, while maintaining classification robustness.

3.3 Landscape-Aware Regularization

Motivated by the observed landscape-robustness contradiction, in this section, we propose a method to decouple classification and explanation robustness. Specifically, we propose a new training algorithm to control the input loss landscape w.r.t explanation robustness and see if models with different flatness will perform differently on explanation robustness. If we can have a training algorithm that can influence explanation robustness while not changing classification robustness, we can conclude that classification robustness and explanation robustness are not strongly correlated.

To explicitly control the input loss landscape w.r.t explanation robustness, we propose **S**eparate **E**xplanation Robustness via **P**GD (SEP) through the following loss function:

$$\mathcal{L}_{\text{SEP}} = \|I(x+\zeta) - I(x)\|_2^2, \tag{6}$$

where ζ is a noise randomly sampled from a standard Gaussian distribution and I is the explanation method. Note that off-the-shelf explanation AT [49, 14] must be executed in a targeted setting: A victim image and a target image are required for the explanation of adversarial attacks. Calculating ζ through a targeted setting may increase the training time and increase the probability that the model overfits the chosen pairs. Therefore, we use randomly sampled noise, like for ζ which does not require a target image.

Algorithm 1 SEP Algorithm

Input: Dataset $\mathcal{D}$, total training iteration T, explanation method I, model weights $\mathbf{w}$, and balancing factor λ.
for $t = 0$ **to** $T-1$ **do**
  for batch x in $\mathcal{D}$ **do**
    Sample a random noise ζ from a standard Gaussian distribution.
    Get adversarial samples (on classification): $x_{adv} = PGD(x, y)$.
    Calculate loss function with Equation (7).
    Update $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \mathcal{L}(f(x), f(x_{adv}), y|\mathbf{w})$
  end for
end for

The new loss function $\mathcal{L}_{\text{SEP}}$ can be incorporated into existing training frameworks, including Madry adversarial training [34], TRADES [60], and normal training. In this paper, we will mainly focus on Madry adversarial training plus the new training loss:

$$\mathcal{L} = \mathcal{L}_{sc}(f(x_{adv}), y) + \lambda \mathcal{L}_{\text{SEP}}. \tag{7}$$

where the hyperparameter λ balances two components of the losses. When $\lambda > 0$, we have SEP_{pos} which guides the loss landscape to become flatter; when $\lambda < 0$, we have SEP_{neg}, which guides the loss landscape to become sharper. The overall training algorithm is shown in Algorithm 1.

In the following section, we show that our method can influence explanation robustness while it does not change classification robustness. Please note that our method is designed and used to explore the relationship between the classification robustness and explanation robustness instead of common dimensions that a normal algorithm will care like the performance. We also visualize the comparison of saliency maps from models trained with different algorithms to provide how our methods influence the saliency maps in Figure 7 in the later section to show how our method works.

4. EXPERIMENTS

We conduct comprehensive experiments across multiple datasets and model architectures to validate our method's ability to decouple explanation robustness from classification robustness. Our evaluation addresses three key research questions:

- **RQ1:** Can we independently control explanation robustness across different datasets and model architectures without changing classification robustness?
- **RQ2:** How do different explanation methods affect this relationship between two robustnesses?
- **RQ3:** Does our method generalize across different training protocols, e.g., different adversarial training methods like TRADES?

4.1 Experimental Settings

Table 3: Performance of models trained with ConvNet and ResNet18 on various datasets is evaluated using four training methods, w.r.t. $\mathcal{L}_e^{\text{end}}$ and Acc_{adv}. Higher $\mathcal{L}_e^{\text{end}}$ indicates better explanation robustness; higher Acc_{adv} indicates better classification robustness. $\mathcal{L}_e^{\text{start}}$ is also included to show our method's influence on explanation robustness. The **best** performance in explanation and classification robustness and the worst performance in explanation robustness are highlighted. There is no positive correlation between explanation and classification robustness achieved through $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$ training methods, compared to MAT.

Model Architecture	ConvNet				ResNet18			
MNIST								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	261.183	204.825	99.29	0.00	266.834	146.16	99.36	0.00
MAT	373.262	298.729	99.00	89.92	916.017	778.003	99.28	**94.60**
$\texttt{SEP}_{pos}$	93.033	61.545	98.8	89.4	92.371	59.278	98.4	91.63
$\texttt{SEP}_{neg}$	806.204	**657.180**	98.97	**90.34**	9356.306	**8248.627**	99.4	93.95
FMNIST								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	106.530	72.198	92.32	0.00	128.640	69.847	91.57	0.00
MAT	386.370	274.267	62.85	73.98	588.610	417.031	79.22	**67.10**
$\texttt{SEP}_{pos}$	35.588	22.465	69.88	**86.81**	32.466	22.512	68.75	56.51
$\texttt{SEP}_{neg}$	1811.969	**994.818**	62.75	76.89	8050.942	**7593.650**	70.23	57.55
CIFAR10								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	10.375	6.206	79.08	0.00	13.982	6.130	81.32	0.00
MAT	16.913	6.906	64.85	35.11	31.959	21.879	67.22	29.09
$\texttt{SEP}_{pos}$	3.565	1.269	64.94	**35.25**	11.962	7.958	66.68	**29.69**
$\texttt{SEP}_{neg}$	19.002	**7.590**	64.56	34.86	70.159	**36.276**	39.17	29.32
CIFAR100								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	10.099	6.140	48.39	0.05	12.044	4.716	41.24	0.00
MAT	20.642	13.650	36.4	17.35	33.456	22.623	36.14	15.70
$\texttt{SEP}_{pos}$	13.650	9.932	37.41	**17.98**	19.217	12.744	34.83	15.16
$\texttt{SEP}_{neg}$	22.506	**14.970**	36.17	17.43	35.525	**24.289**	34.80	**15.87**
TinyImageNet								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	0.966	0.633	28.71	0.00	1.131	0.528	28.34	0.00
MAT	2.426	1.728	25.13	9.55	3.119	2.349	26.34	10.81
$\texttt{SEP}_{pos}$	2.242	1.571	24.83	**9.63**	1.967	1.435	25.96	**10.83**
$\texttt{SEP}_{neg}$	3.873	**2.610**	24.31	9.61	4.413	**3.016**	26.11	10.74

