Spoof Trace Discovery for Deep Learning Based Explainable Face Anti-Spoofing

Haoyuan Zhang[1,2], Xiangyu Zhu[1,2], Li Gao[3], Jiawei Pan[1,2], Kai Pang[4], Guoying Zhao[5], Zhen Lei[1,2,6,7,*]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]China Mobile Financial Technology Co., Ltd., Beijing, China
[4]Guangzhou Pixel Solutions Co., Ltd., Guangzhou, China
[5]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
[6]CAIR, HKSIS, Chinese Academy of Sciences, Hong Kong, China
[7]SCSE, the Faculty of Innovation Engineering, M.U.S.T, Macau, China
{zhanghaoyuan2023, xiangyu.zhu, panjiawei2023, zhen.lei}@ia.ac.cn
gaolids@chinamobile.com, pangkai@pixelall.com, guoying.zhao@oulu.fi

Abstract

With the rapid growth usage of face recognition in people's daily life, face anti-spoofing becomes increasingly important to avoid malicious attacks. Recent face anti-spoofing models can reach a high classification accuracy on multiple datasets but these models can only tell people "this face is fake" while lacking the explanation to answer "why it is fake". Such a system undermines trustworthiness and causes user confusion, as it denies their requests without providing any explanations. In this paper, we incorporate XAI into face anti-spoofing and propose a new problem termed X-FAS (eXplainable Face Anti-Spoofing) empowering face anti-spoofing models to provide an explanation. We propose SPTD (SPoof Trace Discovery), an X-FAS method which can discover spoof concepts and provide reliable explanations on the basis of discovered concepts. To evaluate the quality of X-FAS methods, we propose an X-FAS benchmark with annotated spoof traces by experts. We analyze SPTD explanations on face anti-spoofing dataset and compare SPTD quantitatively and qualitatively with previous XAI methods on proposed X-FAS benchmark. Experimental results demonstrate SPTD's ability to generate reliable explanations.

1. Introduction

Due to the vulnerability of face recognition (FR) systems to attack, academia and industry have paid extensive attention to Face Anti-Spoofing (FAS) technology [47]. Nowadays, FAS technologies [41, 23, 25, 40] have already reached a high level of defense against physical attacks such as print, replay, makeup and 3D masks, etc. However, these technologies can only answer the question "whether the photograph provided was fake" while lacking the evidence to support its results which brings doubts and implicit bias.

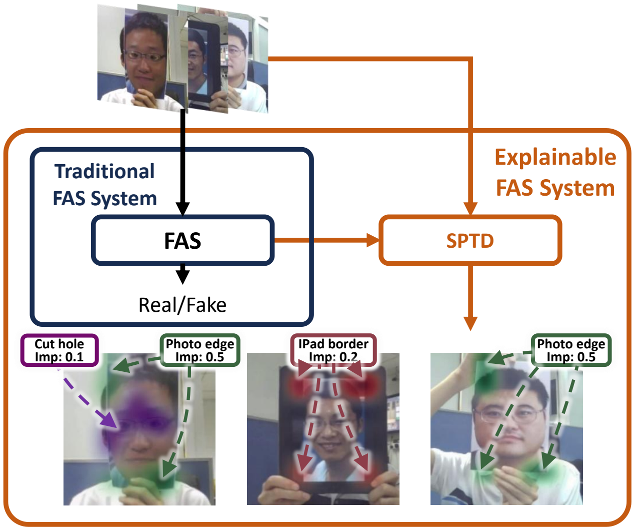

Figure 1. X-FAS method provide explanations on top of classification result. Imp indicates the *importance of the concept*.

When a FR system rejects an image for security reasons, it is considered necessary to provide an explanation. Without such explanation, the interaction can become frustrating and uncomfortable for users, which lack transparency and trustworthiness. Thus an X-FAS (eXplainable Face Anti-Spoofing) system is advocated to provide user-friendly results by generating explanations based on FAS classification models, which is the goal of this paper. We believe X-FAS is crucial as it can make the FAS system more trustworthy

[*] Corresponding author

and significantly enhance the user experiences.

The field of eXplainable Artificial Intelligence (XAI) [6, 20, 46, 8] has emerged to demystify the inner workings of black-box models and offer insights into their decision-making processes. Traditional XAI methods focused on providing a single explanation for a given input, typically in the form of heatmaps that highlight key regions. However, recent advancements in XAI [35, 42] have underscored the importance of concept discovery which can provide not only heatmaps but also the corresponding activated concepts. These breakthroughs in XAI techniques lay a solid foundation for exploring the field of X-FAS.

In this paper, we introduce SPTD (SPoof Trace Discovery), an X-FAS method designed to discover spoof concepts and provide explanations for attack images. Given a well-trained FAS model, SPTD can discover spoof concepts from a given FAS dataset and analyze the importance of each concept. With the help of discovered concepts, SPTD can mark the attention region of each concept if the input image is judged as a fake sample during inference. Notably, the whole process no need to updating the FAS model, thus keep the original performance. Examples are shown in Figure 1, SPTD find multiple activated concepts in attack images and provide the corresponding attention regions. To evaluate X-FAS methods, we present an X-FAS benchmark that includes 13 spoof types and 777 samples, aimed at assessing the quality of explanations generated by X-FAS techniques. Experiments demonstrate that SPTD can identify key spoof concepts and provide heatmaps corresponding to these concepts in FAS tasks, thus enhancing the trustworthiness of the system for users. Our main contributions can be described as follows:

1. We propose a new problem termed X-FAS (eXplainable Face Anti-Spoofing) that generates reliable explanations on top of FAS classification results and introduce an X-FAS method SPTD (SPoof Trace Discovery) which can discover spoof concepts and provide attention region of each concept in attack images.

2. We propose an X-FAS benchmark with expert annotations to evaluate the quality of explanations generated by X-FAS methods.

The quantitative results on the X-FAS benchmark, along with the qualitative analysis of the generated explanations demonstrate the efficacy of SPTD. These results highlight its ability to provide reliable explanations, thereby enhancing the trustworthiness of FAS models.

