Wild Face Anti-Spoofing Challenge 2023: Benchmark and Results

Dong Wang [1] Jia Guo [2] Qiqi Shao [1] Haochi He [1] Zhian Chen [1]
Chuanbao Xiao [1] Ajian Liu [3] Sergio Escalera [4,5] Hugo Jair Escalante [6]
Zhen Lei [3] Jun Wan [3] Jiankang Deng [2*]
[1]MoreDian [2]InsightFace [3]CASIA
[4]Computer Vision Center (UAB) [5]University of Barcelona [6]INAOE
{wangdong,shaoqiqi,hehc,chenzhian,xiaocb}@moredian.com, {guojia,jiankangdeng}@gmail.com
sescalera@cvc.uab.cat, hugojair@inaoep.mx, {ajian.liu,zhen.lei,jun.wan}@ia.ac.cn

Abstract

Face anti-spoofing (FAS) is an essential mechanism for safeguarding the integrity of automated face recognition systems. Despite substantial advancements, the generalization of existing approaches to real-world applications remains challenging. This limitation can be attributed to the scarcity and lack of diversity in publicly available FAS datasets, which often leads to overfitting during training or saturation during testing. In terms of quantity, the number of spoof subjects is a critical determinant. Most datasets comprise fewer than 2,000 subjects. With regard to diversity, the majority of datasets consist of spoof samples collected in controlled environments using repetitive, mechanical processes. This data collection methodology results in homogenized samples and a dearth of scenario diversity. To address these shortcomings, we introduce the Wild Face Anti-Spoofing (WFAS) dataset, a large-scale, diverse FAS dataset collected in unconstrained settings. Our dataset encompasses 853,729 images of 321,751 spoof subjects and 529,571 images of 148,169 live subjects, representing a substantial increase in quantity. Moreover, our dataset incorporates spoof data obtained from the internet, spanning a wide array of scenarios and various commercial sensors, including 17 presentation attacks (PAs) that encompass both 2D and 3D forms. This novel data collection strategy markedly enhances FAS data diversity. Leveraging the WFAS dataset and Protocol 1 (Known-Type), we host the Wild Face Anti-Spoofing Challenge at the CVPR2023 workshop. Additionally, we meticulously evaluate representative methods using Protocol 1 and Protocol 2 (Unknown-Type). Through an in-depth examination of the challenge outcomes and benchmark baselines, we provide insightful analyses and propose potential avenues for future research.

The dataset is released under Insightface[1].

1. Introduction

In recent years, face recognition technologies [1, 13, 15, 16, 88] have become increasingly pervasive in various aspects of our lives, including access control, phone unlocking, digital payment, and attendance systems. Despite their widespread adoption, these systems remain susceptible to significant security risks. For instance, face recognition attendance systems are designed to enhance management efficiency; however, if these systems can be easily deceived by photographs, it would result in disarray in personnel management. Similarly, phone unlocking systems, if successfully attacked, may compromise user data security and tarnish the reputation of manufacturers. Malicious attackers may also exploit vulnerabilities in digital payment systems to steal others' identities for illegal purposes. Consequently, face anti-spoofing (FAS) technologies [33,34,36–38,63,80], a crucial component of automatic face recognition systems, has garnered considerable attention from both academia and industry.

Owing to the rapid progress of deep learning, the FAS technology community has witnessed a surge in outstanding contributions. Deep learning-based FAS methods can be categorized into three groups: 1) classification supervision, 2) auxiliary pixel-wise supervision, and 3) generative pixel-wise supervision. Intuitively, FAS tasks can be framed as classification problems, with many works [5,7,12,22,32,64] directly supervised using binary cross-entropy (CE), while others [53, 71] extend the problem to multiple classification. Classification supervisions are easy to construct, enabling deep FAS models to converge rapidly. In contrast, pixel-wise supervision with auxiliary tasks can ex-

*Corresponding author

[1]https://github.com/deepinsight/insightface/tree/master/challenges/cvpr23-fas-wild

tract more fine-grained cues, with additional information such as pseudo depth maps [65, 78, 80, 82], binary mask maps [43, 66, 74], and reflection maps [30, 81, 84] helping to delineate local live/spoof features. Generative pixel-wise supervisions [19,42,44,54,56], which do not rely on expert-designed guidance and offer more flexible labels, visualize spoof cues in spoof samples, thereby enhancing the interpretability of FAS tasks.

As highlighted in [84], existing FAS methods continue to face challenges in generalizing to real-world scenarios, particularly when employing unimodal RGB sensors without hardware advantages. These sensors are prone to various presentation attacks (PAs), including print, replay, and 3D-model attacks. In comparison to face recognition technology, face anti-spoofing remains an unresolved issue in face recognition systems [62, 76]. The success of recognition technology largely depends on the availability of large-scale, diverse datasets. The 2016 release of the MS1M [23] dataset marked a turning point in the rapid development and industrial application of face recognition algorithms.

The relatively slow progress of face anti-spoofing technology can be attributed to the limitations in the quantity and diversity of publicly accessible FAS datasets, which often leads to overfitting during training or saturation during testing. FAS datasets typically comprise several key components: the number of subjects, presentation attacks (PAs), scenarios, and input sensors. The scale of a dataset is primarily determined by the number of subjects; however, most existing FAS datasets include fewer than 2,000 subjects, with only one dataset (CelebA-Spoof) containing over 10,000 subjects. Moreover, CelebA-Spoof has an average of more than 60 images per subject, which results in high data homogeneity and adversely impacts dataset diversity.