4.1.1 Datasets

To thoroughly demonstrate the impact of our proposed training method and the resulting conclusions, we evaluate on five standard benchmarks:

- **CIFAR10** [26]: consisting of 60k 32×32 color images in 10 classes including 50k training images and 10k test images.
- **CIFAR100** [26]: containing the same images as CIFAR10 but has a more refined label with 100 categories, which makes it a harder dataset.
- **MNIST** [28]: containing 60k training samples and 10k test samples from 10 digit classes. Each digit is a 28×28 grayscale image.
- **Fashion MNIST** [56]: consisting of 60k training samples and $10k$ test samples from 10 classes. Each sample is a 28×28 grayscale image in a clothes category.
- **TinyImageNet** [27]: it is a subset of ImageNet [13] with 64x64 pixels and 200 categories

We also consider using ImageNet [13] and the experiment results for ImageNet can be found in Appendix D.3. The results for ImageNet are similar to the results here.

4.1.2 Architecture of To-be-explained Model

In addition to utilizing diverse datasets, we have also designed four distinct model architectures for training on these datasets. We conduct experiments on ConvNet, ResNet [18], Wide ResNet [59] and MoblieNetV2 [21, 38]. The ConvNet model consists of three convolutional layers and one fully connected layer from Gidaris et al. [16]. For ResNet and Wide ResNet, we use a standard ResNet18 and Wide-ResNet-28, respectively. We also adjust the ResNet, Wide ResNet, and MoblieNetV2 so that they can fit into all datasets we use. All models use Softplus activation for explanation attack compatibility [14], maintaining ReLU-like behavior with improved differentiability.

4.1.3 Explanation Methods

For explanation methods, we mainly use the implementations of five explanation methods from Captum [25]: Gradient [4], Gradient × Input [41], Guided Backpropagation [46], Deep Lift [41] and Integrated Gradients [47]. Their detailed descriptions can be found in Appendix A.

4.1.4 Training Protocols

We consider two baselines: normal training (Normal) and Madry adversarial training (MAT) [34]. We also explore two variants of the proposed method: $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$, as mentioned in Section 3.3. In the rest of this paper, unless specified, we will use $\lambda = 50000$ for $\texttt{SEP}_{pos}$ and $\lambda = -3000$ for $\texttt{SEP}_{neg}$. We use a learning rate of 0.01 for ConvNet and MobileNet while using a learning rate of 0.001 for ResNet and Wide ResNet.

4.1.5 Hyperparameters

For all experiments, we train our models for 25 epochs with 64 as the batch size. We also consider different training

epochs and our conclusion remains the same as shown later. To accelerate the training process, we use Adam [24] as the optimizer. We use the standard settings in adversarial training [36], with $\epsilon = 8/255$ in PGD for RGB images and $\epsilon = 0.3$ for grayscale images, and steps in PGD are set to 10 for all experiments. We list the detailed hyperparameters in the Appendix Table 10.

4.1.6 Metrics

As mentioned in Section *3.1.2*, we measure explanation robustness using the explanation loss in the end after explanation attack $\mathcal{L}_e^{\text{end}}$. A higher $\mathcal{L}_e^{\text{end}}$ indicates a worse attack and thus better explanation robustness. We also report the explanation loss before explanation attack $\mathcal{L}_e^{\text{start}}$ to show the influence of our method on the explanation loss landscape. For classification robustness, we report adversarial accuracy Acc_{adv}, with higher values indicating better classification robustness. Additionally, we include clean accuracy Acc_{clean} to ensure the models function normally in non-adversarial settings.

4.2 Decoupling Robustness Dimensions (RQ1)

We extend our preliminary experiments on CIFAR10 in Section 3.2 to multiple model architectures and datasets with Gradient × Input as the explanation method. The results are shown in Table 3 for ConvNet and ResNet18 with additional results for W-ResNet and MoblieNetV2 provided in the Appendix Table 9. We have the following observations:

- Across all the datasets and model architecture, $\texttt{SEP}_{pos}$ has the lowest $\mathcal{L}_e^{\text{end}}$, and $\texttt{SEP}_{neg}$ has the highest $\mathcal{L}_e^{\text{end}}$. This indicates that $\texttt{SEP}_{pos}$ shows weaker explanation robustness and $\texttt{SEP}_{neg}$ shows the strongest explanation robustness. The different performance w.r.t. explanation loss at end for $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$ is mainly induced by the difference in $\mathcal{L}_e^{\text{start}}$, which is influenced by our training method by setting λ to positive or negative.
- While $\texttt{SEP}_{pos}$, $\texttt{SEP}_{neg}$, and MAT have very similar Acc_{adv}, $\texttt{SEP}_{pos}$ shows the weakest explanation robustness by having the lowest $\mathcal{L}_e^{\text{end}}$, and $\texttt{SEP}_{neg}$ shows the strongest explanation robustness. These results show that despite the classification robustness of $\texttt{SEP}_{pos}$, $\texttt{SEP}_{neg}$, and MAT being similar, their explanation robustness is different. Therefore, we argue that there is no inherent relationship between explanation robustness and classification robustness.
- In the setting of CIFAR10 and ResNet18, increasing the explanation robustness by $\texttt{SEP}_{neg}$ hurts the Acc_{clean} while it still does not change Acc_{adv}, which represents classification robustness. This observation further validates our argument: classification robustness and explanation robustness may not be strongly correlated.
- From the results, we could see that though $\texttt{SEP}_{neg}$ has a much higher $\mathcal{L}_e^{\text{end}}$ compared with the $\texttt{SEP}_{pos}$. However, if we consider $\Delta\mathcal{L} = \mathcal{L}_e^{\text{start}} - \mathcal{L}_e^{\text{end}}$, we can find $\Delta\mathcal{L}$ for $\texttt{SEP}_{neg}$ is much higher than $\mathcal{L}_e^{\text{end}}$. This loss decrease demonstrates that our design is working since the larger loss decrease indicate $\texttt{SEP}_{neg}$ has a sharper loss landscape. Considering all the results, we hypothesis that the explanation robustness actually comes from the increase of the attack difficulty at the starting point instead of a flat loss landscape that is hard for the attacker to optimize. While previous works [57, 29] show that classification robustness comes from the flat loss landscape, this mechanism difference between classification and explanation robustness also further indicate that they might not be strongly correlated.

