

adversarial attacks. This work designs an attack mechanism to find smaller yet viable adversarial perturbations.

Besides malicious adversarial perturbations, DNNs might also encounter *natural perturbations* measured by attributes in the real world. For example, object-level transforms, such as the degree to which a person smiles, or geometric transforms, such as the rotation of images, that commonly appear in the real world are not accounted for by adversarial perturbations. Natural perturbations along these attributes preserve the semantic classification information and are thus *semantic-preserving*. For instance, changing the extent of smiling of a person in a gender classification task, or changing the rotation angle of a digit in a digit classification task, generates a natural-looking image and will not lead to a change of the true class label. Yet, authors in [14], [15] find that intentional perturbations along these attributes are also likely to cause model performance to decline. Although such performance decline is much smaller compared to that of adversarial perturbations, natural perturbations can be used for attacks. Moreover, it is shown that the robustness against natural perturbations is independent of adversarial robustness [16]. That is, classifiers trained with only adversarial samples are not robust against natural perturbations, and vice versa. Therefore, we aim to train a target model that is robust against both types of perturbations.

In this work, we address the robustness against both perturbations by proposing a novel generator-based adversary named Semantic-Preserving Adversarial (SPA) attack that generates *jointly-perturbed* samples. Figure 1 illustrates the idea of the proposed attack. SPA attack maximizes the exposure of the target model to variations in both attribute space and adversarial space. The attack framework consists of an attribute manipulator for natural perturbations and an adversarial noise generator for diverse adversarial perturbations. By modifying the class-irrelevant attributes of the images, SPA attack searches semantic-preserving samples that are more vulnerable to adversarial attacks. Then, SPA attack finds valid adversarial noises under stringent l_p norm-ball constraints by exploring the adversarial diversity variable. Based on the proposed attack mechanism, we further design a robust training approach, which addresses a min-max optimization and adversarially trains the target model against joint perturbations. Empirical studies in Section V verify the effectiveness of our SPA attack, and demonstrate that our SPA training can provide superior protection against joint attacks compared to previous methods.

The major contributions are summarized as follows:

- We present a novel attack mechanism named Semantic-Preserving Adversarial (SPA) attack that considers the problem of robustness against joint perturbations in the pixel space as well as a set of specified attributes in the attribute space.
- We propose SPA training as robust training by solving a min-max optimization problem and jointly exploring the pixel space and the attribute space in novel ways without access to the test domain.

- We introduce two surrogate functions for attribute manipulation on two classical semantic-preserving natural perturbations: geometric transformations and objective-level transformations.
- We empirically demonstrate the effectiveness of our proposed approach on four public datasets.

The rest of this paper is organized as follows. We briefly review the related work in Section II and provide the definition of this problem in Section III. Then, we introduce our approach for attack and defense in Section IV. Finally, we discuss the experiment results on MNIST, FashionMNIST, CelebA, and SICAPv2 datasets in Section V.

II. RELATED WORK

In this section, we investigate the related work on adversarial attacks and training, attribute robust training, and attribute manipulation.

A. Adversarial Attacks and Training

As a pioneering work, Szegedy *et al.* [17] initially observed that deep learning models are vulnerable against imperceptible *adversarial perturbations*. While adversarial training could be formulated as a min-max optimization mathematically, it is a huge challenge to solve the inner maximization (e.g., the generation of adversarial examples). Goodfellow *et al.* [4] proposed Fast Gradient Sign Method (FGSM) to generate adversarial examples with the sign of gradient in a one-step manner. However, the model trained with FGSM-generated adversarial examples suffers from catastrophic overfitting [18], [19]. For further enhancing the strength of attacks, an iterative FGSM variant Basic Iterative Method (BIM), was proposed in [5]. And CW attack [20] then took a direct optimization approach to find adversarial samples, which broke the distillation knowledge defense for the first time. Later, Madry *et al.* [6] proposed PGD attack, which is considered to be one of the most powerful first-order attacks for approximating the optimal value of the inner maximization problem.

More recently, approaches based on learning-to-learn (L2L) [21]–[23] are proposed to improve adversarial training. L2LDA [24] further introduced a diversity variable to generate diverse adversarial noises and update the adversarial noise recursively. Their adversarially trained ResNet was shown to outperform the ResNets trained by CW, L2L and PGD on CIFAR-10. Such adversarial attacks perturb the pixel space under l_p norm-ball constraints, yet may fail when the constraints get much more stringent. On the contrary, we approach the problem of generating jointly-perturbed samples with smaller adversarial perturbations. Following the literature, we use several state-of-the-art attacks from CW [20] and L2LDA [24] as our benchmarks.

B. Attribute Robust Training

Recently, there has been an increasing interest in the robustness against *natural perturbations* in terms of attributes, which are perceptible shifts in the data but still are natural-looking, and enough to fool a classifier [25]. Liu *et al.* [26]

proposed to perturb physical parameters that underlay image formation to produce natural perturbations sensitive to physical concepts like lighting and geometry. [14] proposed to generate natural perturbations by modifying multiple specified attributes with a conditional generative model. [15] proposed to perturb the attribute space to synthesize new images and maximize the exposure of the classifier to the attributes space. However, such attacks only cause *limited* performance decline and do not result in severe model failure as pixel-level adversarial attacks do [15], especially when few attributes are valid. Contrary to these approaches, we address the robustness against both perturbations by considering a novel and powerful attack that jointly leverages attribute perturbations and adversarial perturbations.