In terms of diversity, the remaining elements play a crucial role. PAs can be broadly classified into 2D and 3D forms based on their geometric properties. 2D PAs display facial identity information to sensors through photos or videos, with common attack variants including flat or wrapped printed photos, cut-out photos, images displayed on screens, and video replays. With advancements in 3D manufacturing technology, 3D masks and models have emerged as new PAs that challenge FAS technology. In comparison to traditional 2D PAs, 3D attacks exhibit greater realism in terms of texture and geometric structure. Rigid 3D masks can be made from various materials, such as paper, resin, plaster, or plastic, while flexible soft 3D masks typically consist of silicone or latex. 3D models often exhibit high levels of simulation, including waxworks and adult dolls.

A review of current datasets, as shown in Table 1, reveals that most datasets contain either 2D or 3D PAs exclusively. Datasets such as [4, 9–11, 50, 52, 58, 68, 83] feature 2D print or display PAs, while [18, 20, 28, 39, 41, 57, 60, 79] include 3D mask or model PAs. Some datasets, such as [31, 43, 84], encompass both 2D and 3D PAs; however, their diversity remains unsatisfactory due to deficiencies in other elements, such as the absence of high-fidelity 3D models. Additionally, the spoof samples in nearly all current datasets are collected in controlled and limited scenarios through mechanical and repetitive processes, which we refer to as manually controlled scenarios. This data collection approach results in a lack of scenario richness and leads to significant sample homogenization.

To address the limitations of existing face anti-spoofing datasets, we introduce the Wild Face Anti-Spoofing Dataset (WFAS), a large-scale FAS dataset collected in the wild. To the best of our knowledge, this is the first dataset to extend FAS research to real-world scenarios. Our dataset comprises 853,729 images of 321,751 spoof subjects and 529,571 images of 148,169 live subjects, significantly increasing the quantity of available data. Furthermore, the spoof data is sourced from the internet, encompassing a wide variety of scenarios and commercial sensors.

The internet-derived spoof samples, though not intended to attack face recognition systems, incidentally resemble spoof samples that benefit FAS research. Examples include 2D faces appearing in picture books or on TV screens, as well as 3D waxwork faces at tourist attractions. Our dataset includes 17 PAs, covering both 2D and 3D forms. The 2D PAs comprise print types (*e.g.*, newspapers, posters, photos, albums, picture books, scanned photos, packaging, and cloth) and display types (*e.g.*, phones, tablets, TVs, and computers). The 3D PAs feature five subcategories with varying fidelity: masks, garage kits, dolls, adult dolls, and waxworks.

All spoof samples were captured using photographic equipment, such as various mobile phone brands, digital cameras, and scanners. These optical sensors produce images with a wide range of resolutions, resulting in a richer face quality within our dataset. The live face set, collected from the internet under specific creative commons licenses, is a typical in-the-wild dataset, incorporating diverse scenarios, races, ages, and more. Our dataset significantly enhances FAS data diversity, as illustrated by the spoof samples in Figure 1.

Leveraging our dataset and Protocol 1 (Known Type), we hosted the Wild Face Anti-Spoofing Challenge at the CVPR2023 workshop. The competition attracted 219 teams, with 66 teams advancing to the final round. The top-ranking algorithms were re-run and analyzed by the organizing team. We also thoroughly benchmarked existing representative methods on Protocol 1 and Protocol 2 (Unknown Type). Based on a comprehensive examination of the challenge results and benchmark baselines, we provide insightful analysis and discuss future research directions.

Table 1. Face anti-spoofing datasets recorded by commercial RGB cameras.

Dataset	Year	Subjects	Quantity	Format	PAs
NUAA [58]	2010	15	12,614	image	Print
PRINT-ATTACK [2]	2011	50	400	video	Print
CASIA-FASD [86]	2012	50	400	video	Print, Replay
REPLAY-ATTACK [9]	2012	50	1,200	video	Print, Replay
3DMAD [18]	2014	17	255	video	Mask(paper, hard resin)
MSU-MFSD [68]	2014	35	440	video	Print, Replay
Msspoof [10]	2015	21	4,704	image	Print
UVAD [52]	2015	404	17,076	video	Replay
MSU-USSA [50]	2016	1,140	10,260	image	Print, Replay
REPLAY-Mobile [11]	2016	40	1,030	video	Print, Replay
3DFS-DB [20]	2016	26	520	video	Mask(plastic)
HKBU-MARs V2 [40]	2016	12	1,008	video	Mask(hard resin)
BRSU [57]	2016	6	140	video	Mask(silicon, plastic, resin, latex)
OULU-NPU [4]	2017	55	3,600	video	Print, Replay
Rose-Youtu [31]	2018	20	3,350	video	Print, Replay, Mask(paper, crop-paper)
WFFD [27]	2019	745	4,600/285	image/video	Waxworks(wax)
SiW-M [43]	2019	493	1,628	video	Print, Replay, Mask(hard resin, plastic, silicone, paper, Mannequin)
CASIA-SURF [83]	2019	1,000	21,000	video	Print
SWAX [60]	2020	55	1,812/110	image/video	Waxworks(wax)
CelebA-Spoof [84]	2020	10,177	625,537	image	Print, Replay, Mask(paper)
CASIA-SURF 3DMask [79]	2020	48	1,152	video	Mask(3D print)
CASIA-SURF CeFA [35]	2021	1,607	23,538	video	Print, Replay, Mask(3D print, silica gel)
HiFiMask [39]	2021	75	54,600	video	Mask(transparent, plaster, resin)
Our Dataset (WFAS)	2023	469,920	1,383,300	image	Print(newspaper, poster, photo, album, picture book, scan photo, packging, cloth), Display(phone, tablet, TV, computer), Mask, 3D Model(garage kit, doll, adult doll, waxwork)

2. Related Work

2.1. Face Anti-Spoofing Datasets

As shown in Table 1, we focus on datasets recorded using commercial RGB cameras. The first dataset specifically designed for the face anti-spoofing field is the NUAA Photograph Imposter Database (NUAA) [58], containing only 2D print attack types with 15 subjects and 500 images per subject. In [86], the authors introduced CASIA-FASD, featuring three types of PAs: distorted printed photos, printed photos with perforated eye areas, and video replays. This dataset can be seen as an extension of NUAA, increasing PA diversity.