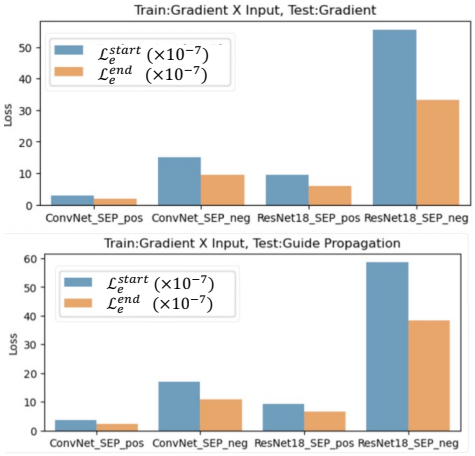


Figure 5: Performance of varying explanation methods in the testing phase, w.r.t. $\mathcal{L}_e^{\text{start}}$, $\mathcal{L}_e^{\text{end}}$, Acc_{clean} and Acc_{adv}. Models are trained with Gradient x Input on CIFAR10 and tested on different explanation methods: Gradient (Top) and Guide Propagation (Bottom). Even if the explanation methods during training and testing are different, $\texttt{SEP}_{pos}$ shows a lower explanation loss compared to $\texttt{SEP}_{neg}$, while they have similar adversarial accuracy.

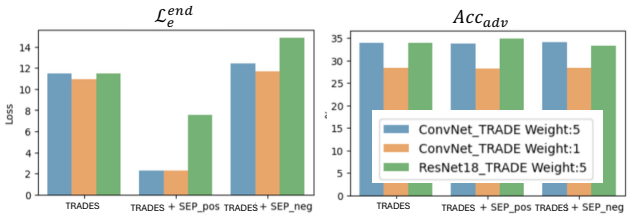


Figure 6: The test results of the model trained using the TRADE training method with CIFAR10, combined with our approach, are presented. The findings indicate that when we apply our method to an alternative adversarial training method TRADES, distinct from MAT, we can still find that the classification robustness and explanation robustness are not inherently interconnected.

4.3 Explanation Method Agnosticism (RQ2)

4.3.1 Training Phase Variations

In the previous experiment, we demonstrated that under the Gradient × Input explanation method, our methods achieve similar classification robustness while exhibiting significantly different explanation robustness. To further investigate whether this conclusion holds for different explanation methods, we changed the explanation method to Gradient and Guide Propagation. The results are based on CIFAR10 and are summarized in Table 4. Extended results using DeepLIFT [41] and Integrated Gradients [47] as the explanation methods on ConvNet and various datasets. The results

Table 4: Performance of different explanation methods (Gradient and Guide Propagation) in the training phase is evaluated w.r.t. $\mathcal{L}_e^{\text{start}}$, $\mathcal{L}_e^{\text{end}}$, Acc_{clean} and Acc_{adv} on CIFAR10. Higher $\mathcal{L}_e^{\text{end}}$ indicates better explanation robustness, while higher Acc_{adv} denotes better classification robustness. The **best** and worst performances in explanation robustness and classification robustness are highlighted. Under various explanation methods, $\texttt{SEP}_{pos}$ shows a lower explanation loss compared to $\texttt{SEP}_{neg}$, with similar adversarial accuracy.

Model Architecture	ConvNet				ResNet18			
Gradient								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	7.977	4.591	79.08	0.00	11.310	4.671	81.32	0.00
MAT	13.810	8.705	64.85	**35.11**	26.899	18.215	67.22	29.09
$\texttt{SEP}_{pos}$	0.876	0.503	52.89	29.68	11.317	6.604	66.76	**37.69**
$\texttt{SEP}_{neg}$	13.964	**9.290**	53.23	29.56	8282.990	**7236.182**	49.38	32.28
Guide Propagation								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	8.075	4.639	79.08	0.00	11.515	4.736	81.32	0.00
MAT	14.012	8.813	64.85	35.11	27.012	18.311	67.22	29.09
$\texttt{SEP}_{pos}$	1.023	0.506	60.27	33.57	12.004	7.593	67.16	30.64
$\texttt{SEP}_{neg}$	14.643	**9.110**	59.74	**33.78**	27.422	**18.940**	66.48	**30.72**

using Guide Propagation with TinyImageNet and FMNIST can be found at Table 8. We have the following observations:

• Our methods achieve similar classification robustness under various explanation methods, yet they exhibit notably different explanation robustness. In most cases, $\texttt{SEP}_{pos}$ shows lower explanation loss compared to $\texttt{SEP}_{neg}$, despite similar adversarial accuracy.

• Compared to MAT, our method $\texttt{SEP}_{pos}$ shows comparable adversarial accuracy, indicating similar classification robustness, but it demonstrates distinct explanation loss characteristics. This suggests that explanation robustness and classification robustness may not be strongly correlated.