C. Attribute Manipulation

There are various approaches for manipulating attributes in images. Conditional Generative Adversarial Networks [27] use an encoder-decoder architecture to learn attribute invariant latent representation and disentangle attributes for attribute editing. Matrix Subspace Projection (MSP) [28] is proposed to factorize the latent space for attribute manipulation effectively. Spatial Transformer Networks (STN) [29] uses a localization net to conduct precise spatial transformation.

III. PROBLEM FORMULATION

We begin by defining a classifier parameterized by θ as $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, where \mathcal{X} denotes the space of the image data and \mathcal{Y} denotes the label space. Labeled data $(\mathbf{x}, \boldsymbol{\alpha}, \mathbf{y})$, where \mathbf{x} denotes the input image, $\boldsymbol{\alpha}$ denotes the annotated attribute value, \mathbf{y} denotes the class label. Table I summarizes the notions used in this paper.

The objective of this paper is to build a classifier f_θ that is robust to both 1) *adversarial* perturbations, which are l_∞ -bounded in the pixel space, and 2) *natural* perturbations along specific attributes $\boldsymbol{\alpha}$ that is specified *a priori*.

We categorize semantic-preserving natural perturbations into two major types - (a) **Geometric transformations**, where images are manipulated by affine transformations such as rotation, scaling, and shifting. (b) **Object-level transformations**, where general attributes of the objects in images are manipulated without changing the semantic information of the class labels, such as changing the extent of smiling or hair color of a person in a gender classification task. Geometric transformations have explicit generative mechanisms, allowing us to conduct accurate natural perturbations. However, we do not have access to the true generative mechanisms for objective-level transformations. Therefore, we need to approximate the objective-level transformations by training a conditional generative model.

The classification loss of the target classifier f_θ can be defined as a cross-entropy loss:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}; \theta) = \mathbf{y}^T \log f(\mathbf{x}; \theta), \quad (1)$$

where $f(\mathbf{x}; \theta)$ outputs the predicted probabilities for \mathbf{x} . We simplify the loss as $\mathcal{L}(\mathbf{x})$ without causing confusion.

TABLE I: Notations

Notations	Descriptions
\mathbf{x}	The input features
\mathbf{y}	The labels
$\boldsymbol{\alpha}$	The attribute values
\mathbf{z}	The diversity variable of adversarial noises
f_θ	The classifier parameterized by θ
g_ϕ	The adversarial noise generator parameterized by ϕ
h_ψ	The attribute manipulator parameterized by ψ
\mathcal{L}	The loss function
ϵ	The l_∞ norm of adversarial noise

This task can be formulated as a min-max optimization of a saddle point problem:

$$\min_{\theta \in \Theta} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \left[\max_{\tilde{\mathbf{x}} \in \mathcal{X}_{spa}(\mathbf{x})} \mathcal{L}(\tilde{\mathbf{x}}, \mathbf{y}; \theta) \right], \quad (2)$$

where $\mathcal{X}_{spa}(\mathbf{x})$ is a set of admissible attacks of \mathbf{x} with natural perturbations in the attribute space along the specified attributes $\boldsymbol{\alpha}$ as well as l_p -bounded adversarial perturbations in the pixel space.

Eq. 2 solves a min-max optimization problem. The inner maximization aims to generate jointly-perturbed samples with both perturbations that maximize the classification loss, and the outer minimization corresponds to finding model parameters that minimize the loss on jointly-perturbed samples found by the inner maximization. The success of finding the optimal θ^* that minimizes Eq. 2 crucially relies on solving the (non-concave) inner optimization problem. In this work, we focus on the inner maximization problem of Eq. 2 to achieve robustness against both perturbations. To conduct attacks, we propose an architecture that generates and effectively optimizes diverse natural and adversarial perturbations.

IV. PROPOSED APPROACH

Conceptually, generating jointly-perturbed samples with both natural and adversarial perturbations of an image can be broken into two sub-problems: (a) effective navigation on the attribute space and the l_p -bounded pixel space, and (b) joint optimization of attribute values and adversarial noises over both attribute and pixel spaces. We address each problem in detail below.

A. Joint Perturbation Models

First, let us consider the problem of generating *jointly-perturbed* samples with both natural and adversarial perturbations, via navigating on the attribute space and the l_p -bounded pixel space. We design a sequential architecture to perform joint perturbation, as shown in Figure 2.

Attribute Manipulation. We adopt a differentiable attribute manipulator $h_\psi : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{X}$, conditioned on a semantic attribute space \mathcal{A} , to conduct attribute manipulation. h_ψ takes an image \mathbf{x} and an attribute vector $\boldsymbol{\alpha}$ as input, and generates perturbed image \mathbf{x}_{attr} with the specified attribute. h_ψ is trained with the annotated training data $(\mathbf{x}, \boldsymbol{\alpha})$.