2. Related works

2.1. Face anti-spoofing (FAS)

FAS aims to differentiate between real and fake facial images, preventing deception in facial recognition systems. Over the years, FAS has advanced across various dimensions, including model architecture, task formulation, and training paradigms. In terms of model architecture, FAS has evolved from Convolution Neural Networks (CNNs) [14, 30] to Vision Transformers (ViT) [7, 29] and now to large models [48, 44]. Regarding task formulation, FAS has expanded from single-modal RGB recognition [49, 3] to multimodal tasks [24] incorporating infrared and depth images, as well as from within-domain [16, 17, 18] to cross-domain tasks [19, 15]. In terms of training paradigms, FAS has shifted from single-image input [50, 43, 27] to pre-trained vision-language contrastive learning [41, 26]. These advancements have made FAS more powerful, but the opacity of black-box neural networks still limits the trustworthiness of FAS systems. Although previous methods varies from training paradigms, they use same backbones (e.g. ResNet [14], ConvNext [30], ViT [7]). Thus, we propose a framework to provide additional explanations for CNN-based FAS methods to enhance system trustworthiness.

2.2. Explainable AI (XAI)

Explainable AI (XAI) aims to improve the interpretability and transparency of deep learning models by providing human-understandable explanations for their decisions. In computer vision, gradient-based methods such as Grad-CAM [38], Ablation-CAM [37], and related approaches [2, 1] leverage activation and gradient information within the model to attribute decisions to specific regions of the input image (e.g., highlighting an airplane when classifying an image as a plane). In contrast, perturbation-based methods like RISE [34] treat the network as a black box, relying solely on input-output pairs. These methods systematically perturb the input image and observe the resulting changes in the model's output to infer decision attribution. Additionally, concept-based approaches such as ACE [11] and CRAFT [10] automatically decompose a single decision into multiple interpretable concepts, enhancing human understanding. In the context of FAS, identifying multiple spoof traces from a single prediction aligns well with the objectives of concept-based methods, making them particularly suitable for improving model explainability in FAS.

2.3. XAI + FAS

Early studies on XAI for FAS aimed to identify key differences between attack and genuine images [32]. Jourabloo *et al.* [21] and Liu *et al.* [28] employed disentanglement techniques to separate spoof noise from facial features, facilitating both detection and image reconstruction. Wang *et al.* [45] and Fang *et al.* [9] utilized frequency decomposition to improve generalization across datasets by capturing distinct attack patterns. Pan *et al.* [33] incorporated Grad-CAM visualization and textual explanations to enhance interpretability through attention-based training.

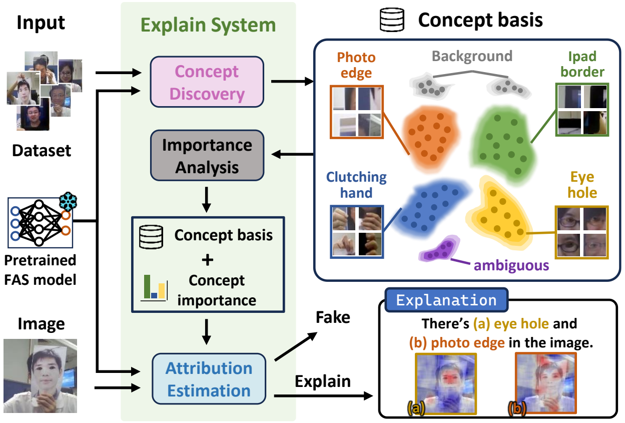

Figure 2. The overall pipeline of the proposed SPTD framework is depicted in this figure. SPTD include three parts: concept discovery, importance analysis and attribution estimation.

Although these methods offer objective explanations that are faithful to the model, the need to train from scratch limits their practical applicability. Recently, Zhang *et al.* proposed I-FAS [48], which incorporates a Large Language Model into the FAS process to generate textual explanations. However, since the generated explanation is still part of the black-box model's output, it carries the same unreliability and does not enhance the trustworthiness of the FAS system. Based on the above, the method proposed in this paper ensures the faithfulness of the generated explanations through XAI methods, without the need for additional training. This enables SPTD to effectively enhance the practical trustworthiness of the FAS system for users.

3. Method

In Face Anti-Spoofing (FAS) tasks, physical attack samples typically contain multiple spoof traces. Therefore, we propose SPTD (SPoof Trace Discovery) to provide a user-friendly explanation, which can necessarily extract various spoof traces from a single attack sample. Given a well trained FAS model, we first discover spoof trace concepts from a group of selected attack data. Secondly, we analyze importance of each concepts through perturbations. Finally, given a single spoof sample, SPTD show activated spoof trace concepts and mark their regions respectively. Thus, we separate SPTD into three parts which are concept discovery, importance analysis and attribution estimation. The whole pipeline can be seen in Figure 2.

3.1. Preliminaries

Consider a general supervised FAS task setting, where the original dataset $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \in \mathcal{X}^N \in \mathbb{R}^{N \times D}$ contains N input images and $(y_1, \cdots, y_n) \in \mathcal{Y}^N$ their associated labels. We are given a well trained predictor $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ which maps the input $\boldsymbol{x}$ to the predicted class $y = \boldsymbol{f}(\boldsymbol{x})$. We decompose the neural network $\boldsymbol{f}$ into two components $\boldsymbol{g}$ and $\boldsymbol{h}$ where $\boldsymbol{g}$ maps input $\boldsymbol{x}$ to intermediate logits $\boldsymbol{g}(\boldsymbol{x})$ and the second maps the intermediate logtis $\boldsymbol{g}(\boldsymbol{x})$ to output $\boldsymbol{h}(\boldsymbol{g}(\boldsymbol{x}))$. The original function $\boldsymbol{f}$ can be reconstructed as $\boldsymbol{f} = \boldsymbol{h} \circ \boldsymbol{g}$. The decomposition of $\boldsymbol{f}$ can occur at any layer of the network, though it is often chosen to be the last layer before classifier, as it contains more semantic information.

