Deep Learning for Face Anti-Spoofing: A Survey

Zitong Yu, *Member, IEEE*, Yunxiao Qin, Xiaobai Li, *Member, IEEE*, Chenxu Zhao, Zhen Lei, *Senior Member, IEEE* and Guoying Zhao, *Fellow, IEEE*

Abstract—Face anti-spoofing (FAS) has lately attracted increasing attention due to its vital role in securing face recognition systems from presentation attacks (PAs). As more and more realistic PAs with novel types spring up, early-stage FAS methods based on handcrafted features become unreliable due to their limited representation capacity. With the emergence of large-scale academic datasets in the recent decade, deep learning based FAS achieves remarkable performance and dominates this area. However, existing reviews in this field mainly focus on the handcrafted features, which are outdated and uninspiring for the progress of FAS community. In this paper, to stimulate future research, we present the first comprehensive review of recent advances in deep learning based FAS. It covers several novel and insightful components: 1) besides supervision with binary label (e.g., '0' for bonafide vs. '1' for PAs), we also investigate recent methods with pixel-wise supervision (e.g., pseudo depth map); 2) in addition to traditional intra-dataset evaluation, we collect and analyze the latest methods specially designed for domain generalization and open-set FAS; and 3) besides commercial RGB camera, we summarize the deep learning applications under multi-modal (e.g., depth and infrared) or specialized (e.g., light field and flash) sensors. We conclude this survey by emphasizing current open issues and highlighting potential prospects.

Index Terms—face anti-spoofing, presentation attack, deep learning, pixel-wise supervision, multi-modal, domain generalization.

1 Introduction

Due to its convenience and remarkable accuracy, face recognition technology [1] has been applied in a few interactive intelligent applications such as checking-in and mobile payment. However, existing face recognition systems are vulnerable to presentation attacks (PAs) ranging from print, replay, makeup, 3D-mask, etc. Therefore, both academia and industry have paid extensive attention to developing face anti-spoofing (FAS) technology for securing the face recognition system. As illustrated in Fig. 1, FAS (namely 'face presentation attack detection' or 'face liveness detection') is an active research topic in computer vision and has received an increasing number of publications in recent years.

In the early stage, plenty of traditional handcrafted feature [2], [3], [4], [5], [6] based methods have been proposed for presentation attack detection (PAD). Most traditional algorithms are designed based on human liveness cues and handcrafted features, which need rich task-aware prior knowledge for design. In term of the methods based on the liveness cues, eye-blinking [2], [7], [8], face and head movement [9], [10] (e.g., nodding and smiling), gaze tracking [11], [12] and remote physiological signals (e.g., rPPG [3], [13], [14], [15]) are explored for dynamic discrimination. However, these physiological liveness cues are usually captured from long-term interactive face videos, which is inconvenient for practical deployment. Furthermore, the liveness cues are easily mimicked by video attacks, making them less reliable. On the other hand, classical handcrafted descriptors (e.g., LBP [4], [16], SIFT [6], SURF [17], HOG [5] and DoG [18]) are designed for extracting effective spoofing patterns from various color spaces (RGB, HSV, and YCbCr). It can be seen from Table-A 1 (in Appendix) that the FAS surveys before 2018 mainly focus on this category.

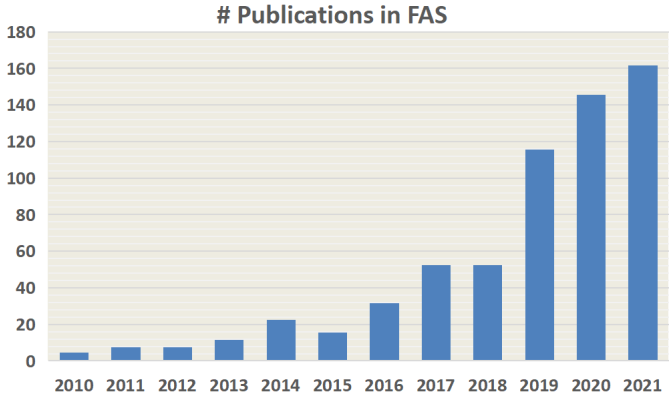


Fig. 1: The increasing research interest in the FAS field, obtained through Google scholar search with key-words: allintitle: "face anti-spoofing", "face presentation attack detection", and "face liveness detection".

Subsequently, a few hybrid (handcrafted+deep learning) [19], [20], [21], [22] and end-to-end deep learning based methods [13], [23], [24], [25], [26], [27], [28] are proposed for both static and dynamic face PAD. Most works [29], [30], [31], [32], [33], [34], [35] treat FAS as a binary classification problem (e.g., '0' for live while '1' for spoofing faces, or vice versa) thus supervised by a simple binary cross-entropy

- *Z. Yu, X. Li and G. Zhao are with Center for Machine Vision and Signal Analysis, University of Oulu, Oulu 90014, Finland. E-mail: {zitong.yu, xiaobai.li, guoying.zhao}@oulu.fi*
- *Y. Qin is with Communication University of China, Beijing 100024, China. E-mail: qinyunxiao@cuc.edu.cn*
- *C. Zhao is with SailYond Technology, Beijing 100000, China. E-mail: zhaochenxu@sailyond.com*
- *Z. Lei is with the National Laboratory of Pattern Recognition (NLPR), Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, SAR. E-mail: zlei@nlpr.ia.ac.cn*