For objective-level transformations, we leverage MSP [28] to build h_ψ . An encoder network encodes the input image

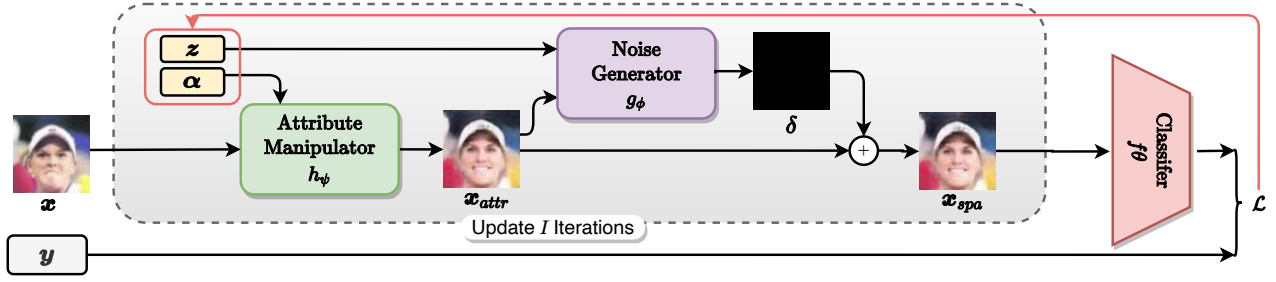


Fig. 2: An illustration of the Semantic-Preserving Adversarial (SPA) attack. The attribute manipulator h_ψ and the noise generator g_ϕ sequentially conduct natural perturbations and adversarial perturbations. The attribute value α (e.g., the extent of smiling) and adversarial diversity variable z are jointly and iteratively optimized to generate jointly-perturbed samples.

into attribute-invariant features and attribute-related features. Matrix subspace projection modifies the attributes of interest to specified values α . Then, a decoder network decodes the modified features into the desired image with specified attributes. For geometric transformations, the attribute manipulator is built with STN [29]. The input images are first normalized by STN with attributes predicted by an attribute predictor, which is pre-trained with the annotated attributes in training data. Then an STN is used to conduct affine transformation on the normalized images to specified attributes.

Adversarial Noise Generation. To explore novel and hard adversarial samples, we explicitly train an adversarial noise generator $g_\phi : X \times \mathcal{Z} \rightarrow \mathcal{S}$ to generate diverse adversarial noises $\delta \in \mathcal{S}$, conditioned on a diversity variable $z \in \mathcal{Z}$, where \mathcal{S} the l_p -ball with ϵ size and \mathcal{Z} the diversity variable space. For a stronger attack, the adversarial noise generator g_ϕ can be further enhanced by adding $\nabla_x \mathcal{L}(x)$ the gradient of the image, y the class label, and δ the noise of the image as input. Therefore, the adversarial noise can be recursively generated by:

$$\tilde{x}^{(t+1)} \leftarrow \text{Proj}_{\mathcal{S} \cup \mathcal{X}}(x^{(t)} + \epsilon_{\text{step}} g_\phi(x, z; y, \delta^{(t)}, \nabla_x \mathcal{L}(\tilde{x}^{(t)}))), \quad (3)$$

where $\text{Proj}_{\mathcal{S} \cup \mathcal{X}}(\cdot)$ denotes the projection of its element to l_p -ball \mathcal{S} and a valid pixel value range, $\delta^{(t)}$ is the noise accumulated up to t -th step, and ϵ_{step} denotes the step size smaller than ϵ . We adopt a diversity loss to encourage generating diverse adversarial noises [24], [30]:

$$\mathcal{L}_{\text{div}} = \frac{1}{T} \sum_{t=1}^T \frac{\|\tilde{x}^{(t)}(z_1) - \tilde{x}^{(t)}(z_2)\|_1}{\|z_1 - z_2\|_1}, \quad (4)$$

where $\tilde{x}^{(t)}(z)$ denotes the adversarial samples generated by Eq. 3 with z , and z_1, z_2 are two i.i.d. samples of z .

Given the above models, we define a semantic-preserving adversarial attack as the process of transforming an input image x via attribute perturbation $x_\alpha = h_\psi(x, \alpha)$ and adversarial perturbation $\delta = g_\phi(x_\alpha, z)$ to produce a new sample $\tilde{x}_\alpha = x_\alpha + \delta$, such that $f(\tilde{x}_\alpha) \neq y$. Therefore, we reformulate Eq. 2 as an optimization of the following problem:

$$\begin{aligned} \min_{\theta \in \Theta} \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[\max_{\alpha \in \mathcal{A}, z \in \mathcal{Z}} \mathcal{L}(x_{spa}, y; \theta) \right] \\ \text{s.t. } x_{spa} = \text{Proj}_{\mathcal{S} \cup \mathcal{X}}(h_\psi(x, \alpha) + \epsilon g_\phi(h_\psi(x, \alpha), z)), \end{aligned} \quad (5)$$

Algorithm 1 Semantic-Preserving Adversarial (SPA) Attack

Input: Image x , ground-truth label y , pretrained attribute manipulator and noise generator h_ψ, g_ϕ .

Output: SPA sample x_{spa} .