• Comparing the results in Table 7, Table 4 and Table 8, we can find that no matter which explanation method we use or which dataset we use, we consistently get the conclusion that we could get quite different explanation robustness when we have the similar classification robustness for $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$. This demonstrates the robustness of our results and demonstrates that our findings are not restricted to one specific explanation method or dataset.

4.3.2 Testing Phase Generalization

To test if our findings hold when using different explanation methods during testing, in this experiment, we use the same model trained with Gradient × Input (thus the classification robustness is the same for different testing phases), but change two different explanation methods (Gradient and Guide Propagation) in the testing phase. The results on CIFAR10 are shown in Figure 5, where the detailed value of this experiment can be found in Appendix Table 11. While with the same classification robustness (as shown in Table 3, under adversarial accuracy in CIFAR10), there is a huge difference between $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$ w.r.t the explanation losses (both at the start and the end). This indicates that even with different explanation methods in the testing phase, the explanation robustness still does not show strong correlations with adversarial robustness.

4.4 Training Protocol Generalization (RQ3)

All previous experiments utilized MAT [34] as the default adversarial training protocol. To assess the generalizability of our approach across different adversarial training protocols, we employed TRADES [60] in this experiment and the results can be found in Figure 6 (details in Appendix Table 12). We can observe that when changing the adversarial training method from MAT to TRADES, a larger α will lead to a larger Acc_{adv}, indicating a better classification robustness. As α increases, $\mathcal{L}_e^{\text{end}}$ of the model trained with SEP varies. This observation indicates that our SEP method impacts explanation robustness without altering classification robustness, suggesting a weak correlation between explanation robustness and classification robustness.

4.5 Parameter Sensitivity Analysis

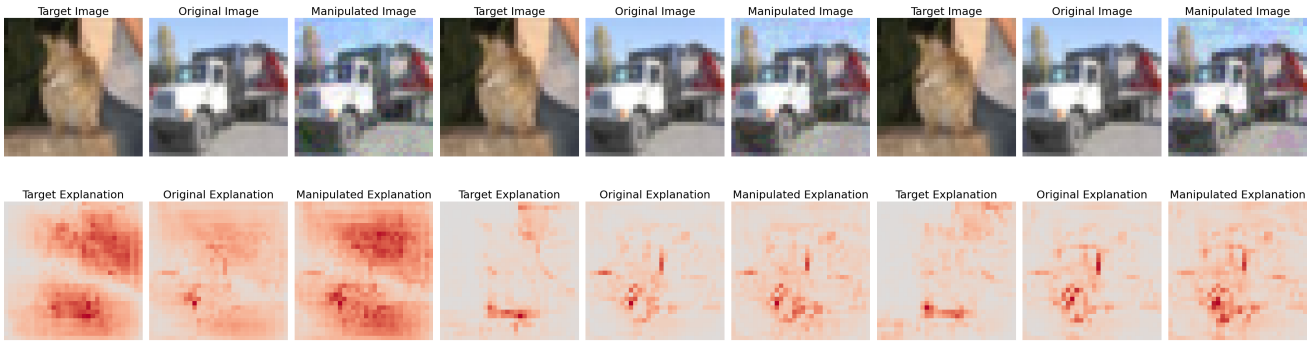
In this section, we examine how different parameters in our experiments affect the results.

Regularization Weights: Firstly, we test how regularization weights λ affect the results. In detail, we trained ConvNet networks on CIFAR10 with various λ values. The test results are presented in Table 5. We observe that the choice of λ influences both the exploration rate at start and end. When λ is greater than 10^4 or less than $-3 * 10^3$, the explanation loss changes intensely. However, from the results, we can see that for different λ from $5{*}10^4$ to $-3{*}10^3$, the classification robustness remains the same, which further validate our hypothesis that explanation robustness and classification robustness may not be highly correlated.

Training Epochs: To demonstrate how training epochs might influence our conclusion, we conducted experiments on the ConvNet network using the CIFAR10 dataset with different training epochs. The results, as presented in Table 6, indicate that the model's performance undergoes only marginal changes after 25 rounds for ConvNet, despite the epoch count continuing to increase. Therefore, we conclude that choosing 25 epochs in the rest experiments does not hurt the reliability of our argument. Besides, the results also support our conclusion. With the increase of training epochs, the classification robustness still increases while the explanation robustness actually decreases, which further validates our conclusion.

4.6 Hessian Analysis

To further analyze how adversarial training and SEP affect the input loss landscape, we perform a Hessian analysis of the loss with respect to the input. Concretely, we study the input Hessian $H_x = \nabla_x^2 L_e(x;\theta)$ with θ fixed after training. The eigenvectors of H_x define principal directions of curvature in input space and the corresponding eigenvalues quantify how rapidly the loss bends along those directions. Large absolute eigenvalues indicate a sharp loss landscape,



(a) $\texttt{SEP}_{pos}$ (b) $\texttt{SEP}_{neg}$ (c) Adversarial Training

Figure 7: Comparison of the saliency maps calculated from gradient x inputs on CIFAR10 with different training methods. Intuitively, $\texttt{SEP}_{pos}$ makes the model consider more input pixels, solely adversarial training makes the model consider only a few input pixels, while $\texttt{SEP}_{neg}$ considers even fewer input pixels compared with adversarial training. However, models trained with these three methods show the same classification robustness.

Table 5: The evaluation of the ConvNet trained on CIFAR10 under different λ conditions reveals that the relationship between explanation and classification robustness is not positively correlated when an appropriate λ is selected during model training.

λ	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
0 (MAT)	16.913	6.206	64.85	35.11
$5 * 10^4$	3.565	1.269	64.94	35.25
10^4	15.436	5.870	64.39	35.18
10^1	17.646	6.819	64.45	35.02
-10^2	17.820	6.934	64.67	35.14
$-3 * 10^3$	19.002	7.590	64.56	34.86

Table 6: Test results of the ConvNet model at different training epochs on the CIFAR10 dataset. As the number of training epochs increases beyond 25, the improvement in performance is marginal. Therefore, 25 epochs are selected as the final number of training epochs for all models to maintain faster training speed without affecting the overall conclusions.