3.2. Concept discovery

The process of concept discovery is illustrated in Figure 3 with **pink arrow**. Firstly, we select a subset of images $\mathcal{X}_{sub} \in \mathbb{R}^{N' \times D}$ from the original dataset $\mathcal{X}^N$ for concept discovery, ensuring that the selected images share common characteristics, such as all being attack samples or belonging to a specific spoof type (e.g., print attacks). In this paper, we randomly sample r frames from each attack video in $\mathcal{X}^N$ to form $\mathcal{X}_{sub}$ where r is a hyper parameter. We assume $\pi(\cdot)$ is a filter function to create candidate spoof traces. It can be a straightforward crop and resize function to create sub-regions candidates. In the SPTD method, $\pi(\cdot)$ generates candidate spoof traces by uniformly sampling patches in both vertical and horizontal directions. Feed $\mathcal{X}_{sub}$ into $\pi(\cdot)$ to obtain an auxiliary dataset $\mathbf{X} \in \mathbb{R}^{N_a \times D}$ which contains all candidate spoof traces.

To automatically discover spoof trace concepts from auxiliary dataset $\mathbf{X}$, we feed it to the network to obtain activation $\mathbf{A} = \boldsymbol{g}(\mathbf{X}) \in \mathbb{R}^{N_a \times hw \times C}$ where hw indicates the shape of activation map and C indicates the number of channel. We apply Semi-NMF (Semi Non-negative Matrix Factorization) [5] to factorize activation maps since the non-negative constraint on the coefficients brings better interpretability while remain the capability to process negative values. Semi-NMF decompose the average pooled activations $\bar{\mathbf{A}} = AvgPool(\mathbf{A}) \in \mathbb{R}^{N_a \times C}$ into a product of concept coefficients $\mathbf{U} \in \mathbb{R}^{N_a \times K}$ and concept basis $\mathbf{W} \in \mathbb{R}^{C \times K}$ by solving:

$$(\mathbf{U}, \mathbf{W}) = \underset{\mathbf{U} \geq 0, \mathbf{W}}{\arg\min} \|\bar{\mathbf{A}} - \mathbf{U}\mathbf{W}^{\intercal}\|_F^2, \tag{1}$$

where K indicates the number of concepts one wish to discover and $\| \cdot \|_F^2$ denotes the Frobenius norm.

Following Ding *et al.* [5], we solve the above objective by iteratively updating $\mathbf{U}$ and $\mathbf{W}$. Specifically, $\mathbf{W}$ is the discovered spoof trace concepts where each column $\mathbf{W}_k \in \mathbb{R}^C$ corresponds to a single spoof trace concept. These concepts will be utilized in the subsequent importance analysis and attribution estimation process.

3.3. Importance analysis

The process of importance analysis is illustrated in Figure 3 with **gray arrow**. We adopt sobol indices [39] to estimate the importance of each spoof trace concept. Given the selected N' images $\mathcal{X}_{sub}$ and the discovered concepts

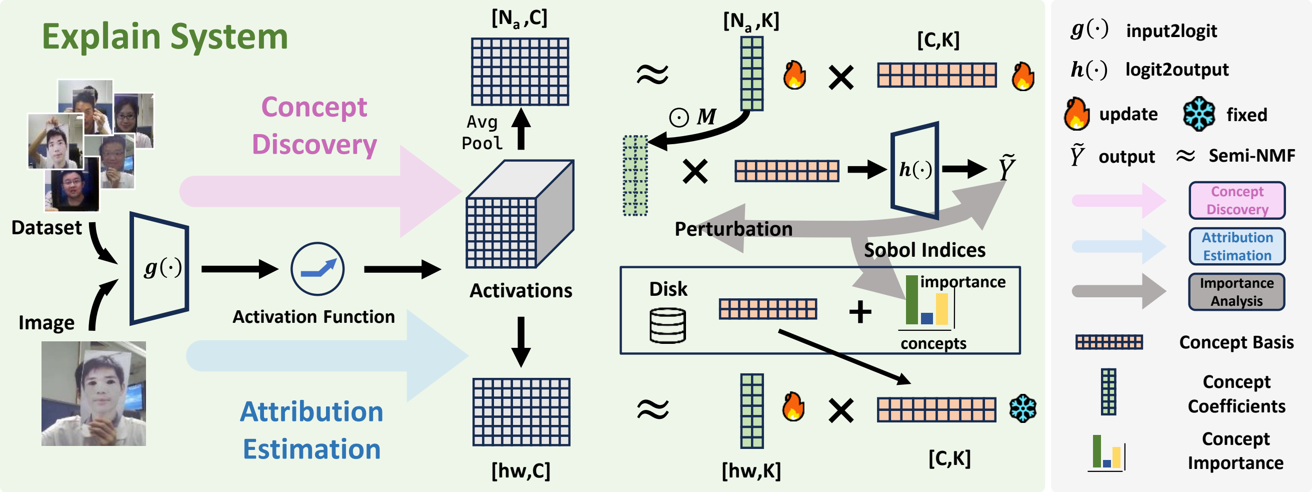

Figure 3. The detailed explain system of SPTD is illustrated above. We first identify concepts from a given dataset to construct a concept basis. Next, we assess the importance of each concept by perturbing its corresponding coefficients. Finally, leveraging the discovered concept basis and importance scores, SPTD estimates the attribution for a given image.

basis $\mathbf{W}$. We first feed images from $\mathcal{X}_{sub}$ to the network to obtain activations $\mathbf{A} = \boldsymbol{g}(\mathcal{X}_{sub}) \in \mathbb{R}^{N' \times hw \times C}$. Then we decompose the activation in each position into several concept coefficients $\mathbf{U}$ corresponding to the basis $\mathbf{W}$ by solving similar objective in Equation 1:

$$(\mathbf{U}^{(i,j)}, \mathbf{W}) = \underset{\mathbf{U}^{(i,j)} \geq 0, \mathbf{W}}{\arg\min} \ \|\mathbf{A}^{(i,j)} - \mathbf{U}^{(i,j)}\mathbf{W}^\intercal\|_F^2, \qquad (2)$$

where $\mathbf{W}$ is fixed based on the values obtained during the concept discovery process, and $\mathbf{A}^{(i,j)} \in \mathbb{R}^{N' \times C}$ represents the feature vector at position (i,j) within the spatial dimensions (h,w). The result $\mathbf{U}^{(i,j)} \in \mathbb{R}^{N' \times K}$ indicates the concept coefficients in position (i,j).