- 1: Initialize $\alpha^{(0)}, z^{(0)}$
- 2: **for** i in range $[0, I]$ **do**
- 3: $x_{attr}^{(i)} \leftarrow h_\psi(x, \alpha^{(i)})$.
- 4: **for** t in range $[0, T]$ **do**
- 5: $\tilde{x}_{attr}^{(i, t+1)} \leftarrow \text{Proj}_{\mathcal{S} \cup \mathcal{X}}(x_{attr}^{(i)} + \epsilon_{\text{step}} g_\phi(x_{attr}^{(i)}, z^{(i)}; y, \delta^{(t)}, \nabla_x \mathcal{L}(\tilde{x}_{attr}^{(i, t)})))$
- 6: **end for**
- 7: $\tilde{x}_{attr}^{(i+1)} \leftarrow \tilde{x}_{attr}^{(i, T)}$
- 8: $\alpha^{(i+1)} \leftarrow \alpha^{(i)} + \nabla_\alpha \mathcal{L}(\tilde{x}_{attr}^{(i+1)})$
- 9: $z^{(i+1)} \leftarrow z^{(i)} + \nabla_z \mathcal{L}(\tilde{x}_{attr}^{(i+1)})$
- 10: **end for**
- 11: $x_{spa} \leftarrow \text{Repeat Step 3-6 with } \alpha = \alpha^{(I)}, z = z^{(I)}.$

B. Iterative Parameter Optimization

Having access to the attribute manipulator and the noise generator, we focus on solving the inner optimization in Eq. 5. Note that the success of robust training relies on generating strong perturbations in terms of attributes and adversarial noises. Algorithm 1 demonstrates the optimization of the attribute value and diversity variable for finding SPA samples x_{spa} . Step 3-6 conducts attribute manipulation and adversarial noise generation sequentially. Then, we project the adversarial loss onto the attribute space and the diversity variable space, by cascading the output of the attribute manipulator, noise generator, and target classifier. The optimization is conducted by back-propagation over the classifier f_θ and perturbation models h_ψ, g_ϕ .

Algorithm 1 efficiently explores a larger parameter space and generates stronger jointly-perturbed samples, compared to previous methods [15], [24], especially under stringent constraints with small l_p -ball and few valid semantic-preserving attributes.

C. Semantic-Preserving Adversarial Training

With our proposed SPA attack, we solve the outer optimization problem in Equation 5 by proposing a unified training framework for the robustness against natural and adversarial

perturbations, by jointly optimizing the noise generator g_ϕ and the target classifier f_θ . Before SPA training, The attribute manipulator h_ψ is pretrained with the annotated training data $(\mathbf{x}, \alpha_{gt})$, where α_{gt} the ground-truth attributes.

Algorithm 2 Semantic-Preserving Adversarial Training

Input: Image \mathbf{x} , and label \mathbf{y} , pretrained attribute manipulator h_ψ .

- 1: $\mathbf{x}_{spa} \leftarrow \text{SPA_Attack}(\mathbf{x}, \mathbf{y}, h_\psi, g_\phi)$
 - 2: Randomly initialize $\alpha, \mathbf{z}_1, \mathbf{z}_2$
 - 3: $\mathbf{x} \leftarrow [\mathbf{x}; \mathbf{x}], \alpha \leftarrow [\alpha; \alpha], \mathbf{y} \leftarrow [\mathbf{y}; \mathbf{y}], \mathbf{z} \leftarrow [\mathbf{z}_1; \mathbf{z}_2]$ ($[\cdot; \cdot]$: concatenation)
 - 4: $\mathbf{x}_{attr} \leftarrow h_\psi(\mathbf{x}, \alpha)$
 - 5: **for** t in range $[0, T)$ **do**
 - 6: $\tilde{\mathbf{x}}_{attr}^{(t+1)} \leftarrow \text{Proj}_{\mathcal{S} \cup \mathcal{X}}(\mathbf{x}_{attr} + \epsilon_{step} g_\phi(\mathbf{x}_{attr}, \mathbf{z}; \mathbf{y}, \delta^{(t)}, \nabla_{\mathbf{x}} \mathcal{L}(\tilde{\mathbf{x}}_{attr}^{(t)})))$
 - 7: **end for**
 - 8: Compute \mathcal{L}_{div} with $\tilde{\mathbf{x}}_{attr}^{(t)}, \mathbf{z}$ following Eq. 4
 - 9: $\mathcal{L}_{cls} \leftarrow \frac{1}{T} \sum_1^T \mathcal{L}(\tilde{\mathbf{x}}_{attr}^{(t)}) + \mathcal{L}(\mathbf{x}_{spa})$
 - 10: $\phi \leftarrow \phi + \nabla_\phi(\mathcal{L}_{cls} + \lambda \mathcal{L}_{div})$
 - 11: $\theta \leftarrow \theta - \nabla_\theta(\mathcal{L}(\mathbf{x}) + \mathcal{L}(\tilde{\mathbf{x}}_{attr}^{(T)}) + \mathcal{L}(\mathbf{x}_{spa}))$
-

The overall procedure is summarized in Algorithm 2. For each batch, we first randomly sample attribute values for attribute manipulation and two batches of diversity variables for diverse adversarial noises. Then, we duplicate the batch with different diversity variables. Then, we conduct attribute manipulation on the training data and generate diverse adversarial noise with different diversity variables. Then, we generate SPA samples of training data following Algorithm 1. We update the noise generator g_ϕ with the classification loss \mathcal{L}_{cls} and the diversity loss \mathcal{L}_{div} . Finally, we update the target classifier f_θ with the classification loss of real images, attribute perturbed images, and SPA images.