PRINT-ATTACK [2] was the first dataset to provide accurate protocols, including training, evaluation, and testing sets, and contains attack videos of 50 printed photos with different identities. REPLAY-ATTACK [9] added more PA types, established a protocol for fair comparison of face anti-spoofing algorithms, and demonstrated the vulnerability of face recognition systems to these attacks. Additional similar 2D print or replay datasets include MSU-MFSD [68], MSU-USSA [50], Msspoof [10], UVAD [52], and REPLAY-Mobile [11].

Compared to 2D PAs, 3D masks offer a more realistic texture and geometric structure, making them more effective at deceiving FAS systems. 3DMAD [18] was the first published 3D mask FAS dataset, featuring 255 videos of 17 subjects. The rigid mask is made of paper and hard resin. Subsequent datasets like 3DFS-DB [20], HKBU-MARs V2 [40], and BRSU [57] improved acquisition equipment, mask types, and lighting environments. WFFD [27] introduced the first waxwork dataset, comprising 450 subjects and 2,200 images, significantly increasing fidelity. Similarly, SWAX [60] included images and videos of waxworks. Recent 3D FAS datasets have considered new attack types, ethnic diversity, and complex recording conditions, such as SiW-M [43] with 13 fine-grained attack types, CASIA-SURF CeFA [35] addressing ethnic bias with three ethnicities, and HiFiMask [39] encompassing six lighting conditions, seven recording devices, and six scenarios.

However, the datasets mentioned above are limited in quantity, particularly in terms of subjects. In this context, the authors of CelebA-Spoof [84] explicitly address the issue of FAS dataset scale, expanding the dataset and in-

creasing the number of subjects to over ten thousand to advance the FAS community. CelebA-Spoof [84] is currently the largest dataset with 625,537 images of 10,177 subjects across eight scenarios, boasting rich annotations. However, the top three teams achieved TPR=100%@FPR = $5 * 10^{-3}$ on this dataset in the ECCV2020 FAS challenge [85], indicating that dataset diversity cannot be neglected while increasing quantity.

To further promote advancements in the FAS dataset with respect to both quantity and diversity, we introduce the first large-scale FAS dataset in the wild, named the Wild Anti-Spoofing Dataset (WFAS). WFAS comprises 853,729 images featuring 321,751 spoof subjects and 529,571 images of 148,169 live subjects, resulting in a substantial increase in quantity. Furthermore, the spoof data is sourced from the internet, encompassing a broad range of scenarios and various commercial sensors. With 17 PAs covering 2D and 3D forms, this innovative data pattern in the wild represents a significant breakthrough in FAS data diversity.

2.2. Face Anti-Spoofing Methods

As CNN architectures develop and FAS datasets are progressively released, end-to-end deep learning-based methods have come to dominate the field of FAS. Deep learning-based FAS methods can be classified into three categories: 1) classification supervision, 2) auxiliary pixel-wise supervision, and 3) generative pixel-wise supervision.

Classification-based methods typically employ binary cross-entropy supervision. In [73], an end-to-end deep learning FAS method based on CNN structure is proposed for the first time. To mitigate overfitting, [7, 22] pre-trained the models on ImageNet. To adapt to the low computing performance of mobile and edge platforms, [26] proposed using the lightweight MobileNetV2 [12]. Some research focused on structural optimization; for instance, [72] suggested employing a shallow fully convolutional network (FCN) to construct a multi-scale structure for learning local discriminative cues in FAS. Other work concentrated on loss function optimization. FAS tasks often exhibit asymmetric intra-distributions, with the living class being more compact and the spoof class being more diverse. [64] introduced an asymmetric angular-margin softmax loss to ease intra-class constraints among PAs. To enhance the prediction of hard samples, binary focal loss was utilized to expand the margin between live/spoof samples, resulting in stronger discrimination for hard samples [49]. However, binary classification models are prone to overfitting and lack robustness to attacks in scenarios with minor domain shifts. Some researchers reframed FAS as a fine-grained classification problem [53, 71], in which type labels are defined as bonafide, print, replay, *etc*. Despite this, models with multi-class CE loss still struggle to distinguish sample distributions between live and spoof, particularly for hard samples.

Pixel-wise supervision with auxiliary tasks skillfully leverages human prior knowledge. Pseudo depth labels [3, 67, 70, 80] take into account that 2D PAs lack facial depth information. These works require deep models to predict genuine depth for live samples while producing zero maps for spoof ones. Pseudo reflection maps [30, 74, 84] consider the discernible cues between the reflection of bonafide samples and PAs. The accuracy of pseudo depth/reflection labels depends on the precision of models with other relevant tasks. In contrast to these labels, binary mask labels [21, 43, 44, 55, 77] are more suitable for PAs with facial depth (*e.g*., waxworks) and easier to generate. However, binary mask labels used in current methods typically assume that all pixels in the face region have the same live/spoof distributions, generating all "one" and "zero" maps for bonafide and PAs, respectively. Such labels are imprecise and challenging to learn when dealing with partial attacks (*e.g*., paper masks with perforated eye areas).