Model Architecture	ConvNet (CIFAR10)			
Training Epoch	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
25	4.388	1.605	64.94	35.25
50	3.885	1.431	65.69	35.94
75	3.671	1.378	66.33	36.27
100	3.557	1.339	66.74	36.50

while small absolute values indicate flatness of the loss landscape. Following previous works [31], we compute the top 20 eigenvalues at evaluation inputs and report their dataset means for the Hessian of explanation loss, which captures both how steep the sharpest direction is and whether sensitivity is concentrated in a few directions or spread across many, as reflected by the spectral decay. In detail, we show the results in Figure 8. The results show that $\texttt{SEP}_{pos}$ has much lower eigenvalues and thus indicates a much flatter loss landscape. Similarly, $\texttt{SEP}_{neg}$ has higher eigenvalues. These results firstly show that the design of SEP is successful because $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$ influence the loss landscape as we want. Secondly, the results of flatness of loss landscape of $\texttt{SEP}_{pos}$ and $\texttt{SEP}_{neg}$ further validate the assumption in the previous section that the explanation robustness is mainly from the high initial loss instead of flat loss landscape, which indicates that there is no internal relationship between classification robustness and explanation robustness.

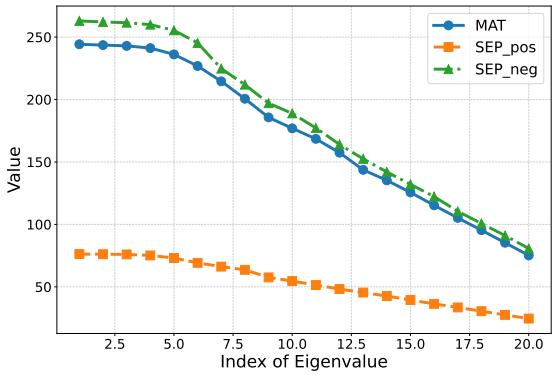


Figure 8: Top-20 Eigenvalue of Hessian Matrix considering the explanation loss. A higher value of eigenvalue indicates a shaper loss landscape. The results show that SEP influence the loss landscape.

4.7 Case Study

In this section, we visualize the comparison of saliency maps from models trained with different algorithms to provide how our methods influence the saliency maps in Figure 7. Specifically, we use the Gradient × Input method [41] to calculate saliency maps on the CIFAR-10 dataset [26] with ConvNet. From Figure 7 we can see that while models trained with $\texttt{SEP}_{pos}$ exhibit broader activation regions in their saliency maps, models trained with standard adversarial training focus on fewer input pixels compared to $\texttt{SEP}_{pos}$. And models trained with $\texttt{SEP}_{neg}$ exhibit the narrowest focus, considering even fewer input pixels than adversarial training. Firstly, these findings highlight the flexibility of our method in shaping explanation patterns and suggest that explanation robustness can be independently controlled through targeted regularization techniques. Secondly, the results show that adjusting the loss landscape can lead to totally different explanation patterns. From the explanation patterns, we can see that if the activated pixels are fewer, the initial distance, which is measured by the $\mathcal{L}_e^{\text{start}}$, will also be

Table 7: Performance of using DeepLift and Integrated Gradients as explanation methods with ConvNet. Higher $\mathcal{L}_e^{\text{end}}$ indicates better explanation robustness, while higher Acc_{adv} denotes better classification robustness. The **best** and worst performances in explanation robustness and classification robustness are highlighted. Under various explanation methods, SEP_{pos} shows a lower explanation loss compared to SEP_{neg}, with similar adversarial accuracy.

Explanation Method	DeepLift				Integrated Gradients			
MNIST								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
MAT	369.153	294.053	99.00	89.92	239.650	224.745	99.00	89.92
SEP_{pos}	82.959	57.408	98.93	95.97	76.284	50.603	98.42	**93.19**
SEP_{neg}	1101.038	**896.157**	98.97	**96.16**	778.663	**534.113**	98.68	92.76
FMNIST								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
MAT	386.377	274.130	62.85	**73.98**	237.727	234.109	62.85	**73.98**
SEP_{pos}	33.824	21.429	60.75	65.52	21.425	16.048	60.85	72.05
SEP_{neg}	4739.769	**3153.331**	60.62	70.05	3624.231	**2748.976**	62.92	70.36
CIFAR10								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
MAT	18.137	11.562	64.85	35.11	16.887	14.007	64.85	**35.11**
SEP_{pos}	2.593	1.273	65.76	**35.38**	3.696	2.523	60.19	32.33
SEP_{neg}	21.329	**13.819**	64.20	34.91	19.568	**14.803**	60.86	32.43
CIFAR100								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
MAT	19.683	12.766	36.40	17.35	16.754	10.779	36.40	**17.35**
SEP_{pos}	14.208	9.201	39.10	18.23	5.320	3.218	34.73	16.74
SEP_{neg}	20.389	**13.694**	39.78	**18.32**	17.011	**11.356**	35.10	16.60

Table 8: Performance of using Guide Propagation in the training phase with Fashion-MNIST and TinyImageNet. Higher $\mathcal{L}_e^{\text{end}}$ indicates better explanation robustness, while higher Acc_{adv} denotes better classification robustness. The **best** and worst performances in explanation robustness and classification robustness are highlighted. Under various explanation methods, SEP_{pos} shows a lower explanation loss compared to SEP_{neg}, with similar adversarial accuracy.