V. EXPERIMENTS

In this section, we report the experimental results of our proposed SPA attack and training on four public datasets.

A. Experimental Setups

Datasets. We evaluate our approach on MNIST [13], FashionMNIST [31], CelebFaces Attributes (CelebA) [32] and SICAPv2 [33] datasets. 1) **MNIST** dataset consists of 70,000 28×28 images with digits from zero to nine. 2) **Fashion-MNIST** dataset consists of 70,000 28×28 images with 10 classes of clothes. For both datasets, we consider rotation in $[-45, 45]$ degrees and scaling in $[0.7, 1.3]$ times as natural perturbations, separately.

We also conduct experiments on two real-world datasets. 3) **CelebA** dataset has more than 200K 64×64 celebrity images, each with 40 attribute annotations. We consider the "smiling" attribute as the target attribute for natural perturbation and train a classifier to predict "gender". All training data are resized to 32×32 . 4) **SICAPv2** is a medical dataset collected for prostate cancer diagnosis. There are 3773 non-cancerous patches and

3641 cancerous patches with Gleason grading equal 4. We resize the images to 64×64 and consider the rotation in $[-45, 45]$ degrees as natural perturbations. We set the size of l_∞ ball as $\epsilon = 0.1$ for MNIST and FashionMNIST, and $\epsilon = 0.01$ for CelebA and SICAPv2 datasets, which is **3** \times smaller than that of previous works.

Model Parameters. We employ ResNet-20 as the target classifier. We randomly initialize the classifier and the generator and jointly train both of them by Adam with a learning rate of 0.01 beta of $[0.9, 0.99]$ and weight decay of 0.0001 for 20K iterations, with a batch size of 100. For SPA attack, after the classifier is pre-trained, the generator is updated for 100K iterations, with a batch size of 100. We set $\epsilon_{step} = 1/4\epsilon$. We update the attribute and adversarial noise to generate SPA samples for ten steps. Both attribute value and diversity variable are optimized by Adam over the cross-entropy loss.

Baselines. To measure the robustness of the SPA-trained classifier, we compare our approach against the classifiers that are trained upon state-of-the-art adversarial training methods. We employ two gradient-based methods CW [20] and PGD [6], and a generator-based method L2LDA [24] for adversarial training. We also compare our approach against an attribute robust training method AGAT [15].

To evaluate the effectiveness of the SPA attack, we compare our SPA attack with a natural perturbation method SAA [14], and three adversarial perturbation approaches CW, PGD, and L2LDA.

Furthermore, to demonstrate the effectiveness of the proposed joint optimization of attribute values and diversity variables, we also compare with hybrid approaches which combine existing adversarial attack approaches, including CW, PGD and L2LDA with random attribute perturbation, denoted as CW-Attr, PGD-Attr and L2LDA-Attr, correspondingly.

TABLE II: Impact of the perturbation models on SPA attack on FashionMNIST dataset with "rotation" attribute, in terms of accuracy. \times denotes removing the model h_ψ or g_ϕ .

Attack \ Defense	Naive	AGAT	CW	SPA
SPA ($h_\psi \times g_\phi \times$)	0.9337	0.9153	0.9241	0.9251
SPA ($h_\psi \checkmark g_\phi \times$)	0.7451	0.8541	0.7544	0.8849
SPA ($h_\psi \times g_\phi \checkmark$)	0.0737	0.1298	0.5190	0.8428
SPA ($h_\psi \checkmark g_\phi \checkmark$)	0.0093	0.0189	0.5092	0.7996

TABLE III: Impact of the parameter optimization on SPA attack on FashionMNIST dataset with "rotation" attribute, in terms of accuracy. α and \mathbf{z} refer to the attribute value and the diversity variable in Algorithm 1. \times denotes using fixed random α or \mathbf{z} .

Attack \ Defense	Naive	AGAT	CW	SPA
SPA ($\alpha \times \mathbf{z} \times$)	0.0541	0.0863	0.5144	0.8327
SPA ($\alpha \checkmark \mathbf{z} \times$)	0.0144	0.0311	0.5102	0.8209
SPA ($\alpha \times \mathbf{z} \checkmark$)	0.0284	0.1045	0.5084	0.8255
SPA ($\alpha \checkmark \mathbf{z} \checkmark$)	0.0093	0.0189	0.5092	0.7996

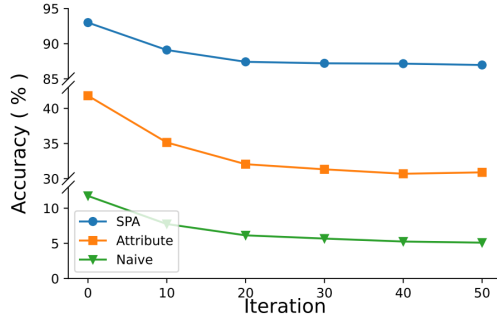


Fig. 3: Impact of attack iterations on classification accuracy on CelebA dataset with "Smiling" attribute. We perform SPA attack method on naive, AGAT [15], and SPA-trained classifiers.