Manuscript received June 26, 2021; revised May 15 and September 2, 2022. (Corresponding author: Guoying Zhao)

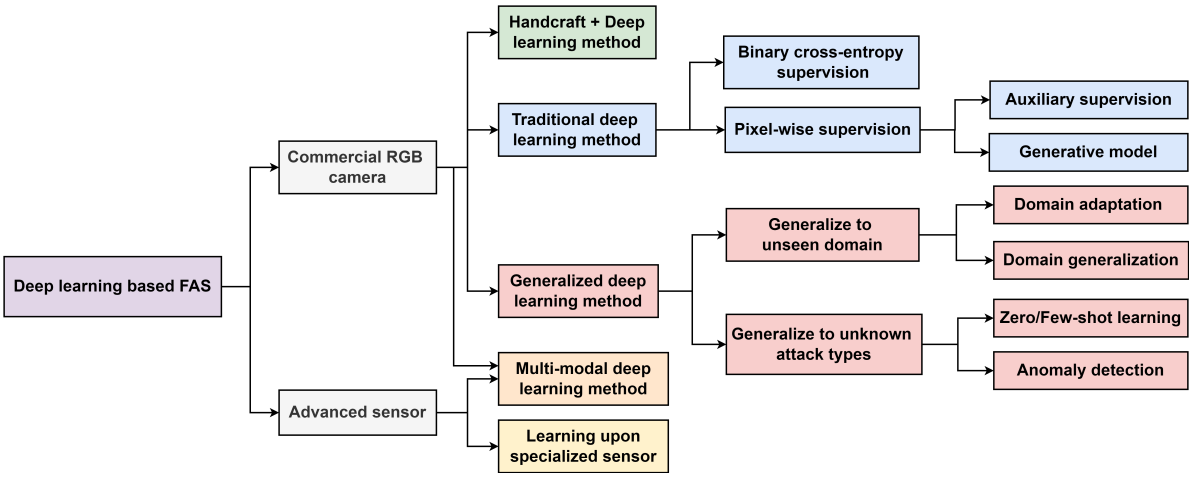


Fig. 2: Topology of the deep learning based FAS methods.

loss. Different from other binary vision tasks, the FAS is a self-evolving problem (i.e., attack vs. defense develop iteratively), which makes it more challenging. Furthermore, other binary vision tasks (e.g., human gender classification) highly rely on the obvious appearance-based semantic clues (e.g., hair style, wearing, facial shape) while the intrinsic features (e.g., material and geometry) in FAS are usually content-irrelevant (e.g., not related to facial attribute and ID), subtle and with fine-grained details, which are very challenging to distinguish by even human eyes. Thus, convolutional neural networks (CNNs) with single binary loss might reasonably mine different kinds of semantic features for binary vision tasks like gender classification but discover arbitrary and unfaithful clues (e.g., screen bezel) for spoofing patterns. Fortunately, such intrinsic live/spoof clues are usually closely related with some position-aware auxiliary tasks. For instance, the face surface of print/replay and transparent mask attacks are usually with irregular/limited geometric depth distribution and abnormal reflection, respectively. Based on these physical evidences, recently, pixel-wise supervision [13], [24], [26], [32], [36], [37] attracts more attention as it provides more fine-grained context-aware supervision signals, which is beneficial for deep models learning intrinsic spoofing cues. On one hand, pseudo depth labels [13], [26], reflection maps [24], [36], binary mask label [32], [38], [39] and 3D point cloud maps [40] are typical pixel-wise auxiliary supervisions, which describe the local live/spoof cues in pixel/patch level. On the other hand, besides physical-guided auxiliary signals, a few generative deep FAS methods model the intrinsic spoofing patterns via relaxed pixel-wise reconstruction constraints [33], [41], [42], [43]. As shown in Table-A 1 (in Appendix), the latest FAS surveys from 2018 to 2020 investigate limited numbers ($<$50) of deep learning based methods, which hardly provide comprehensive elaborations for the community researchers. Note that most data-driven methods introduced in previous surveys are supervised by traditional binary loss, and there is still a blank for summarizing the arisen pixel-wise supervision methods.

Meanwhile, the emergence of large-scale public FAS datasets with rich attack types and recorded sensors also greatly boosts the research community. First, the datasets with vast samples and subjects have been released. For instance, CelebA-Spoof [44], recorded from 10177 subjects, contains 156384 and 469153 face images for bonafide and PAs, respectively. Second, besides the common PA types (e.g., print and replay attacks), some up-to-date datasets contain richer challenging PA types (e.g., SiW-M [38] and WMCA [45] with more than 10 PA types). However, we can find from Table-A 1 (in Appendix) that existing surveys only investigate a handful of ($<$15) old and small-scale FAS datasets, which cannot provide fair benchmarks for deep learning based methods. Third, in terms of modality and hardware for recording, besides commercial visible RGB camera, numerous multimodal and specialized sensors benefit the FAS task. For example, CASIA-SURF [28] and WMCA [45] show the effectiveness of PAD via fusing RGB/depth/NIR information while dedicated systems with multispectral SWIR [46] and four-directional polarized [47] cameras significantly benefit for spoofing material perception. However, previous surveys mostly focus on single RGB modality using a commercial visible camera, and neglect the deep learning applications on the multimodal and specialized systems for high-security scenarios.