Model Architecture	ConvNet				ResNet18			
Fashion-MNIST (FMNIST)								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	30.932	18.451	92.79	0.00	66.131	29.110	91.57	0.00
MAT	97.726	72.402	62.85	73.98	608.486	467.815	79.22	67.10
SEP_{pos}	48.368	34.672	78.46	67.28	97.703	78.354	77.55	62.09
SEP_{neg}	542.540	**425.948**	65.07	**77.17**	4219.351	**3839.408**	80.11	**72.21**
TinyImageNet								
Training Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
Normal	0.559	0.281	28.71	0.00	0.617	0.216	28.34	0.00
MAT	1.356	0.787	25.13	9.55	2.577	1.411	26.33	10.81
SEP_{pos}	0.983	0.625	25.16	5.97	1.767	1.226	28.68	11.47
SEP_{neg}	1.566	**0.977**	24.89	4.99	3.403	**1.761**	26.79	11.23

higher. Though in this case, the loss decrease is higher, the large initial distance ensures a better explanation robustness. On the other hand, if there are more activated pixels, it is more likely that two explanations are close at the beginning. Even though $\texttt{SEP}_{pos}$ has a flat loss landscape, which makes the attacker harder to optimize, the final attack is still successful. Based on the case, we further validate that the mechanism for explanation robustness is not based on the flat loss landscape, indicating there is no strongly correlated relationship between the two robustness.

5. CONCLUSION

This study challenges the widely held assumption that explanation robustness and classification robustness are strongly correlated. By systematically control classification robustness, we demonstrate that increasing explanation robustness does not necessarily result in a flatter input loss landscape for explanation loss. This finding contrasts with the well-established observation that enhancing classification robustness leads to a flatter input loss landscape for classification loss. These results reveal a fundamental difference in how these two types of robustness are influenced by adversarial training. To address this discrepancy, we propose SEP, a novel algorithm that explicitly flattens the input loss landscape for explanation loss. Our experiments show that this approach effectively improves explanation robustness without affecting classification robustness, providing evidence that these two forms of robustness are not inherently linked. This decoupling highlights the importance of separately considering and optimizing explanation and classification robustness to ensure the reliability and trustworthiness of AI systems, particularly in high-stakes domains such as healthcare, autonomous driving, and finance.

Our findings emphasize the need for future research to further explore the relationship between these two types of robustness. In the future, we will theoretically investigate why adversarial training can improve explanation robustness in certain cases and what underlying mechanisms differentiate the behavior of classification and explanation robustness under adversarial attacks. A deeper understanding of these mechanisms will provide valuable insights into designing more robust and interpretable AI systems capable of maintaining both predictive accuracy and trustworthy explanations under adversarial conditions.

Acknowledgment

The work was partially supported by NSF award #2442477. We thank Amazon Research Awards, Cisco Research Awards, Google, and OpenAI for providing us with API credits. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[2] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[3] 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*, 10(7):e0130140, 2015.

[4] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[5] A. Boopathy, S. Liu, G. Zhang, C. Liu, P.-Y. Chen, S. Chang, and L. Daniel. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pages 1014–1023. PMLR, 2020.

[6] W. Brendel, J. Rauber, A. Kurakin, N. Papernot, B. Veliqi, S. P. Mohanty, F. Laurent, M. Salathé, M. Bethge, Y. Yu, et al. Adversarial vision challenge. In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 129–153. Springer, 2020.

[7] N. Carlini, F. Tramer, K. D. Dvijotham, L. Rice, M. Sun, and J. Z. Kolter. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.

[8] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[9] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

[10] J. Chen, X. Wu, V. Rastogi, Y. Liang, and S. Jha. Robust attribution regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] K. Chen, Y. Chen, H. Zhou, X. Mao, Y. Li, Y. He, H. Xue, W. Zhang, and N. Yu. Self-supervised adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2218–2222. IEEE, 2020.

[12] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2020.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

[15] A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[16] S. Gidaris and N. Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019.

[20] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in neural information processing systems*, 32, 2019.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[22] W. Huang, X. Zhao, G. Jin, and X. Huang. Safari: Versatile and efficient evaluations for robustness of interpretability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1988–1998, 2023.

[23] G. Jin, X. Yi, D. Wu, R. Mu, and X. Huang. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16447–16457, 2023.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[26] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[29] L. Li and M. Spratling. Understanding and combating robust overfitting via input loss landscape analysis and regularization. *Pattern Recognition*, 136:109229, 2023.

[30] J. Lin, C. Gan, and S. Han. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

[31] C. Liu, M. Salzmann, T. Lin, R. Tomioka, and S. Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems*, 33:21476–21487, 2020.

[32] G. Liu, I. Khalil, and A. Khreishah. Using single-step adversarial training to defend iterative adversarial examples. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 17–27, 2021.

[33] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[35] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*, 2019.

[36] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.

[37] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[39] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[40] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

[41] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[42] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[43] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arxiv 2013. *arXiv preprint arXiv:1312.6034*, 2019.

[44] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[45] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

[46] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[47] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[48] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[49] S. V. Tamam, R. Lapid, and M. Sipper. Foiling explanations in deep neural networks. *arXiv preprint arXiv:2211.14860*, 2022.

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

[51] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[52] Z. Wang, M. Fredrikson, and A. Datta. Robust models are more interpretable because attributions look normal. *arXiv preprint arXiv:2103.11257*, 2021.

[53] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*, 2023.

[54] M. Wicker, J. Heo, L. Costabello, and A. Weller. Robust explanation constraints for neural networks. *arXiv preprint arXiv:2212.08507*, 2022.

