Exposing Fine-Grained Adversarial Vulnerability of Face Anti-Spoofing Models

Songlin Yang[1,2], Wei Wang[2,*], Chenye Xu[3], Ziwen He[1,2], Bo Peng[2], Jing Dong[2]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Center for Research on Intelligent Perception and Computing, CASIA [3]SenseTime Research
yangsonglin2021@ia.ac.cn, xuchenye@sensetime.com, {wwang, bo.peng, jdong}@nlpr.ia.ac.cn

Abstract

Face anti-spoofing aims to discriminate the spoofing face images (e.g., printed photos and replayed videos) from live ones. However, adversarial examples greatly challenge its credibility, where adding some perturbation noise can easily change the output of the target model. Previous works conducted adversarial attack methods to evaluate the face anti-spoofing performance without any fine-grained analysis that which model architecture or auxiliary feature is vulnerable. To handle this problem, we propose a novel framework to expose the fine-grained adversarial vulnerability of the face anti-spoofing models, which consists of a multitask module and a semantic feature augmentation (SFA) module. The multitask module can obtain different semantic features for further fine-grained evaluation, but only attacking these semantic features fails to reflect the vulnerability which is related to the discrimination between spoofing and live images. We then design the SFA module to introduce the data distribution prior for more discrimination-related gradient directions for generating adversarial examples. And the discrimination-related improvement is quantitatively reflected by the increase of attack success rate, where comprehensive experiments show that SFA module increases the attack success rate by nearly 40% on average. We conduct fine-grained adversarial analysis on different annotations, geometric maps, and backbone networks (e.g., Resnet network). These fine-grained adversarial examples can be used for selecting robust backbone networks and auxiliary features. They also can be used for adversarial training, which makes it practical to further improve the accuracy and robustness of the face anti-spoofing models. Code: https://github.com/Songlin1998/SpoofGAN

1. Introduction

Face anti-spoofing [49, 56] is significantly important to the credibility of face recognition systems, which aims to determine whether a presented face is live or spoofing. If the face anti-spoofing part is unreliable, malicious attackers can use photos and videos of your face to unlock your mobile phone or other biometric authentication systems. The researchers [1] adopted different backbone networks (e.g., Resnet [17] and Transformer [30, 47]) and auxiliary information [60] (e.g., depth maps and facial attributes), and proposed highly complicated face anti-spoofing models [56]. These models improved the performance of face anti-spoofing to high accuracy. However, the emergence of adversarial examples [2, 19, 26] poses a fatal threat to face anti-spoofing models, which can easily mislead the target model, making it output a wrong classification result with high confidence. For example, an image of your photo could have been classified as spoofing input, but this image will be classified as live input, after the modification of the adversarial attacks.

Previous works [34, 48, 54, 57] have tried to conduct adversarial attacks to these models, which only revealed the adversarial vulnerability of these face anti-spoofing models. But these attacks [5, 13, 14, 21, 29, 40, 45] entangled discrimination features (i.e, spoofing-live classification) with auxiliary features and failed to expose the fine-grained adversarial analysis. In other words, they are not able to figure out which part of the target face anti-spoofing model is vulnerable, especially the models using several auxiliary features to assist in discrimination. This makes it difficult for researchers to improve the adversarial robustness of face anti-spoofing models while maintaining the accuracy performance simultaneously.

To expose the fine-grained adversarial vulnerability of face anti-spoofing models, we propose a novel framework that consists of a multitask module and a semantic feature augmentation (SFA) module. The multitask module is a backbone network with several branches to obtain different semantic features corresponding to different auxiliary information for spoofing-live discrimination of face anti-spoofing models. This architecture is flexible to evaluate different auxiliary information and backbone networks systematically. However, only attacking these se-

*Corresponding author.

mantic features without spoofing-live prior fails to reflect the spoofing-live discrimination correlation. This means the low attack success rate towards the spoofing-live classification via attacking the semantic features. We then design the SFA module to introduce the data distribution prior for more discrimination-related gradient directions. The discrimination-related improvement is quantitatively reflected by the increase of attack success rate and our SFA module boosts the attack success rate of adversarial examples even for the one-step attack such as FGSM [19], which provides an efficient way to generate a large number of adversarial examples for quantitative experiments. This fine-grained vulnerability can be used for selecting robust backbone networks and auxiliary features, which provides a significantly important tool for model optimization, further improving the accuracy and robustness of the face anti-spoofing models.

The main contributions of this work are as follows:

- We propose a novel framework to systematically expose the fine-grained adversarial vulnerability of face anti-spoofing models.
- We propose a semantic feature augmentation (**SFA**) module to obtain discrimination-related gradient directions, increasing the attack success rate by nearly 40% on average.
- We conduct comprehensive experiments from three perspectives, which are annotations (facial attributes, spoofing types, and illumination), geometric information (depth and reflection maps), and backbone networks.

2. Related Work

2.1. Face Anti-Spoofing

Face anti-spoofing is an important guarantee for the reliability of face recognition systems, especially in some security scenarios [49, 56, 56]. In early research, face anti-spoofing algorithms were based on handcrafted features, such as LBP [53], HoG [42], and SURF [4]. Temporal features like eye-blinking [37] and lip motion [24] also received attention. Methods based on different color spaces have also been proposed, such as HSV [4], YCbCr [4] and Fourier spectrum [27]. With the popularity of methods based on deep learning, CNNs have been used for feature extraction and classification, and these CNN-based methods achieved excellent performance [3,28,56], nearly 100% accuracy in academic datasets. Auxiliary information including annotations [60] (e.g., facial attributes, spoofing types, and illumination) and geometric maps [22] (depth and reflection maps), were studied to assist the binary classification. However, the emergence of adversarial attacks [34, 48, 54, 57] exposed the vulnerability of face anti-spoofing. Different auxiliary information and backbone networks can indeed improve the classification accuracy of live and spoofing, but for this task related to security, its adversarial robustness should be studied.