Formally, a common way to estimate the importance of a concept k is to measure the variations of the model's output $\boldsymbol{h}(\mathbf{U}\mathbf{W}^\intercal)$ when concept coefficient $\mathbf{U}_{(1,k)}, \cdots, \mathbf{U}_{(N',k)}$ undergo meaningful perturbations. We generate random perturbation mask $\mathbf{M} \sim \mathcal{U}[0,1]^K$ and reconstruct the perturbed activation $\widetilde{\mathbf{A}} = (\mathbf{U} \odot \mathbf{M})\mathbf{W}^\intercal$. Thus, the perturbed output can be denoted as $\widetilde{\mathbf{Y}} = \boldsymbol{h}(\widetilde{\mathbf{A}})$. Simply understanding, the model output will vary substantially when perturbing an important concept, while a less relevant concept will have little to no impact. The importance of concept k can be written as:

$$\mathcal{S}_k = \frac{\mathbb{E}_{\mathbf{M}_{\sim k}}(\mathbb{V}_{\mathbf{M}_k}(\widetilde{\mathbf{Y}}|\mathbf{M}_{\sim k}))}{\mathbb{V}(\widetilde{\mathbf{Y}})}, \qquad (3)$$

where $\mathbb{E}$ indicates expectation and $\mathbb{V}$ indicates variance. We use Sobol Sequence as the random generator of perturbation mask $\mathbf{M}$.

3.4. Attribution estimation

The process of attribution estimation is illustrated in Figure 3 with blue arrow. After concept discovery, we get K concept basis $\mathbf{W} \in \mathbb{R}^{C \times K}$ from the the selected subset of N' images, which serves as an approximation of the spoof trace concepts present in the original dataset containing N images. Given an image $\boldsymbol{x}$ (from the original dataset or any other data source), we factorize the activations $\mathbf{A} = \boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^{hw \times C}$ with the fixed concept basis $\mathbf{W}$ through Equation 2 and get the corresponding concept coefficients $\mathbf{U} \in \mathbb{R}^{hw \times K}$. Specifically, we can regard $\mathbf{U}_k^{(x,y)}$ as the importance of concept k at (x,y) position of the activation map and the $\mathbf{U}_k$ can be seen as an activation map of concept k. In this way, we factorize a single input image into several concepts and its activation map. With the help of concept coefficients $\mathbf{U}$, we can use previous attribution method to enhance the estimation of concept attribution. We introduce C-RISE (Concept RISE) which modified RISE (Randomized Input Sampling for Explanation) [34] to work for concept attribution.

RISE [34] treats the target model as a black box, requiring only input-output pairs without needing access to the internal inference process. Given an image $\mathbf{x}$, RISE applies random masks $\mathbf{M}$ to each pixel and computes the expected output influence under perturbations:

$$S(\mathbf{x}) = \mathbb{E}_{\mathbf{M}}\left[\mathbf{M} \cdot \mathbf{f}(\mathbf{x} \odot \mathbf{M})\right] \qquad (4)$$

where $\odot$ denotes element-wise multiplication.

To estimate the pixel importance of a specific concept k, we focus on the coefficients $\mathbf{U}_k$ instead of the classification logits. The estimation in C-RISE can be formulated as:

$$S_k(\mathbf{x}) = \mathbb{E}_{\mathbf{M}}\left[\mathbf{M} \cdot \mathrm{Avg}(\mathbf{Fac}_k(\mathbf{g}(\mathbf{x} \odot \mathbf{M})))\right] \qquad (5)$$

where $\mathbf{Fac}_k$ denotes Semi-NMF factorization result of concept k and $\mathbf{Avg}$ represents average pooling along the spatial dimension.

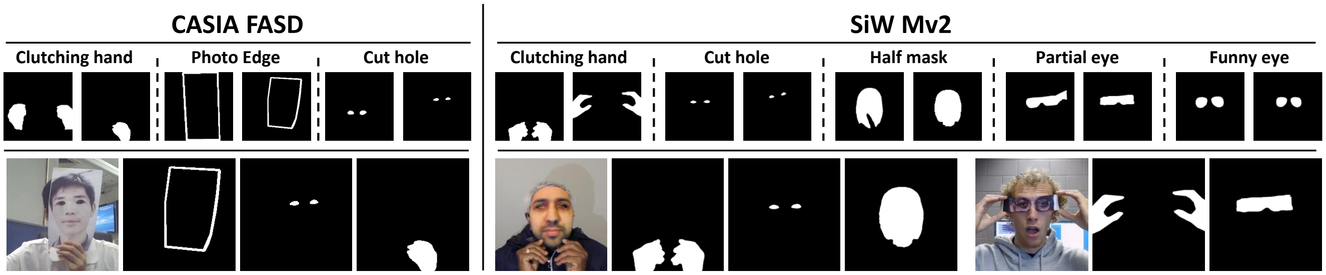

Figure 4. Sample visualization of X-FAS benchmark. We show samples of annotation mask and attack images with multiple spoof traces.

Table 1. Detailed information of the proposed X-FAS benchmark.

Dataset	Spoof Type	Subjects	Number of frame	Spoof Traces	Number of mask
CASIA-FASD	Print	16	16	Photo edge	16
				Cut hole	8
				Clutching hand	16
	Replay	8	8	Ipad border	8
				Clutching hand	8
	Total	24	24	All	56
SiW-Mv2	Makeup Impersonation	1	2	Eye brows	2
				Makeup eye	2
	Makeup Obfuscation	9	18	Makeup mark	18
	Mannequin	40	80	Model_head	80
	Half Mask	66	66	Half mask	66
				Cut hole	63
				Clutching hand	31
	Paper Mask	17	33	Paper mask	34
				Clutching hand	34
	Transparent Mask	58	58	Transparent Mask	58
				Cut hole	58
				Clutching hand	3
	Partial Eye	56	104	Partial eye	104
				Clutching hand	104
	Funny Eye Glasses	176	176	Funny eye	176
				Funny Eyeball	14
	Partial Mouth	29	58	Paper mouth	58
				Clutching hand	58
	Paper Glasses	75	132	Paper glass	132
				Clutching hand	3
	Silicone	14	26	Silicon mask	26
				Cut hole	26
	Total	541	753	All	1150

4. Benchmark

In order to evaluate the explanation quality generated by X-FAS methods, we use expert annotated spoof traces that can repeatably evaluate multiple explainable methods in an unbiased way. Following the annotation manner of Kondapaneni et al. [22] which proposed an expert-defined birds feature dataset to evaluate explainable methods for birds classification, we introduce an X-FAS benchmark for testing X-FAS methods which can measure the accuracy of generated explanation at a fine-grained level.