[55] D. Wu, S.-T. Xia, and Y. Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[56] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[57] C. Xie, M. Tan, B. Gong, A. Yuille, and Q. V. Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

[58] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[59] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[60] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[61] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

APPENDIX

A. CODE AND DATA

This is our open source code link: open source code.
We evaluate five explanation methods from Captum [25]:

- Gradient [4]: Raw input gradients
- Gradient×Input [41]: Element-wise product of inputs and gradients
- Guided Backprop [46]: Filtered gradient visualization
- DeepLIFT [41]: Reference-based difference attribution
- Integrated Gradients [47]: Path integral of gradients

B. MORE VISUALIZATION RESULTS

Firstly, we visualize the input loss landscape w.r.t explanation loss using a normal trained model and model trained with Madry adversarial training in Figure 9. The results show that increasing the explanation robustness does not flatten the input loss landscape. Besides, we also visualize

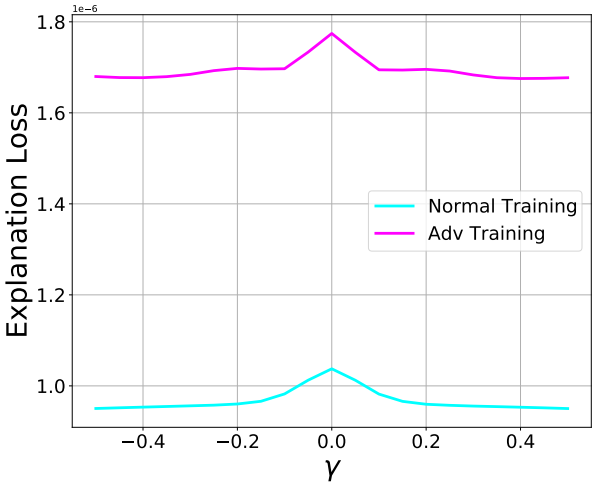


Figure 9: Comparison of input loss landscape w.r.t explanation loss with adversarial training and normal training. The results show that there is no obvious difference in input loss landscape.

more saliency maps with more explanation methods with images from different clusters in Figure 10. They all prove that we can choose the most representative saliency maps.

C. DETAILED HYPERPARAMETER

In this section, we provide the detailed hyperparameter for our CIFAR10 dataset in Table 10.

Table 10: Comparison of explanation loss for intra-cluster sample and inter-cluster sample on CIFAR10. The results show that our cluster method indeed the cluster images with similar explanations.

Models	Learning Rate	λ
SEP_pos		
ConvNet	0.01	5e4
ResNet18	0.001	5e4
Wide ResNet	0.001	5e4
MobileNet	0.01	5e4
SEP_neg		
ConvNet	0.01	-3e3
ResNet18	0.001	-1.9e3
Wide ResNet	0.001	-1.9e3
MobileNet	0.01	-1.25e3

D. MORE EXPERIMENTAL RESULTS

D.1 Experiments on W-ResNet and MobileNet

We list the main results using Gradient X Inputs as training and testing explanation methods for W-ResNet and MobileNetV2 in Table 9. We have the following observations:

- Once again, the adversarial accuracy for MAT, SEP_{pos}, and SEP_{neg} is similar in most scenarios for W-ResNet and MobileNet, while SEP_{pos} always has a smaller explanation loss compared with MAT, and SEP_{neg} always has a larger explanation loss compared with MAT. These results show that influencing explanation robustness does not necessarily change classification robustness.
- For W-ResNet and MobileNet, the adversarial accuracy for CIFAR100 fluctuates. For MobileNet and CIFAR100, compared with MAT, SEP_{pos} increases classification robustness while SEP_{neg} decreases it. However, this observation also indicates that the positive correlation between explanation robustness and classification robustness might not be true since SEP_{pos} decreases explanation robustness while increasing classification robustness.

D.2 Detailed Values for Transferablity Experiments

The detailed values for Transferablity experiments can be found in Table 11 and the detailed values for experiments using TRADES for our method can be found in Table 12. The analysis of these results can be found in the main paper.

D.3 Experiments on ImageNet

Here, we present our experiments on ImageNet with ResNet18 in Table 13. We can find that the conclusion of ImageNet experiments is the same as the main paper: Increasing or decreasing explanation robustness will not necessarily influence the classification robustness.