2.2. Adversarial Attack

The models based on deep learning are vulnerable to adversarial attacks, which is a popular research concern in recent years [2, 58]. By adding the imperceptible noise to the original data, adversarial examples can mislead the classification easily [13]. Adversarial attacks can be divided into white-box and black-box attacks. White-box attacks [6, 19, 33] generated the adversarial perturbation by obtaining the gradient of the model, while black-box attacks [11,32] focus on the transferability [52] of adversarial examples (i.e., using the adversarial examples generated on one model can be used to attack other models).

Besides being studied as a problem [38, 41, 55], adversarial examples can also be used as a tool to expose the vulnerability of the model, as well as measuring the impact of data distribution and network architecture on the adversarial robustness [5, 13, 14, 21, 40, 45]. However, previous methods [48, 57] towards the adversarial attacks of face anti-spoofing merely tackle the fine-grained adversarial vulnerability from the perspective auxiliary information.

Physical Adversary vs. Face Spoofing Attack Some literature [9, 25, 51] has explored the physical adversarial attacks in real-world scenes, making adversarial examples generated in the digital domain can still mislead the target model after being printed. However, the aim of face anti-spoofing models is to distinguish spoofing images captured from the physical domain and real-world diversity has been considered in the dataset, which means the face spoofing attack itself is a physical attack. Our goal is to quantitatively evaluate the template discriminant features of face anti-spoofing models (i.e., efficiently test whether face anti-spoofing models can robustly classify a large number of input images in the digital domain). So we focus on digital adversarial attacks first.

2.3. Class Activation Map (CAM)

Explicability of classification decisions and localization made by the neural networks are obtained via the computation of class activation map (CAM) [43]. Furthermore, CAM and its following works [7, 20, 36, 39, 46] visualize class-relevant features in the form of a heatmap, which is the reason why we select CAM [43] as a visualization tool.

3. Method

We present our framework for exposing the fine-grained adversarial vulnerability of face anti-spoofing models in

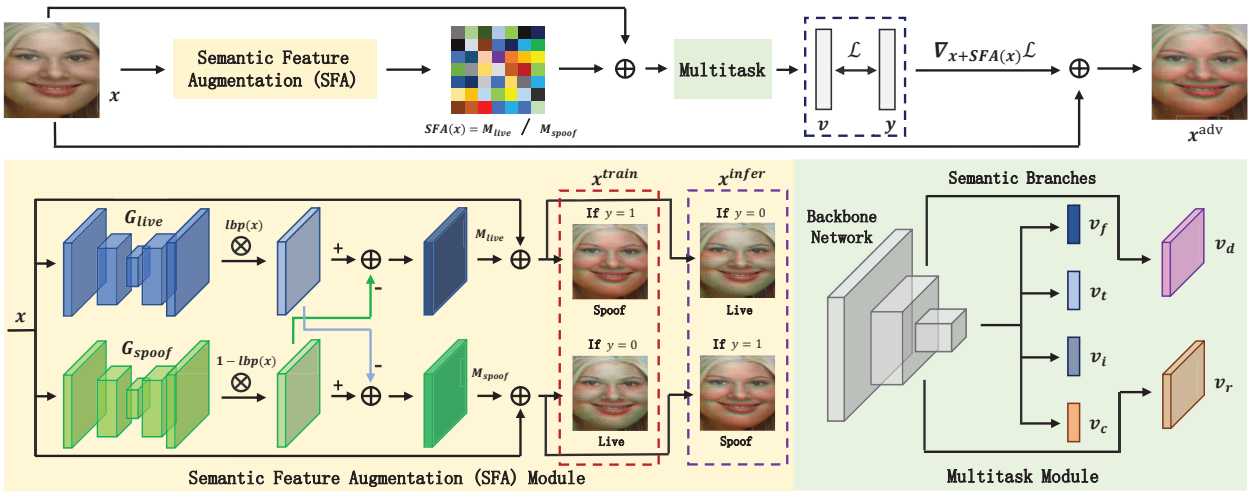

Figure 1. The framework for exposing the fine-grained adversarial vulnerability of face anti-spoofing models. The semantic feature augmentation (SFA) module aims to obtain the live/spoofing activation map $\boldsymbol{M}_{live}/\boldsymbol{M}_{spoof}$, and add the map to the input image. The multitask module is adopted to obtain the semantic features of the input image. With the SFA and multitask module, we are able to conduct adversarial attacks towards different backbone networks and auxiliary information for fine-grained adversarial analysis in the context of face anti-spoofing.

Fig. 1. We first adopt the semantic feature augmentation (SFA) module (Sec. 3.2) to obtain the live/spoofing activation map and add the map to the input image. And then, we feed the processed image into the multitask module to obtain the semantic features. Finally, we conduct fine-grained adversarial attacks (Sec. 3.3) toward different backbone networks and auxiliary information. Before introducing the details of our proposed method, we first take a one-step adversarial attack, FGSM [19], as a representative method to introduce the generation process of adversarial examples in Sec. 3.1.

3.1. Preliminary

The adversarial example $\boldsymbol{x}^{adv}$ is the image modified by the adversarial noise, and it can make the target model predict $f(\boldsymbol{x}^{adv}) = \boldsymbol{v}$ while such output is different from its label $\boldsymbol{y}$. FGSM [19] is a one-step attack, which is formally defined by

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot sign(\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{y})), \tag{1}$$

where $\boldsymbol{x}$, ϵ, and $\mathcal{L}(\cdot)$ denote the input image, max perturbation scale, and loss function. The $sign(\cdot)$ is a mathematical function that extracts the sign of a real number or vector.