From the perspective of evaluation protocols, traditional 'intra-dataset intra-type' and 'cross-dataset intra-type' protocols are widely investigated in previous FAS surveys (see Table-A 1 in Appendix). As FAS is actually an open-set problem in practice, the uncertain gaps (e.g., environments and attack types) between training and testing conditions should be considered. However, no existing reviews consider the issues about unseen domain generalization [48], [49], [50], [51] and unknown PAD [38], [52], [53], [54]. Most reviewed FAS methods design or train the FAS model on predefined scenarios and PAs. Thus, the trained models easily overfit on several specific domains and attack types, and are vulnerable to unseen domains and unknown attacks. To bridge the gaps between academic research and real-world applications, in this paper, we fully investigate deep learning based methods under four FAS protocols, including challenging domain generalization and open-set PAD situations. Compared with existing literatures, the major contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first survey paper to comprehensively cover ($>$100) deep learn-

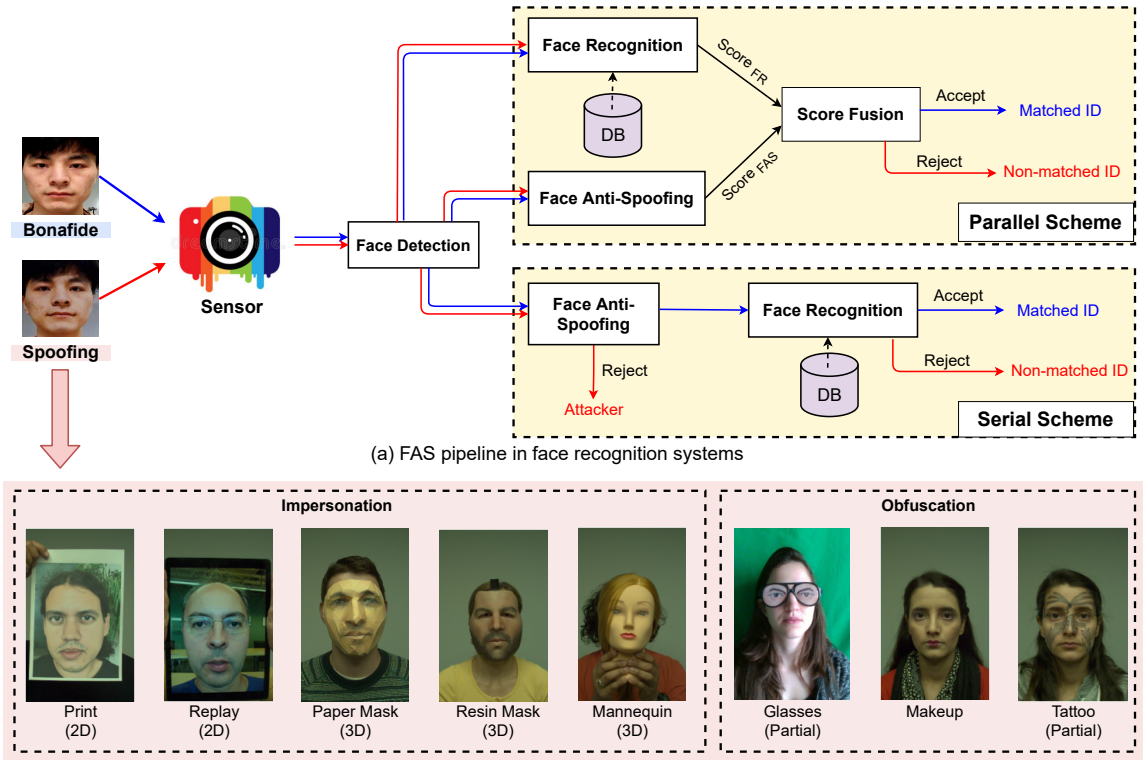


Fig. 3: Typical face spoofing attacks and face anti-spoofing pipeline. (a) FAS could be integrated with face recognition systems with paralled or serial scheme for reliable face ID matching. (b) Visualization of several classical face spoofing attack types [55] in terms of impersonation/obfuscation, 2D/3D, and whole/partial evidences.

ing methods for both single- and multi-modal FAS with generalized protocols. Compared with previous surveys only considering the methods with binary loss supervision, we also elaborate on those with auxiliary/generative pixel-wise supervision.

- As opposed to existing reviews [56], [57], [58] with only limited numbers (<15) of small-scale datasts, we show detailed comparisons among past-to-present 35 public datasets including various kinds of PAs as well as advanced recording sensors.
- This paper covers the most recent and advanced progress of deep learning on four practical FAS protocols (i.e., intra-dataset intra-type, cross-dataset intra-type, intra-dataset cross-type, and cross-dataset cross-type testings). Therefore, it provides the readers with state-of-the-art methods with different application scenarios (e.g., unseen domain generalization and unknown attack detection).
- Comprehensive comparisons of existing deep FAS methods with insightful taxonomy are provided in Tables-A 5, 6, 7, 8, 9, 10, and 11 (in Appendix), with brief summaries and discussions being presented.