B. Ablation Study

To understand (a) how the perturbation models impact the performance of SPA attack, and (b) how the optimization of the attribute value α and the diversity variable z improves the quality of SPA samples, we conduct ablation studies on the FashionMNIST dataset, using STN-based attribute manipulator. Note that STN conducts precise natural perturbation along with attributes and thus will not unintentionally introduce an adversarial noise. Therefore, the noise δ generated by the noise generator is the overall adversarial noise added to the data. This allows us to *isolate* the influence of each perturbation in SPA attacks.

Perturbation Model. Table II demonstrates the impact of each perturbation model on SPA attack. We compare the second row and the third row with the first row, respectively. We observe that the adversarial noise generator g_ϕ causes severe model failure (85% accuracy decline against naive classifier on FashionMNIST), while the attribute manipulator h_ψ only results in 18.86% accuracy decline. Therefore, the adversarial noise generator contributes more accuracy decline than the attribute manipulator against the four defense methods.

TABLE IV: Impact of the noise size ϵ on CelebA dataset with "smiling" attribute, in terms of accuracy. We conduct PGD, L2LDA, and SPA attacks on a naive classifier and a SPA-trained classifier.

Defense	Attack	Noise size ϵ			
		0.00	0.01	0.02	0.03
Naive	PGD	0.9565	0.0935	0.0225	0.0027
	L2LDA	0.9565	0.1055	0.0372	0.0108
	SPA	0.9565	0.0755	0.0214	0.0051
SPA	PGD	0.9430	0.9130	0.7315	0.5792
	L2LDA	0.9430	0.9185	0.5721	0.4744
	SPA	0.9430	0.8910	0.6122	0.4530

We further observe Table III to study the impact of parameter optimization by comparing random initialization with parameter optimization methods. By comparing the second and first row, we note that the optimization of attributes α leads to a 1.18% accuracy decline against the SPA defense on FashionMNIST. Similarly, we observe the performance

TABLE V: Impact of the attribute number on SICAPv2 dataset, in terms of accuracy. Legend of attributes: Re: reshape, Sc: scaling, Ro: rotation. We conduct the attack on a PGD-trained classifier. Rand-Attr is an adversarial attack with randomly generated attributes. As the number of attributes increases, SPA attacks are more effective. Our SPA attack performs better than Rand-Attr, where the attribute values are randomly selected, showing the efficacy of joint optimization in finding strong adversarial samples.

Attack Type	Attributes	SPA	Rand-Attr
Single Attribute	Re	0.3030	0.4136
	Sc	0.5741	0.6277
	Ro	0.3732	0.3830
Double Attributes	Re, Sc	0.2944	0.3541
	Re, Ro	0.2629	0.3842
	Sc, Ro	0.3580	0.3732
Multi Attributes	Re, Sc, Ro	0.2207	0.3317

drops 0.72% by optimizing the diversity variable z . Therefore, optimizing attributes α contributes more than optimizing the diversity variable z .

The above ablation studies show that our method's perturbation models and parameter optimizations benefit the generation of powerful attacks in a complementary way.

C. Impact of Parameters

Noise Size. As demonstrated in Table IV, we explore the noise size ϵ on the CelebA dataset with the "smiling" attribute. We conduct PGD, L2LDA, and SPA attacks on a naive classifier and a classifier adversarially trained via SPA. The size of adversarial noise varies from 0.0 to 0.03. It can be observed that when the size of noise is small (e.g., $\epsilon = 0.01$), our proposed SPA attack outperforms PGD and L2LDA by around 2%. As the noise size increases, the SPA attack can achieve a strong attack but could be outperformed by other approaches (e.g., PGD on a naive classifier).

Attack Iteration. To study the impact of the attack iteration I in Algorithm 1, we observe the change of prediction accuracy in Figure 3. The attack is conducted upon classifiers trained by naive, attribute robust, and SPA training. We observed that parameter optimization could effectively enhance the SPA samples. The accuracy becomes stable after 30 steps. We also notice that the SPA-trained classifier has the smallest relative accuracy decline ($7.2\% = \frac{93.75\% - 86.97\%}{93.75\%}$ for 50 steps) than the other baselines (26.1% for attribute robust training and 56.6% for naive training). For efficiency, we only update 10 steps in SPA training.

Attribute Number. We quantitatively explore the effect of introducing a different number of attributes for the SPA attack. We conduct experiments on the SICAPv2 dataset by attacking a classifier adversarially trained against PGD attack. Table V demonstrates that as the number of attributes increases, the SPA attack can efficiently search stronger natural perturbations by optimizing in larger attribute space. Compared to randomly sampling attributes as in Rand-SPA, our optimization-based SPA attack finds better and can find stronger adversarial samples.

TABLE VI: The result of White-box attacks on MNIST, FashionMNIST, CelebA, and SICAPv2 datasets. For adversarial noise, we set $\epsilon = 0.1$ for MNIST and FashionMNIST datasets and $\epsilon = 0.01$ for CelebA and SICAPv2 datasets. We measure the classification accuracy of the adversarially trained classifiers (rows) against various attack methods (columns).