3.2. Semantic Feature Augmentation (SFA)

We design our SFA module from data and model perspectives: Previous works augmented data without class information or with just their corresponding label [8, 35], while SFA considers both contrastive labels. Taking a live image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ as an example, we augment the live features of a live image, by adding live activation map $\boldsymbol{M}_{live} \in \mathbb{R}^{1\times H\times W}$ to strengthen live features and subtracting spoofing activation map $\boldsymbol{M}_{spoof} \in \mathbb{R}^{1\times H\times W}$ to weaken spoofing features, where H and W are the height and width of the input image. Moreover, CNN-based models are biased to texture changes [16], so we adopt LBP [10] to manipulate textures to further increase the attack success rate. These are reasons why SFA can make processed images more semantic-aware to live/spoofing features and achieve better results than class activation map [43] and vanilla data augmentation methods (e.g., flip and rotation). Next, we introduce each part of the pipeline of SFA module, as shown in Fig. 1:

Generator. Two variational autoencoders (VAE) [23], denoted as G_{live} and G_{spoof}, are used for generating live and spoofing activation maps ($\boldsymbol{M}_{live}$ and $\boldsymbol{M}_{spoof}$) respectively. Then, the activation maps output by G_{live} and G_{spoof} will be multiplied by LBP [10] through Hadamard product denoted as $lbp(\cdot)$ and $1 - lbp(\cdot)$. Thus, region-wise manipulation of texture can be achieved. To be specific, for the input image $\boldsymbol{x}$, the activation maps $\boldsymbol{M}_{live}$ and $\boldsymbol{M}_{spoof}$ can be obtained by the following formula:

$$\boldsymbol{M}_{live}(\boldsymbol{x}) = lbp(\boldsymbol{x}) \odot G_{live}(\boldsymbol{x}) - [1 - lbp(\boldsymbol{x})] \odot G_{spoof}(\boldsymbol{x}), \tag{2}$$

$$\boldsymbol{M}_{spoof}(\boldsymbol{x}) = -lbp(\boldsymbol{x}) \odot G_{live}(\boldsymbol{x}) + [1 - lbp(\boldsymbol{x})] \odot G_{spoof}(\boldsymbol{x}), \tag{3}$$

Discriminator. The pretrained face anti-spoofing model is chosen to be the discriminator, denoted as D. The parame-

Table 1. The attack success rates of different adversarial attacks without and with SFA.

Adversarial Attack	Attack Success Rate ↑	
	without SFA	with SFA
None	0.0002	**0.7129**
FGSM [19] ($\epsilon = 0.06$)	0.6578	**0.9046**
FGSM [19] ($\epsilon = 0.1$)	0.6899	**0.9120**
FGSM [19] ($\epsilon = 0.2$)	0.6983	**0.9161**
C&W [6] ($\epsilon = 0.06$)	0.6925	**0.8812**
C&W [6] ($\epsilon = 0.1$)	0.7239	**0.9122**
C&W [6] ($\epsilon = 0.2$)	0.8039	**0.9258**
Spatial [50] ($\epsilon = 0.06$)	0.4786	**0.9443**
PGD [33] ($\epsilon = 0.1$)	0.7045	**0.9471**
Sparse [12] ($\epsilon = 0.5$)	0.1261	**0.7832**

ters of this target model are fixed in the process of optimizing the generator.

Training Strategy and Loss Function. The modified image $\boldsymbol{x}_i^{train}$ for optimizing G_{live} and G_{spoof}, can be generated by Eq. 4. Note that input image $\boldsymbol{x}_i$ is added by activation map opposite to its label:

$$\boldsymbol{x}_i^{train} = \boldsymbol{x}_i + y_i \cdot \boldsymbol{M}_{live}(\boldsymbol{x}_i) + (1 - y_i) \cdot \boldsymbol{M}_{spoof}(\boldsymbol{x}_i), \quad (4)$$

where $[\boldsymbol{x}_i, y_i]_{i=1}^N$ is a batch of training data with labels. Note that y_i=0 if the input image is live, otherwise y_i=1.

The training objective is to make the $\boldsymbol{x}_i^{train}$, inferred by the discriminator D, consistent with the opposite label. As shown below, the Binary Cross Entropy is adopted:

$$\mathcal{L}_1 = \frac{-1}{N} \sum_{i=1}^{N} [(1-y_i) \cdot log(D(\boldsymbol{x}_i^{train})) + y_i \cdot log(1 - D(\boldsymbol{x}_i^{train}))], \quad (5)$$

To better introduce our method and experimental results in the following sections, we denote the Semantic Feature Augmentation module as $SFA(\cdot)$. Note that $SFA(\boldsymbol{x})$ generate the activation map same as the label of x:

$$SFA(\boldsymbol{x}) = y \cdot \boldsymbol{M}_{spoof}(\boldsymbol{x}) + (1 - y) \cdot \boldsymbol{M}_{live}(\boldsymbol{x}), \quad (6)$$

$$\boldsymbol{x}^{infer} = \boldsymbol{x} + SFA(\boldsymbol{x}). \quad (7)$$

3.3. Fine-Grained Adversarial Attack

Many previous works [22, 60] used auxiliary information (e.g. illumination maps) to help face anti-spoofing models learn the discriminant boundary. We perturb the specific auxiliary feature and evaluate the effect of such perturbation, to figure out which part really affects the discriminant boundary of spoofing/live. The multitask network aims to get learned auxiliary features for generating fine-grained adversarial examples. Such a 'Backbone + Branches' structure is suitable for studying the relationships between auxiliary information, backbone networks, and spoofing/live discriminant features.