4.1. Fine-grained explanation dataset

The benchmark data is spoof images from the CASIA-FASD [49] and SiW-Mv2 [13] dataset. The detailed information of the proposed X-FAS benchmark is illustrated in Table 1. CASIA-FASD consists of two attack types: print attacks and replay attacks, while SiW-Mv2 includes a diverse set of attack types, such as various mask attacks and model head attacks. For both CASIA-FASD and SiW-Mv2, we automatically select frames from original videos using a pretrained CLIP [36] model, following the process outlined in Algorithm 1. Our approach aims to maximize the dissimilarity between sampled frames from each video while ensuring that a face is detected in every selected frame. By analyzing obvious shared spoof traces with human experts' knowledge, we annotate the regions of each spoof traces, finally forming 1206 spoof trace masks of 777 attack images in total to produce targets of X-FAS methods. Samples in the X-FAS benchmark are visualized in Figure 4.

Algorithm 1 Filter frames from video

```
Input: Video $\mathcal{V} = \{F_1, F_2, \cdots, F_N$, where $F_i$ indicates the $i$
    th frame$\}$, Frame number $l$, Random sample iteration $Iter$,
    Pretrained CLIP model $\mathcal{M}_{clip}$, Retina face detection model
    $\mathcal{M}_{face}$
Output: Selected frames $\mathcal{S}$
$\mathcal{V}'$ = $\{F_1, F_2, \cdots, F_{N'}$, where $M_{face}(F_i)$ detects a face$\}$
Extract CLIP Feature $\mathcal{E}$ of all frames from $\mathcal{V}'$ :
    $\mathcal{E} = \mathcal{M}_{clip}(\mathcal{V}') \in \mathbb{R}^{N \times D}$
$max\_sim \leftarrow -\infty$
$L \leftarrow None$
for all $t = 1, \cdots, Iter$ do
    $sim \leftarrow 0$
    Random choose $l$ frames from $\mathcal{V}'$ :
    $L^{sample} = random\_choice(F_1, F_2, \cdots, F_{N'})$
    for all $i = 1, \cdots, l-1$ do
        for all $j = i+1, \cdots, l$ do
            $sim+ = 1 - \frac{\mathcal{E}_{L_i^{sample}} \cdot \mathcal{E}_{L_j^{sample}}}{||\mathcal{E}_{L_i^{sample}}|| \cdot ||\mathcal{E}_{L_j^{sample}}||}$
        end for
    end for
    if $max\_sim < sim$ then
        $max\_sim = sim$
        $L = L_{sample}$
    end if
end for
$\mathcal{S} \leftarrow \{F_{L_1}, F_{L_2}, \cdots, F_{L_l}\}$
return $\mathcal{S}$
```

4.2. Evaluation protocol

In the X-FAS benchmark, CASIA-FASD and SiW-Mv2 subsets should be evaluated separately. To ensure reliable testing, the model used should perform well on the target dataset (e.g., when evaluating the SiW-Mv2 subset, the pretrained model should be trained using the SiW-Mv2 intra-protocol). With a well-trained model, explanations can be

Table 2. Quantitative result on X-FAS benchmark. When calculating IoU and nIoU, we consider the top 30% pixels as the explanation mask. The evaluation protocol follows [22].

Dataset	Method		GradCAM	GradCAM++	EigenGradCAM	AblationCAM	RandomCAM	RISE	SPTD (Ours)
CASIA-FASD	Print	IoU	0.0648	0.0622	0.0681	0.0968	0.0585	0.0801	**0.1078**
		nIoU	0.3354	0.3811	0.4268	0.4688	0.2585	0.3707	**0.6141**
	Replay	IoU	0.1530	0.1023	0.0606	0.1598	0.0962	0.1840	**0.2244**
		nIoU	0.2545	0.1358	0.1169	0.2882	0.1959	0.4048	**0.4837**
	Average	IoU	0.0942	0.0756	0.0656	0.1178	0.0760	0.1110	**0.1467**
		nIoU	0.3084	0.2993	0.3235	0.4086	0.3026	0.3679	**0.5706**
SiW-Mv2	Makeup Impersonation	IoU	**0.0295**	**0.0295**	**0.0295**	**0.0295**	0.0171	**0.0295**	**0.0295**
		nIoU	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.5103	**1.0000**	**1.0000**
	Makeup Obfuscation	IoU	0.1211	**0.1319**	**0.1319**	0.1243	0.0568	0.1019	0.1318
		nIoU	0.9156	0.9999	**1.0000**	0.9372	0.4123	0.7423	0.9991
	Mannequin	IoU	0.4317	0.4580	0.5019	0.4595	0.2828	0.3923	**0.5173**
		nIoU	0.7028	0.7478	0.8105	0.7481	0.4686	0.6405	**0.8453**
	Half Mask	IoU	0.2189	0.2418	0.2494	0.2318	0.0815	0.1648	**0.2629**
		nIoU	0.8243	0.8678	0.8803	0.8583	0.3185	0.7429	**0.9619**
	Paper Mask	IoU	0.2373	0.2362	0.2422	0.2322	0.1022	0.1497	**0.2802**
		nIoU	0.6297	0.6110	0.6663	0.6153	0.3142	0.5631	**0.9296**
	Transparent Mask	IoU	0.1805	0.2750	0.2783	0.2251	0.1255	0.1423	**0.2795**
		nIoU	0.7060	0.9830	0.9887	0.8409	0.4927	0.5741	**0.9922**
	Partial Eye	IoU	0.1358	0.1271	0.1344	0.1517	0.0853	0.1917	**0.2122**
		nIoU	0.6298	0.6153	0.6317	0.6691	0.3199	0.7533	**0.8147**
	Funny Eye Glasses	IoU	0.1077	0.1121	0.1120	0.1080	0.0601	0.1059	**0.1123**
		nIoU	0.9448	0.9977	0.9950	0.9513	0.5296	0.9330	**0.9985**
	Partial Mouth	IoU	0.1885	0.1690	0.1679	0.1951	0.0962	0.2007	**0.2438**
		nIoU	0.6971	0.6283	0.6225	0.7194	0.3312	0.7218	**0.8852**
	Paper Glasses	IoU	0.0914	0.0927	0.0927	0.0927	0.0406	0.0914	**0.0944**
		nIoU	0.9742	0.9924	0.9924	0.9811	0.4223	0.9598	**0.9964**
	Silicone	IoU	0.2082	0.2393	**0.2738**	0.2219	0.1192	0.1822	0.2711
		nIoU	0.7236	0.7708	0.8009	0.7332	0.3631	0.6508	**0.8119**
	Average	IoU	0.1739	0.1859	0.1939	0.1848	0.0974	0.1645	**0.2154**
		nIoU	0.8103	0.8516	0.8644	0.8388	0.4306	0.7868	**0.9345**

generated using X-FAS methods, and their quality can be assessed by computing metrics that compare the generated explanations to the ground truth annotations for each image in the benchmark. If multiple annotation masks exist for an image, the final metric should be computed as the average across all masks.