Attack \ Defense	Naive	SAA	CW	PGD	L2LDA	CW-Attr	L2LDA-Attr	SPA	Min
MNSIT [13] Attribute \leftarrow Scaling									
Plain	0.9684	0.8236	0.1226	0.0983	0.1159	0.0351	0.0308	0.0118	0.0118
AGAT	0.9672	0.9584	0.1328	0.1240	0.1401	0.0741	0.1143	0.0219	0.0219
CW	0.9629	0.8307	0.8172	0.4131	0.7931	0.7497	0.6278	0.6415	0.6278
PGD	0.9891	0.8406	0.7521	0.7842	0.8028	0.7412	0.7749	0.6744	0.6744
L2LDA	0.9677	0.8472	0.8641	0.8140	0.9148	0.8073	0.7946	0.7441	0.7441
PGD-Attr	0.9860	0.9431	0.7581	0.7872	0.7992	0.7528	0.7617	0.7844	0.7581
SPA	0.9661	0.9575	0.9194	0.8789	0.9071	0.8951	0.8784	0.8572	0.8572
FashionMNSIT [31] Attribute \leftarrow Rotation									
Plain	0.9337	0.7496	0.0648	0.0617	0.0737	0.0165	0.0284	0.0093	0.0093
AGAT	0.9153	0.8659	0.0869	0.0794	0.1298	0.0831	0.1045	0.0189	0.0189
CW	0.9241	0.7597	0.6539	0.3642	0.5190	0.6831	0.5084	0.5092	0.5084
PGD	0.9257	0.7642	0.7365	0.7044	0.6531	0.6611	0.6417	0.5371	0.5371
L2LDA	0.9224	0.7827	0.8593	0.7281	0.7417	0.8144	0.7929	0.7731	0.7417
PGD-Attr	0.9217	0.8582	0.7912	0.7380	0.7278	0.7128	0.7417	0.6982	0.6982
SPA	0.9251	0.8845	0.8313	0.8047	0.8428	0.8127	0.8255	0.7996	0.7996
CelebA [32] Attribute \leftarrow Smiling									
Plain	0.9565	0.8830	0.1375	0.1035	0.1055	0.1035	0.0950	0.0775	0.0775
AGAT	0.9400	0.9420	0.3740	0.3577	0.4180	0.3725	0.3580	0.3415	0.3415
CW	0.9518	0.9151	0.5798	0.4966	0.5419	0.5541	0.5221	0.5077	0.5077
PGD	0.9479	0.9068	0.6431	0.9047	0.5541	0.5681	0.5328	0.5041	0.5041
L2LDA	0.9523	0.9094	0.6740	0.6937	0.7240	0.6192	0.6891	0.6390	0.6192
PGD-Attr	0.9401	0.9268	0.6458	0.8806	0.5618	0.5828	0.5527	0.5541	0.5527
SPA	0.9430	0.9405	0.9125	0.8973	0.9185	0.9035	0.8975	0.8910	0.8910
SICAPv2 [33] Attribute \leftarrow Rotation									
Plain	0.8971	0.8070	0.0014	0.0142	0.0217	0.0028	0.0121	0.0217	0.0014
AGAT	0.7624	0.8791	0.0430	0.0841	0.0402	0.0830	0.0402	0.0531	0.0402
CW	0.8732	0.7832	0.3044	0.2831	0.2629	0.3188	0.2370	0.2207	0.2207
PGD	0.8890	0.7890	0.6731	0.3347	0.4141	0.6542	0.3030	0.3732	0.3030
L2LDA	0.8877	0.8021	0.5844	0.5821	0.5741	0.5830	0.5703	0.4785	0.4785
PGD-Attr	0.8941	0.8741	0.6525	0.5506	0.4633	0.6277	0.4527	0.4071	0.4071
SPA	0.8920	0.8837	0.6207	0.8533	0.7685	0.6183	0.7051	0.6295	0.6183

D. Comparisons

In this section, we compare our approach to existing works in defense and attack. We use four popular defense methods in the literature, AGAT [15], CW [20], PGD [6] and L2LDA [24], as the baselines for adversarial training. The results on MNIST, FashionMNIST, CelebA, and SICAPv2 datasets are presented in Table VI.

First, we observe that the SPA-trained target classifier outperforms the others by a significant margin in all four datasets against jointly-perturbed samples. Compared to AGAT and SPA training, the two classifiers trained by CW and L2LDA cannot defend natural perturbations from SAA in most cases. SPA training achieves comparable performance to AGAT against natural perturbations. However, the SPA-trained classifier is robust against adversarial perturbations, while the AGAT-trained classifier is not. According to the fifth to the seventh columns in Table VI, our proposed SPA training is robust against both perturbations.