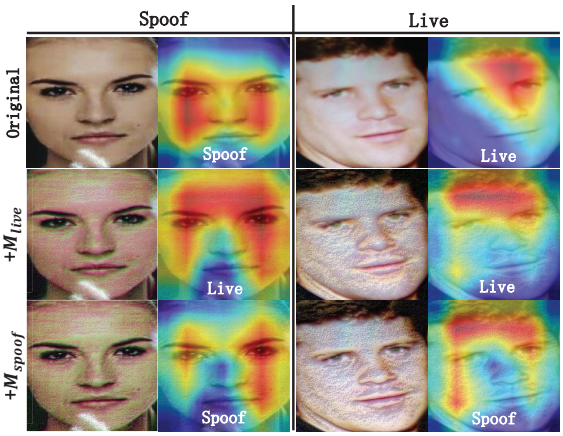

Figure 2. Visualization and perturbation effect on the classification of live and spoofing maps generated by SFA. An original spoofing (replay attack) and live sample with their class activation maps (CAM) are shown in 1st and 2nd columns, while 2nd and 3rd rows are the results added by $\boldsymbol{M}_{live}$ and $\boldsymbol{M}_{spoof}$ respectively. The spoofing activation region tends to distribute around the face in the form of vertical lines, while the live activation region distributes in the forehead area of the face.

In order to fully mine the impact of data annotation and auxiliary information on the adversarial robustness of face anti-spoofing, we adopt the multitask network with auxiliary information [60] as a basic model. As shown in Fig. 1, the backbone network is marked as gray, and the last logits are output to four fully-connected layers, to obtain the vector $\boldsymbol{v}_f$, $\boldsymbol{v}_t$, $\boldsymbol{v}_i$ and $\boldsymbol{v}_c$, which represents facial attribute, spoofing type, illumination and binary classification of live and spoofing. The three annotation vectors are compared with ground-truth label $\boldsymbol{y}_f$, $\boldsymbol{y}_t$ and $\boldsymbol{y}_i$, thus we get the semantic loss function $\mathcal{L}_a$ as follows:

$$\mathcal{L}_a = \lambda_f \cdot \mathcal{L}_f(\boldsymbol{v}_f, \boldsymbol{y}_f) + \lambda_t \cdot \mathcal{L}_t(\boldsymbol{v}_t, \boldsymbol{y}_t) + \lambda_i \cdot \mathcal{L}_i(\boldsymbol{v}_i, \boldsymbol{y}_i), \quad (8)$$

where $\mathcal{L}_t$ and $\mathcal{L}_i$ use Softmax Cross Entropy loss, and $\mathcal{L}_f$ uses Binary Cross Entropy loss.

Furthermore, reflection map $\boldsymbol{v}_r$ and depth map $\boldsymbol{v}_d$ are captured by two geometric map generators. According to [60], for the depth map of the sample labeled as live, its ground truth $\boldsymbol{y}_d$ is obtained by PRNet [15], while the depth map of the sample labeled as spoofing is zero. The method proposed in [59] is adopted to generate the ground truth of the reflection map of the sample labeled as spoofing, and the reflection map of the sample labeled as live is zero. Therefore, we can get the following geometric loss function $\mathcal{L}_g$:

$$\mathcal{L}_g = \lambda_d \cdot \mathcal{L}_d(\boldsymbol{v}_d, \boldsymbol{y}_d) + \lambda_r \cdot \mathcal{L}_r(\boldsymbol{v}_r, \boldsymbol{y}_r), \quad (9)$$

where $\mathcal{L}_d$ and $\mathcal{L}_r$ are Mean Square Error loss. We use Softmax Cross Entropy loss as $\mathcal{L}_c$, thus we can get the final

Table 2. The attack success rates (ASR) of FGSM [19] with SFA, data augmentation and class activation map. Note that 'Ensemble' means ensemble with different data augmentation operations.

Data Augmentation		Class Activation Map	
Method	ASR ↑	Method	ASR ↑
None	0.6578	None	0.6578
Vertical Flip	0.7313	Gradcam [43]	0.7375
Horizontal Flip	0.7063	Gradcam++ [7]	0.7625
Rotation(30°)	0.7563	Scorecam [46]	0.2875
Brightness	0.6938	Ablationcam [39]	0.7375
Hue	0.7250	Eigencam [36]	0.8000
Ensemble	0.8152	Layercam [20]	0.7625
SFA	**0.9046**	SFA	**0.9046**

optimization objective:

$$\mathcal{L}_2 = \mathcal{L}_c(\boldsymbol{v}_c, \boldsymbol{y}_c) + \mathcal{L}_a + \mathcal{L}_g, \tag{10}$$

Then, from the perspective of auxiliary information (annotation vectors and geometric maps), we attack different parts of the last layer and analyze the adversarial robustness of the target model. The adversarial vulnerability of different auxiliary information can be illustrated by the change in classification accuracy. However, only attacking these semantic features fails to reflect the discrimination-related vulnerability, which means a low attack success rate towards the spoofing-live classification via attacking the semantic features. So adversarial perturbation is generated based on the images modified by SFA to obtain gradient directions related to spoofing-live discrimination.

We adopt FGSM [19] as the basic method to present the formulation of generating these fine-grained adversarial examples:

$$\boldsymbol{x}_s^{adv} = \boldsymbol{x} + \epsilon \cdot sign(\nabla_{\boldsymbol{x}+SFA(\boldsymbol{x})}\ \mathcal{L}_s(\boldsymbol{v}_s, y_s)), \tag{11}$$

where $s \in \{f, t, i, d, r, c\}$ respectively represent the facial attribute, spoofing type, illumination, depth map, reflection map, and classification. ϵ indicates the step of the attack.

Furthermore, replacing the backbone network with different state-of-art models can provide us with a new perspective to study the adversarial vulnerability of face anti-spoofing models, from the effects of backbone structures.