We summarize the topology of deep learning based FAS methods with the commercial monocular RGB camera and advanced sensors in Fig. 2. On one hand, as commercial RGB camera is widely used in many real-world applicational scenarios (e.g., access control system and mobile device unlocking), there are richer research works based on this branch. It includes three main categories: 1) hybrid learning methods combining both handcrafted and deep learning features; 2) traditional end-to-end supervised deep learning based methods; and 3) generalized deep learning methods to both unseen domain and unknown attack types. Besides the commercial RGB camera, researchers have also developed sensor-aware deep learning methods for efficient FAS using specialized sensors/hardwares. Meanwhile, as multi-spectrum imaging systems with acceptable costs are increasingly used in real-world applications, multi-modal deep learning based methods become hot and active in the FAS research community.

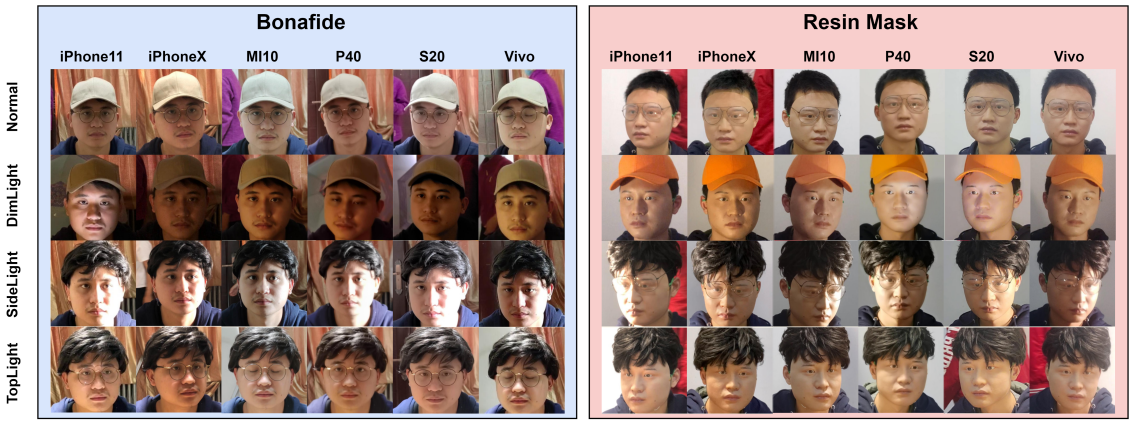
The structure of this paper is as follows. Section 2 introduces the research background, including presentation attacks, datasets, evaluation metrics, and protocols for the FAS task. Section 3 reviews the methods for visible RGB based FAS according to two kinds of supervision signals (i.e., binary loss and pixel-wise loss) as well as generalized learning for unseen domains and unknown attacks. Section 4 gives a comparison about the recording sensors as well as modalities, and then presents the methods for specific recorded inputs. Section 5 discusses the current issues of deep FAS, and indicates the future directions. Finally, conclusions are given in Section 6. Researchers can track an up-to-date list at https://github.com/ZitongYu/DeepFAS.

2 BACKGROUND

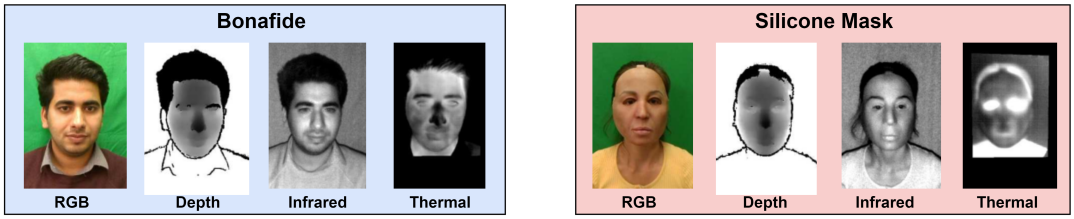
In this section, we will introduce the common face spoofing attacks first, and then investigate the existing FAS datasets as well as their evaluation metrics and protocols.

2.1 Face Spoofing Attacks

Attacks on automatic face recognition (AFR) system usually divide into two categories: digital manipulation [60], [61]



(a) Dataset recorded with various kinds of sensors under different environmental conditions



(b) Dataset with multiple modalities

Fig. 4: Visualization of the bonafide and spoofing samples from the HiFiMask dataset [59] (a) with various cameras under different lighting conditions [59]; and the WMCA dataset [45] (b) with multiple modalities such as visible RGB, depth, infrared, and thermal.

and physical presentation attacks [62]. The former one fools the face system via imperceptibly visual manipulation in the digital virtual domain, while the latter usually misleads the real-world AFR systems via presenting face upon physical mediums in front of the imaging sensors. In this paper, we focus on the detection of physical face presentation attacks, whose pipeline is illustrated in Fig. 3(a). It can be seen that there are two kinds of schemes [63] for integrating FAS with AFR systems: 1) *parallel* fusion [64] with the predicted scores from FAS and AFR systems. The combined new final score is used to determine if the sample comes from a genuine user or not; and 2) *serial* scheme [65] for early face PAs detection and spoof rejection, thus avoiding the spoof face accessing the subsequent face recognition phase.