Unlike auxiliary pixel-wise supervision, pixel-wise supervision with generative models [19, 29, 42, 44, 51, 54, 56] does not impose expert-designed hard constraints. The pixel-wise labels of these generative models are softer, allowing for a wider space for implicit spoof cue discovery. These works typically reframe FAS as a spoof noise modeling problem and design an encoder-decoder architecture to estimate the underlying spoof patterns with relaxed pixel-wise supervision (*e.g*., zero-noise maps for live faces). Generative pixel-wise supervision is visually insightful and more interpretable. However, such soft pixel-wise supervision may easily become trapped in local optima and overfit to unexpected interference (*e.g*., sensor noise), resulting in poor generalization under real-world scenarios.

3. Wild Face Anti-Spoofing Dataset

3.1. Data Construction

Live Data. The live samples in our dataset were obtained from the internet under a specific creative commons license, making it a typical face dataset in the wild that includes a variety of scenarios, races, ages, *etc*. All live faces were clustered using RetinaFace [14] and ArcFace [15], resulting in a total of 529,571 images from 148,169 live subjects.

Spoof Data. Our spoof data boasts unique characteristics that make the dataset progressive compared to current alternatives. Instead of manual collection, our spoof images were sourced from the internet, encompassing a wide range of scenarios and various commercial sensors. These images were not created to intentionally attack face recognition systems but happened to share similar presentation characteristics to spoof samples, benefiting our FAS research. Examples include 2D faces appearing in picture books or on TV screens and 3D waxwork faces at tourist attractions.

The PAs in our dataset encompass both 2D and 3D forms

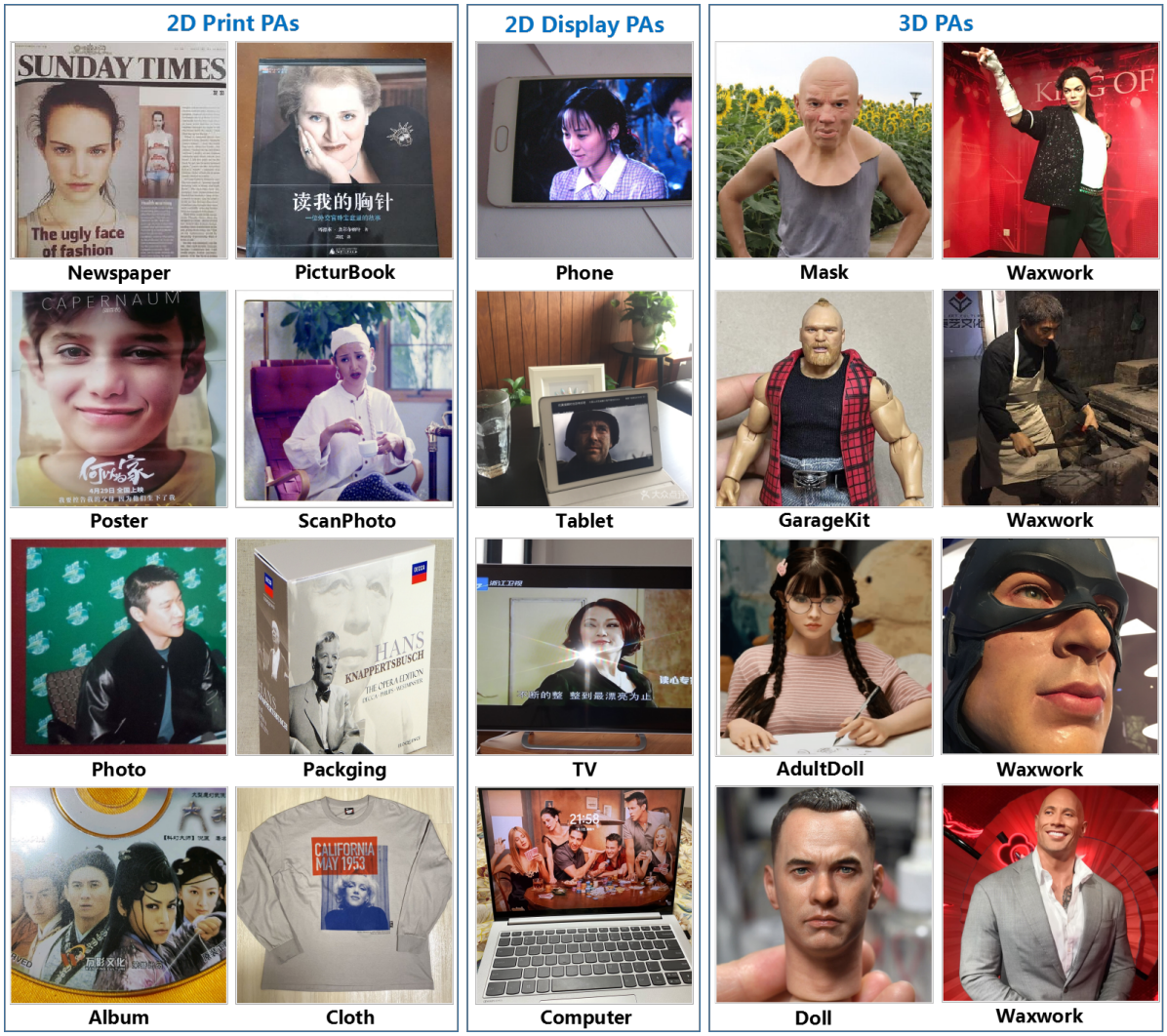

Figure 1. Examples of spoof faces in our dataset.

with diverse materials. The 2D PAs include 2D print and 2D display PAs. The former has four styles (*i.e.*, bending, folding, cutting, and plane) and eight carriers (*i.e.*, newspaper, poster, photo, album, picture book, scanned photo, packaging, and cloth). The latter comprises four subcategories (*i.e.*, phone, tablet, TV, and computer) with screen types including LCD, IPS, OLED, and VA, among others.