4.3. Evaluation metric

Intersection over Union (IoU) is a widely used metric and can also be applied in the X-FAS benchmark. However, directly averaging the IoU of multiple spoof traces is unfair, as the theoretical maximum IoU varies due to differences in the pixel count of ground truth annotation masks. To address this, we propose a fairer metric derived from IoU, termed *normalized Intersection over Union* (nIoU). Given a annotated mask $\mathbf{M}_G$ and an explanation $\mathbf{M}_I$, we first obtain a processed explanation mask $\mathbf{M}_I^x$ by setting the top x percent of values to 1 and the rest to 0. The nIoU metric is then formulated as follows:

$$nIoU(\mathbf{M}_G, \mathbf{M}_I, x) = \frac{IoU(\mathbf{M}_G, \mathbf{M}_I^x) * max(x, y)}{min(x, y)} \in [0, 1] \quad (6)$$

where y is the useful pixel percentage of annotated traces $\mathbf{M}_G$. $\frac{min(x,y)}{max(x,y)}$ is the optimum value of $IoU(\mathbf{M}_G, \mathbf{M}_I^x)$

5. Experiments

To demonstrate the effectiveness of SPTD, we primarily evaluate it on the Face Anti-Spoofing (FAS) task using the proposed X-FAS benchmark. We compare SPTD with several representative XAI methods, analyze the discovered spoof concepts, and visualize fine-grained explanations to highlight its strengths in interpreting FAS models. In addition, to further verify the fidelity of SPTD, we conduct supplementary experiments on a general vision task (ImageNet classification) by comparing it with CRAFT [10]. This evaluation provides an objective perspective on how well SPTD reflects the model's true decision-making process.

5.1. Qualitative Results on X-FAS Benchmark

5.1.1 Baselines

We consider three categories of XAI methods: gradient-based methods, perturbation-based methods, and concept-based methods. Gradient-based methods include GradCAM [38], GradCAM++ [2], EigenGradCAM [31], AblationCAM [37], RandomCAM (code from [12]). For the perturbation-based category, we adopt RISE [34] as a representative method. The concept-based method is SPTD proposed in this paper. Following the evaluation protocol in [22], we calculate mean IoU and mean nIoU (detail in Section 4.3) metric on different spoof types of X-FAS benchmark.

5.1.2 Experimental settings

We adopt the widely recognized FLIP [41] method and train two FLIP-V models separately on the CASIA-FASD [49] and SiW-Mv2 [13] datasets. The model trained on CASIA-FASD achieves an HTER of 0.11% on the test split, while the model trained on SiW-Mv2 achieves 2.49%, demonstrating their strong performance. All previous XAI methods, along with the proposed SPTD method, are evaluated using these two well-trained FLIP-V models. For all XAI methods, we select the last ResBlock as the target layer. In SPTD, we set the number of concepts to $K = 15$ and use vanilla estimation. The subset for concept discovery and importance analysis is generated by randomly selecting two frames from each video in the original dataset. For a fair comparison, we use vanilla attribution, which has been employed by previous XAI methods. Vanilla attribution follows Collins *et al.* [4] by marking the attention regions of each concept through scaling the activation maps (keeping 10% of the maximum value and setting the rest to zero) and mapping them back to the input shape, as activation maps preserve spatial correlations with the input image.

5.1.3 Results

Table 2 presents the overall results on the X-FAS benchmark. The results indicate that among all previous XAI methods, AblationCAM and RISE outperform other methods on the CASIA-FASD dataset, while EigenGradCAM achieves the best results on the SiW-Mv2 dataset. EigenGradCAM excels in three test items related to Makeup and Silicone, whereas our proposed SPTD method outperforms on the remaining nineteen scenarios. Furthermore, in terms of average IoU and nIoU, SPTD achieves the highest results among all methods. These findings demonstrate that SPTD surpasses previous XAI methods on X-FAS benchmark, highlighting the superior quality of its explanations.

5.2. Visualization and Analysis

With the same experimental settings in Section 5.1, we visualize the discovered concepts and the generated explanations of SPTD on CASIA-FASD [49] dataset.

5.2.1 Spoof Concept and Explanation Visualization

We visualize the discovered top four important spoof concepts on CASIA-FASD dataset with their analyzed importance in Figure 5. As we can see, the discovered concepts can be easily understood by users since they are represented as multiple patches. We further attempt to summarize these four discovered concepts and find clear semantic meanings where c3, c4, c7 and c13 expresses *clutching hand*, *cut hole*, *photo edge* and *iPad border* respectively.

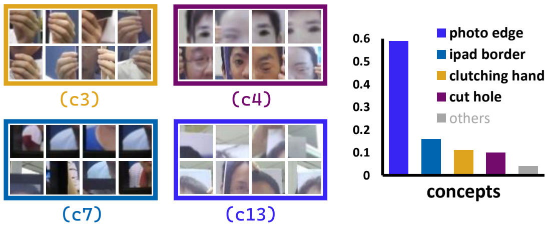

Figure 5. Discovered spoof concepts in CASIA-FASD dataset. Concept c3, c4, c7 and c13 are top four important concepts.

With the help of these discovered concepts, SPTD can generate fine-grained explanations which mark attention regions of activated corresponding concepts using C-RISE attribution estimation mentioned in Section 3.4, as shown in Figure 6. In the explanation of Figure 6(a), SPTD find activated concepts c3, c4, c13 and mark the specific attention region in red which is consistent with concepts in Figure 5. While in Figure 6(b), SPTD only find activated concept c3 and c7.