4. Experiments

4.1. Experimental Settings

Network Architectures. We adopt the VAE architecture [23] as G_{live} and G_{spoof} of SFA module, which has five hidden units with $\{32, 64, 128, 256, 512\}$ dimensions. Each hidden unit consists of vanilla 2D convolution, batch normalization, and LeakyReLU activation layer. In multi-task module, the two geometric map generators consist of a Conv 3×3 followed by an upsample to 14×14. We set $\lambda_f = 1$, $\lambda_t = 0.1$, $\lambda_i = 0.01$, $\lambda_d = 0.1$ and $\lambda_r = 0.1$.

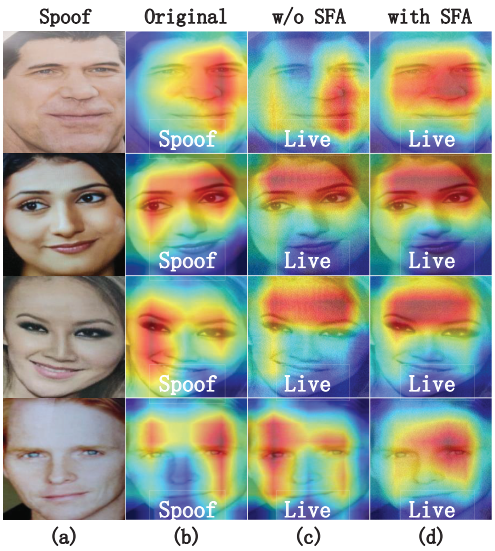

Figure 3. Differences of adversarial attacks without and with SFA module. (a) Original spoofing images obtained from CelebA-spoof [60]. (b) CAM of the original images. (c) CAM of the adversarial images without SFA. (d) CAM of the adversarial images with SFA (completely different from Fig. 4(b)). Note that (c) and (b) are perturbed by the FGSM [19] with the same attack step ($\epsilon = 0.1$) and backbone network (Resnet-18 [17]). SFA makes the adversarial perturbation more concentrated on the semantic features and boosts adversarial attacks, so attacks with SFA do change discrimination-related information of input images. Such effects and differences are illuminated by CAM, which can explicitly locate the classification decision.

Dataset. The purpose of our paper is to expose the adversarial vulnerability of face anti-spoofing from the fine-grained perspective and figure out which part of the target models is vulnerable. So our experiments use CelebA-spoof [60] as the dataset, which has three significant advantages: (a) Large-Scale: CelebA-spoof comprises 625,537 pictures of 10,177 subjects. (b) Diversity: The spoofing images are captured from 8 scenes (2 environments $\times$ 4 illumination conditions) with more than 10 sensors. (c) Annotation richness: CelebA-spoof contains 10 spoofing type annotations, as well as the 40 attribute annotations inherited from the original CelebA [31] dataset. These three advantages are helpful for us to construct the required experimental scenarios, and the experimental results based on this large-scale dataset are more general. We randomly sample 600,000 images from different categories to train the SFA module, and we randomly sample 10,000 images from the left data in every evaluation.

Metrics. The attack success rate (ASR) and classification accuracy are metrics for the performance of adversarial attacks. The results predicted by the target model are correct when the output after Softmax is the same as the ground

Table 3. The ablation study of SFA without and with LBP using different adversarial attacks.

Attack Methods	Attack Success Rate ↑	
	without LBP	with LBP
FGSM [19] ($\epsilon = 0.1$)	0.7121	**0.9120**
C&W [6] ($\epsilon = 0.1$)	0.8403	**0.9122**
Spatial [50] ($\epsilon = 0.06$)	0.8001	**0.9443**
PGD [33] ($\epsilon = 0.1$)	0.8832	**0.9471**
Sparse [12] ($\epsilon = 0.5$)	0.6755	**0.7832**

truth. Mean Square Error is taken to measure the change of each feature output. Each quantitative result is tested at least three times and then averaged.

4.2. Effectiveness Analysis of SFA Module

The semantic feature augmentation (SFA) module is designed to provide better gradients related to the spoofing-live discrimination. Its characteristic is to learn the activation maps of the target model towards both classes of live and spoofing data, by constructing positive and negative samples weighted by the texture filter. With the help of SFA, the adversarial perturbation can be added to more semantic-aware to live and spoofing features. In this way, it can make gradients more related to the spoofing-live decision boundary, thus improving the attack success rate of the fine-grained adversarial attacks.

To thoroughly evaluate the effectiveness of SFA, in this section, we first show that SFA can generate two different maps for one image (See Fig. 2), and study the effect of SFA on different adversarial attacks, to verify that SFA can improve the attack success rate (See Fig. 3 and Tab. 1). Then, we compare SFA with different methods of class activation map (CAM) and data augmentation (See Tab. 2), to illustrate the advantages of SFA over previous methods. Finally, ablation experiments on the texture filter are carried out to show its necessity (See Tab. 3).

Contrastive Activation Maps of Live and Spoofing. SFA can generate two different maps for one image, corresponding to live and spoofing respectively, as shown in Fig. 2. According to numerous empirical observations on the visualization of experimental results, spoofing activation tends to distribute at the edge and has the law of a quadrilateral grid. For live activation, the inner region of the face is stronger, especially the forehead. Such contrastive activation maps extract the discrimination information, which can be adopted for data augmentation to boost the adversarial attacks. We use the $SFA(\cdot)$ denoted in Eq. 6 to add an activation map with the same label to the input image to obtain better gradient directions to generate adversarial examples.

Different Adversarial Attacks with SFA. As shown in Tab. 1, the attack success rates of different adversarial attacks [6, 12, 19, 33, 50] based on Resnet-18 [17] have improved a lot with SFA. To qualitatively demonstrate the influence of the SFA module on discrimination of the target face anti-spoofing model, as shown in Fig. 3, we then visualize a spoofing example with its adversarial examples generated without and with SFA based on the Resnet-18 [17].

Table 4. The accuracy of Resnet-18 [17] when different annotation vectors and geometric maps are attacked.