3D PAs feature five subcategories with varying levels of fidelity, such as masks, garage kits, dolls, adult dolls, and waxworks, made from materials like resin, plaster, plastic, silicone, latex, and wax. Our dataset contains a total of 17 PAs covering both 2D and 3D forms. Detailed subject and image numbers for all PAs are provided in Table 2.

All spoof samples were captured using a range of photographic equipment, including various mobile phone brands, digital cameras, and even scanners. The optical sensors of this equipment captured images with diverse resolutions, resulting in a richer face quality within our dataset. This new data pattern in the wild represents a significant breakthrough in FAS data quantity and diversity. Adopting such an FAS data production method saves considerable labor, material resources, and time.

4. Experimental Settings

The dataset is divided into training, development, and testing subsets with an approximate ratio of 4:1:5. As shown in Table 2, each PA does not appear in the same subset simultaneously. For instance, some PAs (*e.g.*, Newspaper, Poster, Photo, and Album) belonging to the 2D Print category appear only in the training set, while others are found in the development or testing set. This arrangement leverages the advantages of data diversity and examines the algorithm's robustness in cases of slight domain shifts.

Table 2. Dataset details including subjects/images distribution of training/development/testing settings.

Category	PAs	Subjects	Images	Train	Dev	Test
2D-Print	Newspaper	9,046	14,425	✓		
	Poster	40,858	15,439	✓		
	photo	61,990	102,826	✓		
	Album	21,122	56,490	✓		
	PictureBook	118,355	349,232		✓	✓
	ScanPhoto	1,161	2,484		✓	✓
	Packaging	3,866	19,136		✓	✓
	Cloth	138	266		✓	✓
2D-Display	Phone	20,813	34,907	✓		
	Tablet	8,089	15,431	✓		
	TV	28,184	75,606		✓	✓
	Computer	13,938	25,291		✓	✓
3D	Mask	268	1,454	✓		
	GarageKit	1,488	4,505	✓		
	AdultDoll	165	12,021	✓		
	Doll	15,406	91,954		✓	✓
	Wax	2,283	6,843		✓	✓

4.1. Evaluation Metrics

We adopt the widely used standardized ISO/IEC 30107-3 metrics for performance evaluation. The evaluation metric comprises Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER). APCER and BPCER measure the error rates of spoof and live samples, respectively. ACER is calculated as the mean of APCER and BPCER. The formulas are as follows:

$$APCER = FP/(TN + FP), \tag{1}$$

$$BPCER = FN/(TP + FN), \tag{2}$$

$$ACER = (APCER + BPCER)/2, \tag{3}$$

where FN, FP, TN, and TP represent false negatives, false positives, true negatives, and true positives, respectively. Additionally, Equal Error Rate (EER) [48] is used in the development set to obtain the threshold, which is then employed to calculate APCER and BPCER in the testing set.

4.2. Evaluation Protocols

Known-Type Protocol. In contrast to the widely used intra-dataset and intra-type protocol, which evaluates each PA type individually, we introduce a new protocol called Known-Type Protocol. This protocol employs all PA types for training, development, and testing, offering a more global, compatible, and realistic application scenario.

Unknown-Type Protocol. This protocol designates one PA type to appear exclusively in the testing stage to assess whether the algorithms have learned generalized spoof cues for unknown attack types. In this work, 2D PAs are utilized in the training and development stages, while 3D PAs are employed in the testing stage.

4.3. Baselines

We evaluate various representative algorithms on Protocol 1 and Protocol 2, including classification supervision (*i.e.*, ResNet-50 [25], PatchNet [64], MaxVit [59]), auxiliary pixel-wise supervision (*i.e.*, CDCN++ [80], CDCN++binarymask [75], DCN [81], DC-CDN [78]) and generative pixel-wise supervision (*i.e.*, LGSC [19]). For a fair comparison, we do not use pre-trained models, and other parameters are reproduced according to the original works. Additionally, to eliminate the influence of arbitrary and unfaithful cues (*e.g.*, screen bezel) on spoofing patterns, we use the face bounding box [14] as the input scale, forcing the models to focus on the face area for feature learning and prediction. The results are listed in Table 3, where we find that classification supervision-based models perform better in both Protocol 1 and Protocol 2. The performances of models with greater designability (*i.e.*, auxiliary pixel-wise supervision) fall short of expectations. Generative pixel-wise supervision significantly outperforms auxiliary pixel-wise supervision in Protocol 1 and demonstrates exceptional potential in Protocol 2, achieving the top performance.

4.4. Challenge Results

Based on our dataset and Protocol 1 (Known-Type), we hosted the Wild Face Anti-Spoofing Challenge at the CVPR 2023 workshop. A total of 219 teams participated in the competition, with 66 teams advancing to the final round. The top-ranking algorithms were re-run and analyzed by the organizing team. Table 4 presents the top-10 results of the Wild Face Anti-Spoofing Challenge at the CVPR 2023 workshop. In contrast to the baseline evaluation, we relaxed the face scale constraint for the Challenge. The face scale was randomly set to between 1.0 and 1.2 times the side length, making it more suitable for real-world application scenarios (*e.g.*, exposing part of the screen bezel or book edge). Ultimately, the teams from China Telecom, Meituan, and NetEase secured the top three positions in the competition. Here is a brief introduction to their solutions:

China Telecom. As illustrated in Figure 2, the champion solution consists of two learning phases, *i.e.*, self-supervised [6] stage and supervised learning stage. In the first stage, the model passes two different views with random transformations of an input image, without labels, to the student and teacher networks. Both networks have the same architecture but different parameters. Each network outputs a K-dimensional feature that is normalized with a temperature softmax over the feature dimension. The output of the teacher network is centered with a mean computed over the batch. Their similarity is then measured with

Table 3. Baseline performance under Protocol 1 and Protocol 2.