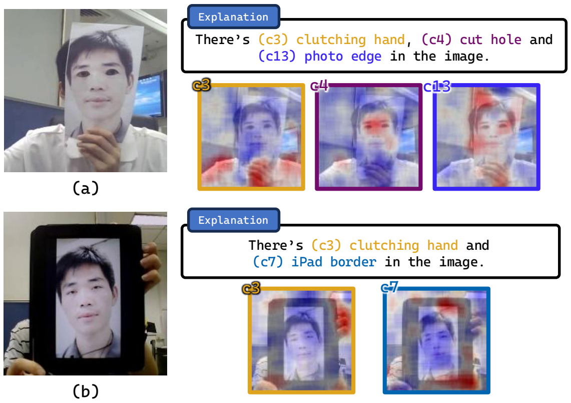
Explanation
There's (c3) clutching hand, (c4) cut hole and (c13) photo edge in the image.
c3
c4
c13
(a)
Explanation
There's (c3) clutching hand and (c7) iPad border in the image.
c3
c7
(b)

Figure 6. Explanations on CASIA-FASD samples of concepts in Figure 5. Sample (a) activated concept c3 (*clutching hand*), c4(*cut hole*) and c13 (*photo edge*) while (b) activated c3 (*clutching hand*) and c7 (*iPad border*). Heatmaps show pixel level attention region of each activated concept.

The result shows that SPTD has the ability to discover spoof concepts which are easy to be understood by users and provide corresponding attention regions of each concept on top of face anti-spoofing models.

5.2.2 Explanation Comparison

Figure 7 shows the visualization comparison between multiple XAI methods and SPTD (using vanilla estimation) on CASIA-FASD [49] dataset. Previous XAI methods show a single heatmap where some cover the whole face missing finer-level information and some pay attention to partial spoof traces. In contrast, SPTD provides multiple heatmaps where each of them corresponds to a specific activated spoof

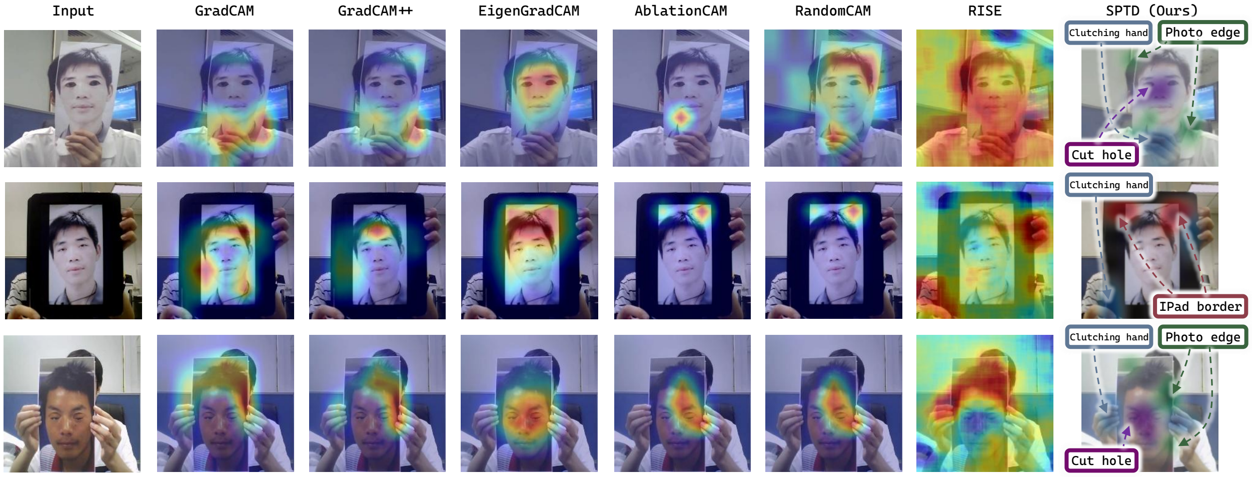

Figure 7. Visualization of multiple XAI methods on CASIA-FASD samples. We compare multiple previous XAI methods with SPTD, previous methods give a single heatmap while SPTD gives multiple attention regions of corresponding activated spoof concepts.

concept. These results prove that with the help with SPTD explanations, we can surely increase the trustworthiness to face anti-spoofing system users.

5.3. Fidelity Evaluation

To further demonstrate the fidelity of SPTD, we adopt two metrics to compare SPTD and CRAFT [10] (both are concept-based method) on several random selected classes of ImageNet. *Deletion* and *Insertion* are a pair of the most popular evaluation methods to check an explanation's fidelity. *Deletion* measures the drop in model confidence by progressively removing the most important features. A stronger explanation is expected to cause a larger drop in confidence. In contrast, *Insertion* evaluates the confidence increase by gradually adding the same features back where larger confidence increase is preferred. For both Deletion and Insertion metrics, we calculate the area under the curve (AUC) which represents the quality of explanation method. A lower deletion AUC and higher insertion AUC is appreciated.

We compare SPTD and CRAFT on class 101 and 413 of ImageNet using the same pretrained ResNet18 model. Since CRAFT also analyze importance of each discovered concepts, the top three important concepts are fetched to calculate fidelity metrics.

Table 3. Fidelity results on class 101 and 413 of ImageNet. Top@k indicates the k th important concept. *Ins* and *Del* indicates *Insertion* and *Deletion* metric correspondingly.

Class	metrics	top@k	CRAFT	SPTD (Ours)	Class	metrics	top@k	CRAFT	SPTD (Ours)
101	Ins (↑)	Top1	0.46335	**0.48351**	413	Ins (↑)	Top1	0.48587	**0.48697**
		Top2	0.44794	**0.48701**			Top2	0.38627	**0.45782**
		Top3	0.36773	**0.41768**			Top3	0.31657	**0.43796**
		Avg	0.42634	**0.46273**			Avg	0.39624	**0.46092**
	Del (↓)	Top1	0.23237	**0.21507**		Del (↓)	Top1	0.14864	**0.14722**
		Top2	0.24420	**0.21876**			Top2	0.23849	**0.17079**
		Top3	0.30649	**0.30107**			Top3	0.30824	**0.19949**
		Avg	0.26102	**0.24497**			Avg	0.23179	**0.17250**

Table 3 presents the quantitative results of CRAFT and SPTD. As shown, SPTD outperforms CRAFT in terms of fidelity across all settings, indicating that SPTD is more effective in discovering concepts and estimating attributions. Notably, CRAFT is only applicable when activation maps are non-negative. In contrast, SPED is free from this constraint, and the results demonstrate its ability to generate superior explanations without relying on such restrictions.