To well evaluate the performance of our proposed SPA attack, we use three existing attack methods (SAA [14], CW,

and L2LDA) and two hybrid attack methods (CW-Attr and L2LDA-Attr) as baselines. We observe that the SPA attack outperforms other natural-perturbation-based attacks, gradient-based adversarial attacks, generator-based adversarial attacks, and hybrid attacks in most cases. Although random attribute perturbation can help enhance adversarial samples, the SPA attack generates stronger samples thanks to parameter optimization. Therefore, the SPA attack is able to generate powerful jointly-perturbed samples for improving the robustness against joint perturbations.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a new adversarial training strategy named Semantic-Preserving Adversarial (SPA) training for enhancing robustness against joint perturbations in the attribute and pixel spaces by designing a novel attack mechanism. To make the classifier more robust against joint adversarial and natural perturbations, we leverage an attribute manipulator for natural perturbation and a noise generator to generate diverse adversarial noises, then optimize both attribute values and adversarial diversity variables. The SPA attack causes a larger

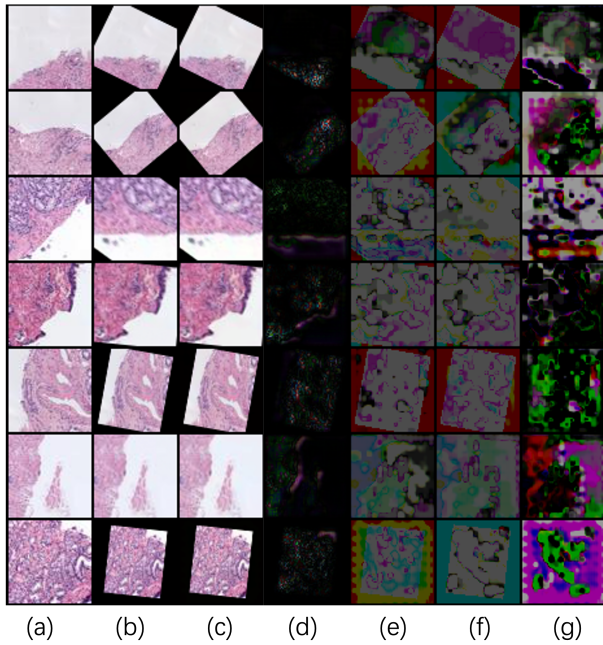


Fig. 4: Semantic-preserving adversarial (SPA) examples generated with multiple attributes on the SICAPv2 dataset. Column (a) shows original images. Column (b) shows images with optimized attributes. Column (c) shows SPA images with joint perturbations. Column (d) shows the gradient of images in column (c). Column (e) and column (f) show the adversarial noises generated with different diversity variables. Column (g) shows the difference between the two noises in columns (e) and (f). Images in columns (e), (f), (g) are amplified 30 times.

performance decline under small l_∞ norm-ball constraints compared to existing approaches. We extensively evaluate SPA attacks and training on four benchmarks and achieve state-of-the-art performance. Besides, we empirically demonstrate that SPA training applies to multiple types of natural perturbations and can be used with different surrogate functions for attribute manipulation. In the future, we are going to explore joint-perturbation space with a unified generator more effectively and adapt our SPA training to other robustness problems, not limited to classification.

REFERENCES

- [1] 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*, 2016, pp. 770–778.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.
- [4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv e-prints*, p. arXiv:1607.02533, Jul. 2016.
- [6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.
- [7] 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*. PMLR, 2018, pp. 274–283.
- [8] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=Xb8xvrtB8Ce>
- [9] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1633–1645. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf>
- [10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations*, 2017.
- [11] 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*. PMLR, 2019, pp. 7472–7482.
- [12] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=rkOg6EFwS>
- [13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.
- [14] A. Joshi, A. Mukherjee, S. Sarkar, and C. Hegde, "Semantic adversarial attacks: Parametric transformations that fool deep classifiers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4773–4783.
- [15] T. Gokhale, R. Anirudh, B. Kailkhura, J. J. Thiagarajan, C. Baral, and Y. Yang, "Attribute-guided adversarial training for robustness to natural perturbations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 7574–7582.
- [16] A. Laugros, A. Caplier, and M. Ospici, "Are adversarial robustness and common perturbation robustness independent attributes?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- [17] 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.
- [18] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=BJx040EFvH>
- [19] H. Kim, W. Lee, and J. Lee, "Understanding catastrophic overfitting in single-step adversarial training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 8119–8127.
- [20] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy*. IEEE, 2017, pp. 39–57.
- [21] H. Jiang, Z. Chen, Y. Shi, B. Dai, and T. Zhao, "Learning to defense by learning to attack," *AISTATS*, vol. 130, 2021.
- [22] H. Wang and C.-N. Yu, "A direct approach to robust deep learning using adversarial networks," in *International Conference on Learning Representations*, 2019.
- [23] K. R. Mopuri, U. Ojha, U. Garg, and R. V. Babu, "Nag: Network for adversary generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 742–751.
- [24] Y. Jang, T. Zhao, S. Hong, and H. Lee, "Adversarial defense via learning to generate diverse attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2740–2749.
- [25] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2018.
- [26] H.-T. D. Liu, M. Tao, C.-L. Li, D. Nowrouzezahrai, and A. Jacobson, "Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer," in *International Conference on Learning Representations*, 2018.
- [27] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

- [28] X. Li, C. Lin, R. Li, C. Wang, and F. Guerin, "Latent space factorisation and manipulation via matrix subspace projection," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5916–5926.
- [29] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2017–2025, 2015.
- [30] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee, "Diversity-sensitive conditional generative adversarial networks," *International Conference on Learning Representations*, 2019.
- [31] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
- [32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.
- [33] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105637, oct 2020. [Online]. Available: <https://doi.org/10.1016%2Fj.cmpb.2020.105637>