In Fig. 3(b), some representative spoofing attack types are illustrated. According to the attackers' intention, face PAs [66] can be divided into two typical cases: 1) *impersonation*, which entails the use of spoof to be recognized as someone else via copying a genuine user's facial attributes to special mediums such as photo, electronic screen, and 3D mask; and 2) *obfuscation*, which entails the use to hide or remove the attacker's own identity using various methods such as glasses, makeup, wig, and disguised face.

Based on the geometry property, PAs are broadly classified into *2D* and *3D* attacks. 2D PAs [67] are carried out by presenting facial attributes using photo or video to the sensor. Flat/wrapped printed photos, eye/mouth-cut photos, and digital replay of videos are common 2D attack variants. With the maturity of 3D printing technology, face 3D mask [57] has become a new type of PA to threaten AFR systems. Compared with traditional 2D PAs, face masks are more realistic in terms of color, texture, and geometry structure. 3D masks are made of different materials, e.g., hard/rigid masks can be made from paper, resin, plaster, or plastic while flexible soft masks are usually composed of silicon or latex.

In consideration of the facial region covering, PAs can be also separated to *whole* or *partial* attacks. As shown in Fig. 3(b), compared with common PAs (e.g., print photo, video replay, and 3D mask) covering the whole face region, a few partial attacks only placed upon specific facial regions (e.g., part-cut print photo, funny eyeglass worn in the eyes region and partial tattoo on the cheek region), which are more obscure and challenging to detect.

2.2 Datasets for Face Anti-Spoofing

Large-scale and diverse datasets are pivotal for deep learning based methods during both training and evaluating phases. We summarize prevailing public FAS datasets in Table 1 in terms of data amount, subject numbers, modality/sensor, environmental setup, and attack types. We also visualize some samples under different environmental conditions and modalities in Fig. 4(a) and (b), respectively.

It can be seen from Table 1 that most datasets [18], [68], [69], [70], [71], [72], [73] contain only a few attack types (e.g., print and replay attacks) under simple recording conditions (e.g., indoor scene) from the early stage (i.e., year 2010-2015), which have limited variations in samples for generalized FAS training and evaluation. Subsequently, there are three main trends for dataset progress: 1) *large scale data amount*. For example, the recently released datasets CelebA-Spoof [44] and HiFiMask [59] contain more than 600000 images and 50000 videos, respectively, where most of them are with PAs; 2) *diverse data distribution*. Besides common print and replay attacks recorded in controllable indoor scenario, more and more novel attack types as well

TABLE 1: A summary of **public available datasets** for face anti-spoofing. The upper part of the table lists the datasets recorded via *commercial RGB camera* while the half bottom investigates the datasets with *multiple modalities or specialized sensors*. In the column '#Live/Spoof', 'I' and 'V' denotes 'images' and 'videos', respectively. '#Sub.' is short for Subjects.