Prot.	SOTA Method	Flops	Dev		Test		
			EER threshold	EER(%)	APCER(%)	BPCER(%)	ACER(%)
1	Res-50	4.13G	0.6736	4.55	6.45	8.96	7.71
	PatchNet	1.82G	0.8415	5.78	8.13	8.94	8.53
	MaxVit	4.46G	0.4964	3.57	5.44	7.72	**6.58**
	CDCN++binarymask	79.79G	0.4374	10.96	11.15	11.84	11.50
	CDCN++	50.97G	0.2692	7.75	7.96	9.52	8.74
	DC-CDN	354.94M	0.2656	8.99	9.26	11.96	10.61
	DCN	49.31G	0.9915	18.58	19.93	18.88	19.40
	LGSC	9.56G	4.56e-5	5.32	7.60	9.39	8.50
2	Res-50	4.13G	0.5647	4.02	47.10	7.76	27.43
	PatchNet	1.82G	0.6957	3.99	58.74	6.07	32.40
	MaxVit	4.46G	0.4516	3.13	51.50	6.74	29.12
	CDCN++binarymask	79.79G	0.3774	10.50	49.76	11.68	30.72
	CDCN++	50.97G	0.2406	6.85	51.69	8.52	30.11
	DC-CDN	354.94M	0.2559	9.90	54.86	11.69	33.28
	DCN	49.31G	0.9877	14.40	15.71	68.63	42.17
	LGSC	9.56G	1.91e-4	4.65	45.38	7.55	**26.47**

Table 4. The top-10 results of the Wild Face Anti-Spoofing Challenge at CVPR2023 workshop.

Rank	Affiliation	Team	ACER(%)	APCER(%)	BPCER(%)
1	ChinaTelecom	xuyaowen	**1.6010**	**1.2960**	**1.9060**
2	Meituan	hexianhua	2.2210	1.3770	3.0640
3	Netease	buccellati	2.5540	2.3390	2.7690
4	SCUT	xmj	2.8940	1.4440	4.3450
5	-	luoman	3.0700	1.7450	4.3950
6	-	Sicks	3.1450	1.7250	4.5640
7	KiwiTech	KiwiTech_LeoDu	3.1800	2.2060	4.1540
8	XMU	Iverson	3.1890	3.2890	3.0900
9	-	admin123	3.5300	2.7530	4.3060
10	SJTU	iKunCTRL	3.5430	3.2420	3.8440

a cross-entropy loss. The stop-gradient [8] operator is applied to the teacher to propagate gradients only through the student network. The teacher network's parameters are updated with an exponential moving average (EMA) of the student parameters.

In the second stage, self-supervised weights are used to initialize ViT [17] for supervised learning. Meanwhile, a series of data augmentation strategies (*e.g.*, color jitter, clahe, Gaussian blur, random fog, random crop, random flip) are adopted to improve the generalization ability. The AdamW [47] optimizer is used for 20 epochs of training, and the initial learning rate is set to 2e-6. The CosineAnnealingLR [46] is adopted to reduce the learning rate, and cross-entropy loss is used for supervised learning.

MeiTuan. Firstly, the unlabeled training and development sets are merged to form the self-supervised data. The MoCoV2 [24] self-supervised method is then employed to train two pre-trained models, *i.e.*, SwinV2-huge and SwinV2-tiny [45]. These pre-trained models are fine-tuned for the

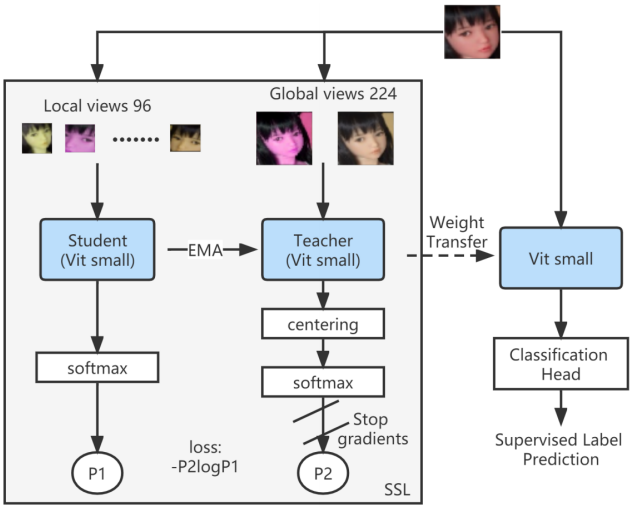

Figure 2. Method diagram of ChinaTelecom.

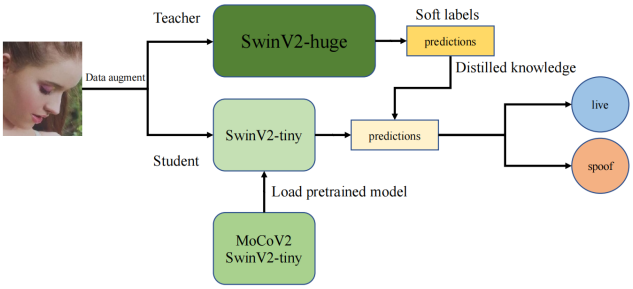

Figure 3. Method Diagram of MeiTuan.

face anti-spoofing task, resulting in significant performance improvements. Secondly, the team utilizes the knowledge distillation method DKD [87] to further enhance the performance of the face anti-spoofing model. They use SwinV2-huge as the teacher model and SwinV2-tiny as the student model. Thirdly, appropriate data augmentation techniques are applied to improve the model's generalization. The main process of the method is shown in Figure 3.