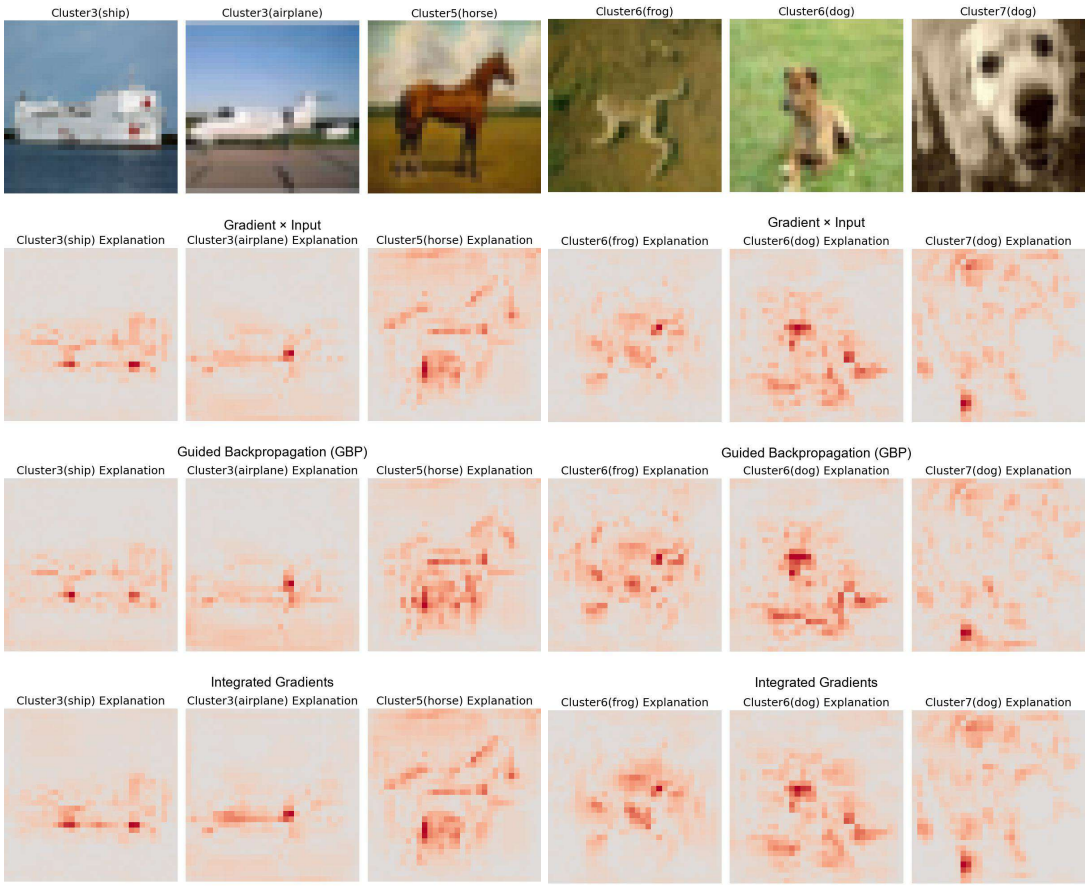


Figure 10: Explanation of images from different clusters. The results show that images from the same cluster that even have different labels still have similar saliency maps on various explanation methods. Besides, images with the same label from different clusters still have different explanations. These results show that our method can sample the most representative subset of explanations.

Table 9: Test results of models trained by Wide ResNet network and MobileNet network on various data sets according to four training methods. The results presented indicate that the performance of models trained using the Wide ResNet network and MobileNet network on different datasets suggests that there is no positive correlation between the model's explanation robustness and classification robustness achieved through the SEP_{pos} and SEP_{neg} training methods, as compared to the MAT training method.

	Wide ResNet				MobileNet			
MNIST								
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Adv Acc	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{adv} (%)	Acc_{adv} (%)
Normal	267.050	206.194	99.58	0.00	287.061	188.700	99.08	0.02
MAT	842.648	736.839	98.92	**82.82**	4328.176	3356.135	98.29	94.19
SEP_pos	109.383	99.891	99.01	82.77	319.629	273.256	98.36	**94.25**
SEP_neg	937.845	**744.698**	98.87	82.71	8134.157	**4454.656**	98.33	94.23
FMNIST								
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{adv} (%)	Acc_{adv} (%)
Normal	120.037	69.593	92.79	0.00	180.159	103.941	91.93	0
MAT	328.817	257.523	78.10	**68.26**	4470.448	3571.210	68.72	57.19
SEP_pos	109.996	74.324	77.69	67.79	236.547	172.200	65.11	57.42
SEP_neg	398.006	**304.927**	78.21	68.05	6032.190	**4809.288**	66.86	**58.16**
CIFAR10								
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{adv} (%)	Acc_{adv} (%)
Normal	17.920	8.029	85.47	0.16	14.797	6.551	77.48	0
MAT	41.513	27.136	60.01	24.22	21.502	13.223	51.51	**23.81**
SEP_pos	26.343	16.217	59.87	24.89	14.756	7.907	49.91	23.27
SEP_neg	43.278	**27.575**	60.15	25.08	26.811	**16.420**	35.43	15.30
CIFAR100								
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{adv} (%)	Acc_{adv} (%)
Normal	13.677	5.606	59.13	0	17.015	9.351	43.91	0
MAT	30.027	18.389	36.69	**16.12**	20.054	10.836	21.19	8.64
SEP_pos	22.046	13.704	33.88	13.19	15.234	8.510	21.82	**10.05**
SEP_neg	31.889	**20.045**	35.74	15.55	21.544	**13.843**	21.35	7.88

Table 11: Test results for transferability of explanation robustness. Models are trained with Gradient x Input and tested on different explanation methods.All models are trained on CIFAR10. Even if the interpretation methods during training and testing are different, comparing the training results of our proposed method with the AT training method of the corresponding configuration in Table 3, we can still draw our previous conclusions, which also shows that our conclusions are transferable.

	ConvNet		ResNet18	
Train:Gradient X Input, Test:Gradient				
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)
SEP_{pos}	3.054	1.901	9.555	5.903
SEP_{neg}	15.093	9.513	55.526	33.176
Train:Gradient X Input, Test:Integrated_Grad				
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)
SEP_{pos}	3.767	2.404	9.209	6.720
SEP_{neg}	17.066	10.923	58.730	38.433

	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{adv} (%)
Normal	114.70	63.52	0.00
AT	1281.71	742.43	19.36
SEP_{pos}	287.64	156.16	17.63
SEP_{neg}	1427.33	905.25	17.44

Table 13: Experiments for ImageNet on ResNet18. The results are aligned with the conclusion made in the main paper.

Table 12: The test results of the model trained using the TRADE training method, combined with our approach. The findings indicate that when we apply our method to TRADE, an alternative adversarial training method distinct from MAT, we can still deduce that classification robustness and interpretation robustness are not inherently interconnected.

ConvNet				
CIFAR10, TRADE Weight:5				
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
TRADE	18.278	11.470	64.5	33.98
TRADE + SEP_pos	3.878	2.285	63.84	33.85
TRADE + SEP_neg	19.781	12.424	64.37	34.07
ConvNet				
CIFAR10, TRADE Weight:1				
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
TRADE	17.271	10.965	72.63	28.31
TRADE + SEP_pos	4.089	2.296	72.41	28.20
TRADE + SEP_neg	18.504	11.662	72.90	28.34
ResNet18				
CIFAR10, TRADE Weight:5				
Method	$\mathcal{L}_e^{\text{start}}$ ($\times 10^{-7}$)	$\mathcal{L}_e^{\text{end}}$ ($\times 10^{-7}$)	Acc_{clean} (%)	Acc_{adv} (%)
TRADE	18.278	11.469	64.50	33.98
TRADE + SEP_pos	12.232	7.527	63.49	34.93
TRADE + SEP_neg	22.571	14.881	63.42	33.30