Dataset & Reference	Year	#Live/Spoof	#Sub.	M&H	Setup	Attack Types
NUAA [18]	2010	5105/7509(I)	15	VIS	N/R	Print(flat, wrapped)
YALE_Recaptured [68]	2011	640/1920(I)	10	VIS	50cm-distance from 3 LCD minitors	Print(flat)
CASIA-MFSD [69]	2012	150/450(V)	50	VIS	7 scenarios and 3 image quality	Print(flat, wrapped, cut), Replay(tablet)
REPLAY-ATTACK [70]	2012	200/1000(V)	50	VIS	Lighting and holding	Print(flat), Replay(tablet, phone)
Kose and Dugelay [71]	2013	200/198(I)	20	VIS	N/R	Mask(hard resin)
MSU-MFSD [72]	2014	70/210(V)	35	VIS	Indoor scenario; 2 types of cameras	Print(flat), Replay(tablet, phone)
UVAD [73]	2015	808/16268(V)	404	VIS	Different lighting, background and places in two sections	Replay(monitor)
REPLAY-Mobile [74]	2016	390/640(V)	40	VIS	5 lighting conditions	Print(flat), Replay(monitor)
HKBU-MARs V2 [75]	2016	504/504(V)	12	VIS	7 cameras from stationary and mobile devices and 6 lighting settings	Mask(hard resin) from Thatsmyface and REAL-f
MSU USSA [6]	2016	1140/9120(I)	1140	VIS	Uncontrolled; 2 types of cameras	Print(flat), Replay(laptop, tablet, phone)
SMAD [76]	2017	65/65(V)	-	VIS	Color images from online resources	Mask(silicone)
OULU-NPU [77]	2017	720/2880(V)	55	VIS	Lighting & background in 3 sections	Print(flat), Replay(phone)
Rose-Youtu [78]	2018	500/2850(V)	20	VIS	5 front-facing phone camera; 5 different illumination conditions	Print(flat), Replay(monitor, laptop), Mask(paper, crop-paper)
SiW [13]	2018	1320/3300(V)	165	VIS	4 sessions with variations of distance, pose, illumination and expression	Print(flat, wrapped), Replay(phone, tablet, monitor)
WFFD [34]	2019	2300/2300(I) 140/145(V)	745	VIS	Collected online; super-realistic; removed low-quality faces	Waxworks(wax)
SiW-M [38]	2019	660/968(V)	493	VIS	Indoor environment with pose, lighting and expression variations	Print(flat), Replay, Mask(hard resin, plastic, silicone, paper, Mannequin), Makeup(cosmetics, impersonation, Obfuscation), Partial(glasses, cut paper)
Swax [79]	2020	Total 1812(I) 110(V)	55	VIS	Collected online; captured under uncontrolled scenarios	Waxworks(wax)
CelebA-Spoof [44]	2020	156384/ 469153(I)	10177	VIS	4 illumination conditions; indoor & outdoor; rich annotations	Print(flat, wrapped), Replay(monitor, tablet, phone), Mask(paper)
RECOD-Mtablet [80]	2020	450/1800(V)	45	VIS	Outdoor environment and low-light & dynamic sessions	Print(flat), Replay(monitor)
CASIA-SURF 3DMask [37]	2020	288/864(V)	48	VIS	High-quality identity-preserved; 3 decorations and 6 environments	Mask(mannequin with 3D print)
HiFiMask [59]	2021	13650/40950(V)	75	VIS	three mask decorations; 7 recording devices; 6 lighting conditions (periodic/random); 6 scenes	Mask(transparent, plaster, resin)
3DMAD [81]	2013	170/85(V)	17	VIS, Depth	3 sessions (2 weeks interval)	Mask(paper, hard resin)
GUC-LiFFAD [82]	2015	1798/3028(V)	80	Light field	Distance of 1.5∼2 m in constrained conditions	Print(Inkjet paper, Laserjet paper), Replay(tablet)
3DFS-DB [83]	2016	260/260(V)	26	VIS, Depth	Head movement with rich angles	Mask(plastic)
BRSU Skin/Face/Spoof [46]	2016	102/404(I)	137	VIS, SWIR	multispectral SWIR with 4 wavebands 935nm, 1060nm, 1300nm and 1550nm	Mask(silicon, plastic, resin, latex)
Msspoof [84]	2016	1470/3024(I)	21	VIS, NIR	7 environment conditions	Black&white Print(flat)
MLFP [85]	2017	150/1200(V)	10	VIS, NIR, Thermal	Indoor and outdoor with fixed and random backgrounds	Mask(latex, paper)
ERPA [86]	2017	Total 86(V)	5	VIS, Depth, NIR, Thermal	Subject positioned close (0.3∼0.5m) to the 2 types of cameras	Print(flat), Replay(monitor), Mask(resin, silicone)
LF-SAD [87]	2018	328/596(I)	50	Light field	Indoor fix background, captured by Lytro ILLUM camera	Print(flat, wrapped), Replay(monitor)
CSMAD [88]	2018	104/159(V+I)	14	VIS, Depth, NIR, Thermal	4 lighting conditions	Mask(custom silicone)
3DMA [89]	2019	536/384(V)	67	VIS, NIR	48 masks with different ID; 2 illumination & 4 capturing distances	Mask(plastics)
CASIA-SURF [90]	2019	3000/ 18000(V)	1000	VIS, Depth, NIR	Background removed; Randomly cut eyes, nose or mouth areas	Print(flat, wrapped, cut)
WMCA [45]	2019	347/1332(V)	72	VIS, Depth, NIR, Thermal	6 sessions with different backgrounds and illumination; pulse data for bonafide recordings	Print(flat), Replay(tablet), Partial(glasses), Mask(plastic, silicone, and paper, Mannequin)
CeFA [91]	2020	6300/ 27900(V)	1607	VIS, Depth, NIR	3 ethnicities; outdoor & indoor; decoration with wig and glasses	Print(flat, wrapped), Replay, Mask(3D print, silica gel)
HQ-WMCA [55]	2020	555/2349(V)	51	VIS, Depth, NIR, SWIR, Thermal	Indoor; 14 'modalities', including 4 NIR and 7 SWIR wavelengths; masks and mannequins were heated up to reach body temperature	Laser or inkjet Print(flat), Replay(tablet, phone), Mask(plastic, silicon, paper, mannequin), Makeup, Partial(glasses, wigs, tatoo)
PADISI-Face [92]	2021	1105/924(V)	360	VIS, Depth, NIR, SWIR, Thermal	Indoor, fixed green background, 60-frame sequence of 1984 × 1264 pixel images	Print(flat), Replay(tablet, phone), Partial(glasses,funny eye), Mask(plastic, silicone, transparent, Mannequin)