NetEase. The team from NetEase proposes a two-stage training strategy to improve the model's capability without pre-training. First, they train a convnext [69] model on the training set based on binary cross-entropy loss and then use the model to infer the soft labels of the training set. Second, they construct a compound loss using focal loss and triplet loss, and then fit the soft labels with the maxvit [59] model. Different input formats are used for the two stages of model training. For convnext, the field of view is fixed at 1.2 times the face box's outer expansion, while maxvit employs random outer expansion during training. They observe that maxvit learns more about face anti-spoofing from the soft labels of convnext. Additionally, data enhancements such as mixup, cutmix, and color adjustment are utilized to improve the model's generalization.

5. Conclusion and Discussion

In this paper, we construct a large-scale FAS dataset in the wild, named Wild Face Anti-Spoofing (WFAS) Dataset. The WFAS Dataset represents a significant breakthrough in FAS data quantity and diversity, and most importantly, it paves the way for a vast scale of wild FAS data in the future. Based on our dataset and Protocol 1, we host the Wild Face Anti-Spoofing Challenge at the CVPR2023 workshop and analyze the top-ranking algorithms. Moreover, we thoroughly benchmark representative methods on Protocol 1 and Protocol 2.

An analysis of the baseline results reveals that state-of-the-art algorithms in recent years do not demonstrate robustness to changes in scenarios (*i.e.*, dataset changes). Interestingly, the simpler classification supervision-based method achieved better results. This raises the question of whether current models with complex designs can effectively mine intrinsic FAS features. We believe that generative pixel-wise supervision methods, which offer more interpretability for visual spoof patterns, have greater potential for future developments, as evidenced by their performance across both protocols in our large-scale WFAS dataset.

Additionally, upon further investigation, we find that most current generative pixel-wise supervision methods focus on the spoof class, such as spoof noise modeling or spoof cue generation. FAS tasks typically exhibit asymmetric intra-distributions, with the live class being more compact and the spoof class being more diverse. As a result, we argue that the current definition of generative pixel-wise supervision revolving around the spoof class lacks rationality, since the PAs of the spoof class are constantly evolving and expanding. Minimizing compact live cues in spoof samples is a more reasonable approach than minimizing diverse spoof cues in live samples.

In reviewing the challenge, the top-3 solutions all utilized the Transformer [61] architecture, which has been proven to outperform CNN in several tasks. Notably, two of the winners applied self-supervised learning to build pre-trained models as a solid initialization for FAS tasks, which was one of the key factors for their success. This approach addresses the problem of the Transformer architecture's difficulty converging in computer vision tasks. Although the competition results surpassed the baseline results, we have yet to see the emergence of new methods with better interpretability for the FAS task.

References

[1] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4051, 2022. 1

[2] André Anjos and Sébastien Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7, 2011. 3

[3] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328, 2017. 4

[4] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618, 2017. 2, 3

[5] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C Kot. Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Trans-*

actions on Information Forensics and Security, 16:937–951, 2020. 1

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[7] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M Robertson, and Stan Z Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593, 2019. 1, 4

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 7

[9] Ivana Chingovska, André Anjos, and Sébastien 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, 2012. 2, 3

[10] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. Face recognition systems under spoofing attacks. *Face Recognition Across the Imaging Spectrum*, pages 165–194, 2016. 2, 3

[11] Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel. The replay-mobile face presentation-attack database. In *2016 international conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2016. 2, 3

[12] Debayan Deb and Anil K Jain. Look locally infer globally: A generalizable face anti-spoofing approach. *IEEE Transactions on Information Forensics and Security*, 16:1143–1157, 2020. 1, 4

[13] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *ECCV*, pages 741–757, 2020. 1

[14] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 4, 6

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 4

[16] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11906–11915, 2021. 1

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[18] Nesli Erdogmus and Sebastien Marcel. Spoofing face recognition with 3d masks. *IEEE transactions on information forensics and security*, 9(7):1084–1097, 2014. 2, 3

[19] Haocheng Feng, Zhibin Hong, Haixiao Yue, Yang Chen, Keyao Wang, Junyu Han, Jingtuo Liu, and Errui Ding. Learning generalized spoof cues for face anti-spoofing. *arXiv preprint arXiv:2005.03922*, 2020. 2, 4, 6

[20] Javier Galbally and Riccardo Satta. Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models. *IET Biometrics*, 5(2):83–91, 2016. 2, 3

[21] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019. 4

[22] Anjith George and Sébastien Marcel. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021. 1, 4

[23] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102, 2016. 2

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[26] Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):399–409, 2020. 4

[27] Shan Jia, Chuanbo Hu, Guodong Guo, and Zhengquan Xu. A database for face presentation attack using wax figure faces. In *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pages 39–47, 2019. 3

[28] Shan Jia, Xin Li, Chuanbo Hu, Guodong Guo, and Zhengquan Xu. 3d face anti-spoofing with factorized bilinear coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4031–4045, 2020. 2

[29] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 290–306, 2018. 4

[30] Taewook Kim, YongHyun Kim, Inhan Kim, and Daijin Kim. Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 4

[31] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 2, 3

[32] Sheng Li, Xun Zhu, Guorui Feng, Xinpeng Zhang, and Zhenxing Qian. Diffusing the liveness cues for face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1636–1644, 2021. 1