Attacked Parts	Attack Success Rate ↑	
	without SFA	with SFA
Facial Attribute	0.6341	0.6201
Spoofing Type	0.6455	**0.7559**
Illumination	0.6475	**0.7679**
Depth Map	0.0041	**0.6694**
Reflection Map	0.0001	**0.6657**
Classification	0.6578	**0.9046**

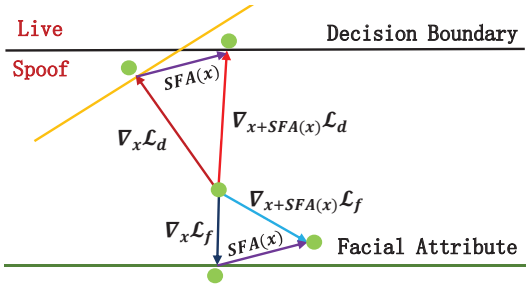

Figure 4. The interpretation of differences between facial attribute boundary and depth map boundary.

Differences among SFA, Data Augmentation and Class Activation Map. Since the SFA module is used to enhance the semantic information of images by introducing data prior, its function is similar to that of data augmentation and class activation map. As shown in Tab. 2, we select typical methods of data augmentation and class activation map (CAM) [7, 20, 36, 39, 43, 46] to compare with SFA and take FGSM [19] as the adversarial attack. To compare with the data augmentation, we not only evaluate operations separately, but also present the average of three combinations: (a) Vertical flip + Horizontal flip + Brightness, (b) Vertical flip + Horizontal flip + Rotation + Hue, and (c) Vertical flip + Horizontal flip + Rotation + Brightness + Hue. The geometric transformations made by data augmentations diversify the gradients but have little effect on the enhancement of semantic-aware features. Both CAM and SFA have studied the activation of the model towards specific classes, but CAM only considers the unique class. When SFA generates activation maps, it not only enhances the target class but also considers the opposite class.

Ablation of Texture Filter. The CNN-based networks are biased to texture features [16] (i.e., the discrimination results of the CNN-based models are easily effected by texture manipulation), which is the vulnerability of methods based on deep learning. To introduce this model prior into

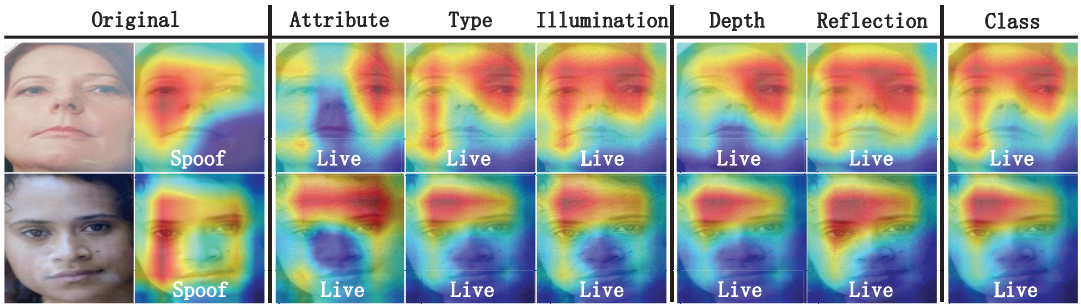

Figure 5. Visualization of adversarial attacks on different annotation vectors and geometric maps of spoofing data. For facial attributes, although the adversarial attack has perturbed the feature vector of facial attributes, it does not change the pattern of spoofing activation, different from other annotation vectors and geometric maps.

Table 5. The accuracy of different backbone networks when their annotation vectors and geometric maps are attacked. Note that we use the decline of accuracy to represent which part is more vulnerable to adversarial attacks.

Backbone Network		VGG [44]	Resnet [17]	Densenet [18]	Swin Transformer [30]
Original Accuracy		0.9416	0.9988	0.9971	0.9989
Annotation	Facial Attribute	0.7849	0.5799	0.6598	0.4623
	Spoofing Type	0.3462	0.2441	0.3838	0.1684
	Illumination	0.2736	0.2321	0.2999	0.2965
Geometric Map	Depth Map	0.4483	0.3306	0.3686	0.3302
	Reflection Map	0.6484	0.3343	0.3304	0.3408
Classification		0.2876	0.0954	0.2744	0.0962

the generation process of semantic feature augmentation, we adopt the LBP as the texture filter for the SFA module. As shown in Tab. 3, we conduct the ablation study of the texture filter, and demonstrate that considering changing texture features can improve the success rate of different adversarial attacks.

4.3. Fine-Grained Adversarial Analysis

Analysis of Auxiliary Information. Take the Resnet-18 [17] trained on CelebA-spoof [60] as the target model, and FGSM [19] is used as the method of adversarial attacks. As shown in Tab. 4, SFA can improve the success attack rate of all parts except facial attributes. As shown in Fig. 5, for facial attributes, its pattern of activation is different from other annotation vectors and geometric maps. This shows that spoofing types, illumination, depth map, and reflection map are more relevant to the ground truth decision boundary. As shown in Fig. 4, we take the depth map as an example to interpret such differences. Considering a spoofing sample, the depth map boundary has a higher correlation with the decision boundary, so SFA can make it pass through such a decision boundary. However, the facial attribute boundary has a low correlation with the decision boundary of the live and spoofing. Even if SFA biases the example towards that decision boundary, it fails to cross such boundary of classification of the live and spoofing.