6. Conclusion

In this paper, we propose a new problem termed X-FAS to provide reliable face anti-spoofing results by generating explanations on top of face anti-spoofing classification results to cope with the vulnerability of black-box models. We introduce SPTD (SPoof Trace Discovery), an X-FAS method which can discover spoofing concepts that is easy to be understood by users and provide a heatmap of activated concepts of attack images. To evaluate the quality of X-FAS methods, we present an X-FAS benchmark with expert annotations on two FAS datasets. In our experiments, both quantitative and qualitative results show the efficacy and reliability of SPTD. We hope that this work will guide further efforts in the research for X-FAS which can eliminate user's doubts of face anti-spoofing models and make it more transparent, trustworthy and effective.

Acknowledgement

This work was supported in part by Chinese National Natural Science Foundation Projects U23B2054, 62276254, 62306313, the Beijing Science and Technology Plan Project Z231100005923033, Beijing Natural Science Foundation L221013, the Science and Technology Development Fund of Macau Project 0140/2024/AGJ, and InnoHK program.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[2] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[3] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.

[4] E. Collins, R. Achanta, and S. Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352, 2018.

[5] C. H. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.

[6] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

[9] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3722–3731, 2022.

[10] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, and T. Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023.

[11] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[12] J. Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[13] X. Guo, Y. Liu, A. Jain, and X. Liu. Multi-domain learning for updating face anti-spoofing models. In *European conference on computer vision*, pages 230–249. Springer, 2022.

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

[15] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *European conference on computer vision*, pages 37–54. Springer, 2022.

[16] P.-K. Huang, C.-H. Chiang, T.-H. Chen, J.-X. Chong, T.-L. Liu, and C.-T. Hsu. One-class face anti-spoofing via spoof cue map-guided feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 277–286, 2024.

[17] P.-K. Huang, C.-H. Chiang, J.-X. Chong, T.-H. Chen, H.-Y. Ni, and C.-T. Hsu. Ldcformer: Incorporating learnable descriptive convolution to vision transformer for face anti-spoofing. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 121–125. IEEE, 2023.

[18] P.-K. Huang, M.-C. Chin, and C.-T. Hsu. Face anti-spoofing via robust auxiliary estimation and discriminative feature learning. In *Asian Conference on Pattern Recognition*, pages 443–458. Springer, 2021.

[19] P.-K. Huang, J.-X. Chong, H.-Y. Ni, T.-H. Chen, and C.-T. Hsu. Towards diverse liveness feature representation and domain expansion for cross-domain face anti-spoofing. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1199–1204. IEEE, 2023.

[20] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[21] A. Jourabloo, Y. Liu, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 290–306, 2018.

[22] N. Kondapaneni, M. Marks, O. Mac Aodha, and P. Perona. Less is more: Discovering concise network explanations. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

[23] A. Liu, H. Ma, J. Zheng, H. Yuan, X. Yu, Y. Liang, S. Escalera, J. Wan, and Z. Lei. Fm-clip: Flexible modal clip for face anti-spoofing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8228–8237, 2024.

[24] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1179–1187, 2021.

[25] A. Liu, Z. Tan, Z. Yu, C. Zhao, J. Wan, Y. Liang, Z. Lei, D. Zhang, S. Z. Li, and G. Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4775–4786, 2023.

[26] A. Liu, S. Xue, J. Gan, J. Wan, Y. Liang, J. Deng, S. Escalera, and Z. Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 222–232, 2024.

[27] Y. Liu, Y. Chen, M. Gou, C.-T. Huang, Y. Wang, W. Dai, and H. Xiong. Towards unsupervised domain generalization for

face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20654–20664, 2023.

[28] Y. Liu, J. Stehouwer, and X. Liu. On disentangling spoof trace for generic face anti-spoofing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 406–422. Springer, 2020.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[31] M. B. Muhammad and M. Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[32] P. C. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso. Causality-inspired taxonomy for explainable artificial intelligence. *arXiv preprint arXiv:2208.09500*, 2022.

[33] S. Pan, S. Hoque, and F. Deravi. An attention-guided framework for explainable biometric presentation attack detection. *Sensors*, 22(9):3365, 2022.

[34] V. Petsiuk. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[35] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, and E. Baralis. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[37] H. G. Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 983–991, 2020.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[39] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

[40] E. Solomon and K. J. Cios. Fass: Face anti-spoofing system using image quality features and deep learning. *Electronics*, 12(10):2199, 2023.

[41] K. Srivatsan, M. Naseer, and K. Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19685–19696, 2023.

[42] A. Sun, P. Ma, Y. Yuan, and S. Wang. Explain any concept: Segment anything meets concept-based explanation. *Advances in Neural Information Processing Systems*, 36, 2024.

[43] J. Wang, J. Zhang, Y. Bian, Y. Cai, C. Wang, and S. Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2746–2754, 2021.

[44] X. Wang, K.-Y. Zhang, T. Yao, Q. Zhou, S. Ding, P. Dai, and R. Ji. Tf-fas: twofold-element fine-grained semantic guidance for generalizable face anti-spoofing. In *European Conference on Computer Vision*, pages 148–168. Springer, 2024.

[45] Y.-C. Wang, C.-Y. Wang, and S.-H. Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1955–1964, 2022.

[46] Z. Wang, C. Huang, and X. Yao. A roadmap of explainable artificial intelligence: Explain to whom, when, what and how? *ACM Transactions on Autonomous and Adaptive Systems*, 19(4):1–40, 2024.

[47] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5609–5631, 2023.

[48] G. Zhang, K. Wang, H. Yue, A. Liu, G. Zhang, K. Yao, E. Ding, and J. Wang. Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. *arXiv preprint arXiv:2501.01720*, 2025.

[49] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31, 2012.

[50] Q. Zhou, K.-Y. Zhang, T. Yao, X. Lu, R. Yi, S. Ding, and L. Ma. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20453–20463, 2023.