[33] Xuan Li, Jun Wan, Yi Jin, Ajian Liu, Guodong Guo, and Stan Z Li. 3dpc-net: 3d point cloud network for face anti-spoofing. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2020. 1

[34] Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biometrics*, 10(1):24–43, 2021. 1

[35] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan 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. 3

[36] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772, 2021. 1

[37] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[38] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 814–823, 2021. 1

[39] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 2, 3

[40] Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 100–106, 2016. 3

[41] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 85–100, 2016. 2

[42] Yaojie Liu and Xiaoming Liu. Physics-guided spoof trace disentanglement for generic face anti-spoofing. *arXiv preprint arXiv:2012.05185*, 2020. 2, 4

[43] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 2, 3, 4

[44] Yaojie Liu, Joel Stehouwer, and Xiaoming 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, 2020. 2, 4

[45] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[46] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[48] Sébastien Marcel, Mark S Nixon, Julian Fierrez, and Nicholas Evans. *Handbook of biometric anti-spoofing: Presentation attack detection*, volume 2. 2019. 6

[49] Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel. Improving cross-dataset performance of face presentation attack detection systems using face recognition datasets. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2947–2951, 2020. 4

[50] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016. 2, 3

[51] Dongmei Peng, Jing Xiao, Rong Zhu, and Ge Gao. Tsfen: Probing feature selection strategy for face anti-spoofing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2942–2946, 2020. 4

[52] Allan Pinto, William Robson Schwartz, Helio Pedrini, and Anderson de Rezende Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Transactions on Information Forensics and Security*, 10(5):1025–1038, 2015. 2, 3

[53] Tong Qiao, Jiasheng Wu, Ning Zheng, Ming Xu, and Xiangyang Luo. Fgdnet: Fine-grained detection network towards face anti-spoofing. *IEEE Transactions on Multimedia*, 2022. 1, 4

[54] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6311–6326, 2021. 2, 4

[55] Koushik Roy, Md Hasan, Labiba Rupty, Md Sourave Hossain, Shirshajit Sengupta, Shehzad Noor Taus, and Nabeel

Mohammed. Bi-fpnfas: Bi-directional feature pyramid network for pixel-wise face anti-spoofing by leveraging fourier spectra. *Sensors*, 21(8):2799, 2021. 4

[56] Joel Stehouwer, Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Noise modeling, synthesis and classification for generic object anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7294–7303, 2020. 2, 4

[57] Holger Steiner, Andreas Kolb, and Norbert Jung. Reliable face anti-spoofing using multispectral swir imaging. In *2016 international conference on biometrics (ICB)*, pages 1–8, 2016. 2, 3

[58] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. *ECCV (6)*, 6316:504–517, 2010. 2, 3

[59] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. pages 459–479, 2022. 6, 8

[60] Rafael Henrique Vareto, Araceli Marcia Saldanha, and William Robson Schwartz. The swax benchmark: attacking biometric systems with wax figures. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 986–990, 2020. 2, 3

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 8

[62] Jun Wan, Sergio Escalera, Hugo Jair Escalante, Guodong Guo, and Stan Z Li. Special issue on face presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):282–284, 2021. 2

[63] Jun Wan, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Multi-modal face presentation attack detection. *Synthesis Lectures on Computer Vision*, 9(1):1–88, 2020. 1

[64] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20281–20290, 2022. 1, 4, 6

[65] 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. 2

[66] Zhuming Wang, Yaowen Xu, Lifang Wu, Hu Han, Yukun Ma, and Guozhang Ma. Multi-perspective features learning for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4116–4122, 2021. 2

[67] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5042–5051. 4

[68] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 2, 3

[69] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 8

[70] Hangtong Wu, Dan Zeng, Yibo Hu, Hailin Shi, and Tao Mei. Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4626–4638, 2021. 4

[71] Xiang Xu, Yuanjun Xiong, and Wei Xia. On improving temporal consistency for online face liveness detection system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 824–833, 2021. 1, 4

[72] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *2015 3rd IAPR asian conference on pattern recognition (ACPR)*, pages 141–145, 2015. 4

[73] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 4

[74] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 557–575, 2020. 2, 4

[75] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 650–651, 2020. 6

[76] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[77] Zitong Yu, Yunxiao Qin, Xiaqing Xu, Chenxu Zhao, Zezheng Wang, Zhen Lei, and Guoying Zhao. Auto-fas: Searching lightweight networks for face anti-spoofing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 996–1000, 2020. 4

[78] Zitong Yu, Yunxiao Qin, Hengshuang Zhao, Xiaobai Li, and Guoying Zhao. Dual-cross central difference network for face anti-spoofing. *arXiv preprint arXiv:2105.01290*, 2021. 2, 6

[79] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020. 2, 3

[80] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5295–5305, 2020. 1, 2, 4, 6

[81] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Shice Liu, Bangjie Yin, Shouhong Ding, and Jilin Li. Structure destruction and

content combination for face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2021. 2, 6

[82] Licheng Zhang, Nan Sun, Xihong Wu, and Dingsheng Luo. Advanced face anti-spoofing with depth segmentation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2022. 2

[83] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 2, 3

[84] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85, 2020. 2, 3, 4

[85] Yuanhan Zhang, Zhenfei Yin, Jing Shao, Ziwei Liu, Shuo Yang, Yuanjun Xiong, Wei Xia, Yan Xu, Man Luo, Jian Liu, et al. Celeba-spoof challenge 2020 on face anti-spoofing: Methods and results. *arXiv preprint arXiv:2102.12642*, 2021. 4

[86] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31, 2012. 3

[87] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 8

[88] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 1