Analysis of Backbone Networks. According to the latest survey [56], we select VGG-13 [44], Resnet-18 [17], Densenet-121 [18] and Swin Transformer [30] as four representative backbone networks. The checkpoints of the target models are trained on CelebA-spoof [60], and the accuracy of each model is close to 100%. Analysis of different backbone networks in the task of face anti-spoofing will be carried out in three perspectives: (a) The changes of accuracy when different annotation and geometric parts are attacked. (b) The changes of each annotation and geometric part when the binary classification of the live and spoofing is attacked. (c) We use the adversarial examples generated by one network to attack the others to evaluate the adversarial transferability of different backbone networks.

a. Adversarial Attacks on Different Annotation and Geometric Maps

We attack the different annotations (facial attributes, spoofing types, and illumination), geometric maps (depth and reflection), and binary classification in turn, using FGSM [19] and the same max perturbation scale $\epsilon = 0.2$. The changes in the accuracy of each backbone network are shown in Tab. 5. Through the comparison of columns, the correlation of each feature vector towards the binary classification of the live and spoofing can be reflected. The accuracy of facial attributes deserves attention while attacking facial attributes has less impact on the results of binary classification. Such a phenomenon exists in all four backbone networks. This shows that the annotation of facial attributes is redundant annotation for the binary classification of live and spoofing. Then, we analyze other attacked parts

Table 6. The changes of annotation vectors and geometric maps when the binary classification of the target model is attacked.

Backbone Network		VGG [44]	Resnet [17]	Densenet [18]	Swin Transformer [30]
Annotation	Facial Attribute	0.39	0.20	0.07	0.20
	Spoofing Type	0.70	2.40	0.03	0.20
	Illumination	12.80	6.60	0.70	8.40
Geometric Map	Depth Map	4.21	0.10	0.04	0.00
	Reflection Map	2.30	13.60	1.80	0.40

Table 7. The accuracy of different backbone networks when attacked by other networks.

Backbone Network	VGG [44]		Resnet [17]		Densenet [18]		Swin Transformer [30]	
	w/o SFA	with SFA	w/o SFA	with SFA	w/o SFA	with SFA	w/o SFA	with SFA
Original Accuracy	0.9416		0.9988		0.9971		0.9989	
VGG [44]	0.6121	**0.2876**	0.7863	**0.1161**	0.81008	**0.3738**	0.2679	**0.2001**
Resnet [17]	0.4212	**0.1190**	0.3101	**0.0954**	0.3245	**0.1394**	0.4016	**0.1297**
Densenet [18]	0.4521	**0.2289**	0.5025	**0.1062**	0.5763	**0.2744**	0.5538	**0.2840**
Swin Transformer [30]	0.3192	**0.1224**	0.2989	**0.1018**	0.2296	**0.0904**	0.1977	**0.0962**

of different backbone networks from the rows. Interestingly, VGG and Densenet have better adversarial robustness when the accuracy is approximately equal. Therefore, when selecting backbone networks for specific tasks, we should handle the trade-off between accuracy and adversarial robustness.

b. Adversarial Attacks on Spoofing-Live Classification

As shown in Tab. 6, we present the changes of each annotation vector and geometric map of different backbone networks when the classification of live and spoofing is attacked by FGSM ($\epsilon = 0.2$) [19]. The Mean Square Error is used as a metric of change. To have better comparability among various feature vectors, the error is divided by l_2-norm of output vectors and maps. Observing the rows, the structure of VGG can enhance the correlations between different feature parts, resulting in obvious changes in other features when the binary classification is attacked. Through the comparison of columns, the parts related to light tend to change a lot, including illumination and reflection maps. According to [22] and human perception, the differences in light between live and spoofing samples are obvious, especially in spoofing types such as replay. Furthermore, for RGB images, although the depth information is used in our experiments, the change in the depth map is small. It is worth noting that in the application scenario where most existing commercial cameras collect RGB images, the way of making the networks learn to use depth information is an important problem when tackling binary classification of live and spoofing.

c. The Adversarial Transferability among Different Backbone Networks

We use FGSM [19] and the same max perturbation scale $\epsilon = 0.2$ to generate adversarial examples based on different backbone networks, and evaluate the accuracy of the other networks when they are attacked by these adversarial examples. As shown in Tab. 7, the adversarial examples have some ability to transfer adversarial attacks to the other networks. With the help of our SFA module, the adversarial transferability of adversarial examples can be strengthened.

5. Conclusions and Ethical Concerns

In this paper, we propose a novel framework to expose the fine-grained adversarial vulnerability of face anti-spoofing models. Our semantic feature augmentation (SFA) module is able to provide more discrimination-related gradients and increase the attack success rate by nearly 40% on average. We use these tools to evaluate three annotations, two geometric maps, and four backbone networks, drawing several meaningful and practical results. These novel perspectives can help our community to select annotations, geometric information, and backbone networks with better adversarial robustness when solving the task of face anti-spoofing, so as to improve the reliability of biometric authentication systems. In the future, we will explore adversarial learning based on our adversarial attacks.

Ethical Concerns. Before launching defense, studying attacks is significantly necessary. This paper guides the adversarial examples as a tool to analyze face anti-spoofing models. We aim to improve the positive impact of the adversarial examples. The methods proposed in our paper can deepen our understanding of data and models, instead of undermining the security of current face recognition systems.

6. Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2021YFC3320103, the National Natural Science Foundation of China (NSFC) under Grant 61972395, 62272460, a grant from Young Elite Scientists Sponsorship Program by CAST (YESS), and sponsored by CAAI-Huawei MindSpore Open Fund.