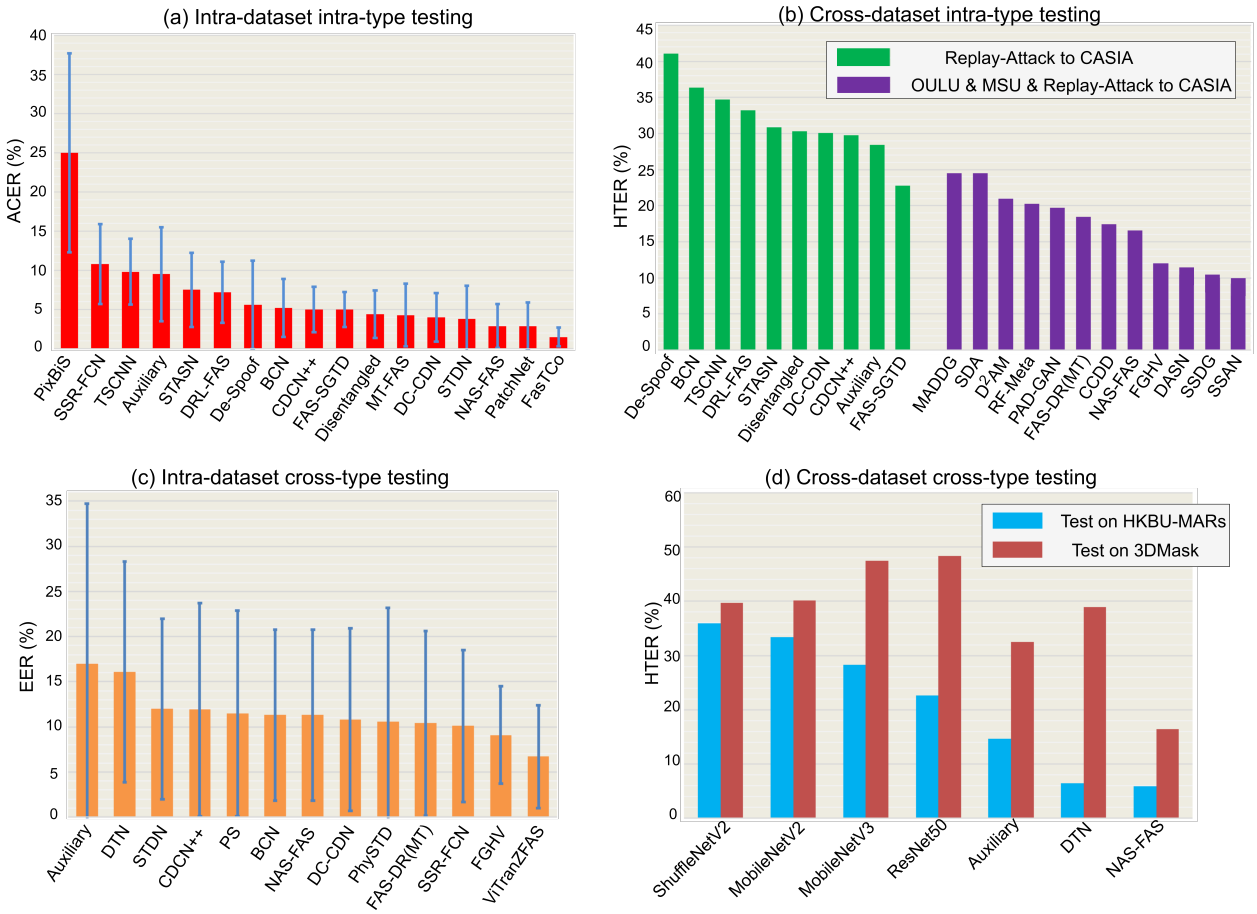


Fig. 5: The performance of deep FAS approaches on four mainstream evaluation protocols. The lower ACER, HTER and EER, the better performance. (a) Intra-dataset intra-type testing on the Protocol-4 of OULU-NPU. (b) Cross-dataset intra-type testing on CASIA-MFSD when training on single Replay-Attack dataset (see green columns) or multiple datasets including OULU-NPU, MSU-MFSD, and Replay-Attack (see purple columns). (c) Intra-dataset cross-type testing on SiW-M with leave-one-type-out setting. (d) Cross-dataset cross-type testing on 3D mask FAS datasets including HKBU-MARs [75] and CASIA-SURF 3DMask when training on OULU-NPU and SiW datasets with only 2D attacks.

as complex recording conditions are considered in recent FAS datasets. For example, there are 13 fine-grained attack types in SiW-M [38] while HiFiMask [59] consists of 3D masks attacks with three kinds of materials (transparent, plaster, resin) recorded under six lighting conditions and six indoor/outdoor scenes; and 3) *multiple modalities and specialized sensors*. Apart from traditional visible RGB camera, some recent datasets also consider various modalities (e.g., NIR [45], [55], [90], [91], Depth [45], [55], [90], [91], Thermal [45], [55], and SWIR [55]) and other specialized sensors (e.g., Light field camera [82], [87]). All these advanced factors facilitate the area of FAS in both academic research and industrial deployment.