References

[1] Faseela Abdullakutty, Eyad Elyan, and Pamela Johnston. A review of state-of-the-art in face presentation attack detection: From early development to advanced deep learning and multi-modal fusion methods. *Information fusion*, 75:55–69, 2021. 1

[2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 1, 2

[3] Davide Belli, Debasmit Das, Bence Major, and Fatih Porikli. A personalized benchmark for face anti-spoofing. In *WACV*, pages 338–348, 2022. 2

[4] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2016. 2

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv*, 2019. 1, 2

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 2, 4, 6

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pages 839–847. IEEE, 2018. 2, 5, 6

[8] Zih-Ching Chen, Lin-Hsi Tsao, Chin-Lun Fu, Shang-Fu Chen, and Yu-Chiang Frank Wang. Learning facial liveness representation for domain generalized face anti-spoofing. In *ICME*, pages 1–6. IEEE, 2022. 3

[9] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *ECCV*, pages 514–532. Springer, 2022. 2

[10] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, pages 1–7. IEEE, 2012. 3

[11] Antonio Emanuele Cinà, Alessandro Torcinovich, and Marcello Pelillo. A black-box adversarial attack for poisoning clustering. *Pattern Recognition*, 122:108306, 2022. 2

[12] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *ICCV*, pages 4724–4732, 2019. 4, 6

[13] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *CVPR*, pages 321–331, 2020. 1, 2

[14] Steffen Eger and Yannik Benz. From hero to z\'eroe: A benchmark of low-level adversarial attacks. *arXiv*, 2020. 1, 2

[15] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 4

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv*, 2018. 3, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5, 6, 7, 8

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 7, 8

[19] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[20] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.*, 30:5875–5888, 2021. 2, 5, 6

[21] Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. *arXiv*, 2019. 1, 2

[22] Taewook Kim, YongHyun Kim, Inhan Kim, and Daijin Kim. Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In *ICCVW*, pages 0–0, 2019. 2, 4, 8

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013. 3, 5

[24] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security*, 2(3):548–558, 2007. 2

[25] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *ICPR*, pages 819–826. IEEE, 2021. 2

[26] Zixiao Kong, Jingfeng Xue, Yong Wang, Lu Huang, Zequn Niu, and Feng Li. A survey on adversarial attack in the age of artificial intelligence. *Wireless Communications and Mobile Computing*, 2021, 2021. 1

[27] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. In *Biometric technology for human identification*, volume 5404, pages 296–303. International Society for Optics and Photonics, 2004. 2

[28] Mingxin Liu, Jiong Mu, Zitong Yu, Kun Ruan, Baiyi Shu, and Jie Yang. Adversarial learning and decomposition-based domain generalization for face anti-spoofing. *Pattern Recognition Letters*, 155:171–177, 2022. 2

[29] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *CVPR*, pages 15105–15114, 2022. 1

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *In ICCV*, 2021. 1, 7, 8

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 5

[32] Chen Ma, Li Chen, and Jun-Hai Yong. Simulating unknown target models for query-efficient black-box attacks. In *CVPR*, pages 11835–11844, 2021. 2

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv*, 2017. 2, 4, 6

[34] Junjie Mao, Bin Weng, Tianqiang Huang, Feng Ye, and Liqing Huang. Research on multimodality face antispoofing model based on adversarial attacks. *Security and Communication Networks*, 2021, 2021. 1, 2

[35] Abdelrahamn Ashraf Mohamed, Marwan Mohamed Nagah, Mohamed Gamal Abdelmonem, Mohamed Yasser Ahmed, Mahmoud El-Sahhar, and Fatma Helmy Ismail. Face liveness detection using a sequential cnn technique. In *CCWC*, pages 1483–1488. IEEE, 2021. 3

[36] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *IJCNN*, pages 1–7. IEEE, 2020. 2, 5, 6

[37] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, pages 1–8. IEEE, 2007. 2

[38] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *arXiv*, 2020. 2

[39] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *WACV*, pages 983–991, 2020. 2, 5, 6

[40] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020. 1, 2

[41] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, pages 8093–8104. PMLR, 2020. 2

[42] William Robson Schwartz, Anderson Rocha, and Helio Pedrini. Face spoofing detection through partial least squares and low-level descriptors. In *IJCB*, pages 1–8. IEEE, 2011. 2

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2, 3, 5, 6

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 7, 8

[45] Liang Tong, Zhengzhang Chen, Jingchao Ni, Wei Cheng, Dongjin Song, Haifeng Chen, and Yevgeniy Vorobeychik. Facesec: A fine-grained robustness evaluation framework for face recognition systems. In *CVPR*, pages 13254–13263, 2021. 1, 2

[46] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 2, 5, 6

[47] Zhuo Wang, Qiangchang Wang, Weihong Deng, and Guodong Guo. Face anti-spoofing using transformers with relation-aware mechanism. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):439–450, 2022. 1

[48] Haibin Wu, Andy T Liu, and Hung-yi Lee. Defense for black-box attacks on anti-spoofing models by self-supervised learning. *arXiv*, 2020. 1, 2

[49] Hangtong Wu, Dan Zeng, Yibo Hu, Hailin Shi, and Tao Mei. Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. *IEEE TCSVT*, 2021. 1, 2

[50] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *In ICLR*, 2018. 4, 6

[51] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *CVPR*, pages 11845–11854, 2021. 2

[52] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739, 2019. 2

[53] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *ICB*, pages 1–6. IEEE, 2013. 2

[54] Linxi Yang, Jiezhi Yang, Mingjie Peng, Jiatian Pi, Zhiyou Wu, Xunyi Zhou, and Jueyou Li. Sparse adversarial attack based on lq-norm for fooling the face anti-spoofing neural networks. *Journal of Electronic Imaging*, 30(2):023023, 2021. 1, 2

[55] Songlin Yang, Wei Wang, Yuehua Cheng, and Jing Dong. A systematical solution for face de-identification. In *Chinese Conference on Biometric Recognition*, pages 20–30. Springer, 2021. 2

[56] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE TPAMI*, 2022. 1, 2, 7

[57] Bowen Zhang, Benedetta Tondi, and Mauro Barni. Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability. *Computer Vision and Image Understanding*, 197:102988, 2020. 1, 2

[58] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *CVPR*, pages 15115–15125, 2022. 2

[59] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, pages 4786–4794, 2018. 4

[60] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *ECCV*, pages 70–85. Springer, 2020. 1, 2, 4, 5, 7