2.3 Evaluation Metrics

As FAS systems usually focus on the concept of bonafide and PA acceptance and rejection, two basic metrics False Rejection Rate (FRR) and False Acceptance Rate (FAR) [93] are widely used. The ratio of incorrectly accepted spoofing attacks defines FAR, whereas FRR stands for the ratio of incorrectly rejected live accesses [94]. FAS follows ISO/IEC DIS 30107- 3:2017 [95] standards to evaluate the performance of the FAS systems under different scenarios. The most commonly used metrics in both intra- and cross-testing scenarios is Half Total Error Rate (*HTER*) [94], Equal Error Rate (*EER*) [67], and Area Under the Curve (*AUC*). HTER is found out by calculating the average of FRR (ratio of incorrectly rejected bonafide score) and FAR (ratio of incorrectly accepted PA). EER is a specific value of HTER at which FAR and FRR have equal values. AUC represents the degree of separability between bonafide and spoofings.

Recently, Attack Presentation Classification Error Rate (*APCER*), Bonafide Presentation Classification Error Rate (*BPCER*) and Average Classification Error Rate (*ACER*) suggested in ISO standard [95] are also used for intra-dataset testings [13], [77]. BPCER and APCER measure bonafide and attack classification error rates, respectively. ACER is calculated as the mean of BPCER and APCER, evaluating the reliability of intra-dataset performance.

2.4 Evaluation Protocols

To evaluate the discrimination and generalization capacities of the deep FAS models, various protocols have been established. We summarize the development of deep FAS approaches on four representative protocols in Fig. 5 and Tables-A 2, 3 and 4 (in Appendix).

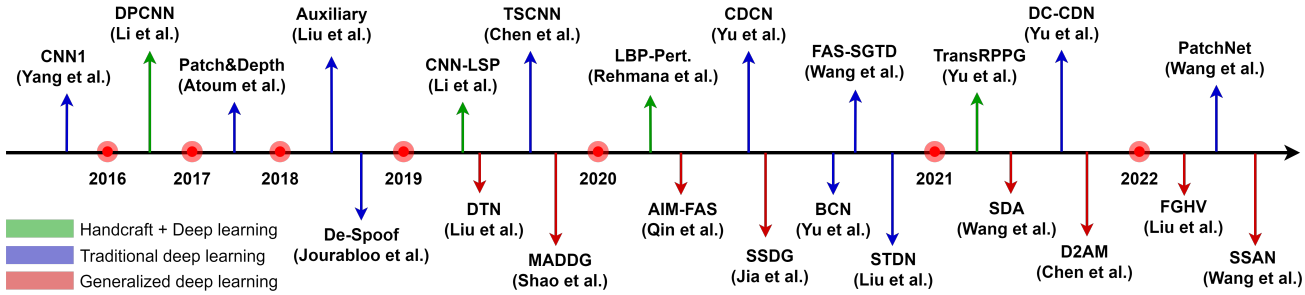


Fig. 6: Chronological overview of the milestone deep learning based FAS methods using commercial RGB camera.

Intra-Dataset Intra-Type Protocol. Intra-dataset intra-type protocol has been widely used in most FAS datasets to evaluate the model's discrimination ability for spoofing detection under scenarios with slight domain shift. As the training and testing data are sampled from the same datasets, they share similar domain distribution in terms of the recording environment, subject behavior, etc. (see Fig. 4(a) for examples). The most classical intra-dataset intra-type testing is the Protocol-4 of OULU-NPU dataset [77], and the performance comparison of recent deep FAS methods on this protocol is shown in Fig. 5(a). Due to the strong discriminative feature representation ability via deep learning, many methods (e.g., CDCN++ [23], FAS-SGTD [96], Disentangled [97], MT-FAS [43], DC-CDN [98], STDN [99], NAS-FAS [37], FasTCo [100] and PatchNet [101]) have reached satisfied performance (<5% ACER) under small domain shifts in terms of external environment, attack mediums and recording camera variation. More intra-dataset intra-type results on OULU-NPU (4 sub-protocols) and SiW (3 sub-protocols) datasets are listed in Table-A 2 (in Appendix).

Cross-Dataset Intra-Type Protocol. This protocol focuses on cross-dataset level domain generalization ability measurement, which usually trains models on one or several datasets (source domains) and then tests on unseen datasets (shifted target domain). We summarize recent deep FAS approaches on two favorite cross-dataset testings [23], [48] in Fig. 5(b). It can be seen from green columns that, when trained on Replay-Attack and tested on CASIA-MFSD, most deep models perform poorly (>20% HTER) due to the serious lighting and camera resolution variations. In contrast, when trained on multiple source datasets (i.e., OULU-NPU, MSU-MFSD, and Replay-Attack), domain generalization based methods achieve acceptable performance (especially SSDG [51] and SSAN [102] with 10.44% and 10.00% HTER, respectively). In real-world cross-testing cases, small amount of target domain data are easily obtained, which can also be utilized for domain adaptation [103] to mitigate domain shifts further. More cross-dataset intra-type testings among OULU-NPU, CASIA-MFSD, Replay-Attack, and MSU-MFSD datasets with different numbers of source domains for training can be found in Table-A 3 (in Appendix).

Intra-Dataset Cross-Type Protocol. The protocol adopts 'leave one attack type out' to validate the model's generalization for unknown attack types, i.e., one kind of attack type only appears in the testing stage. Considering the rich (13 kinds) attack types, SiW-M [38] is investigated in this protocol, and the corresponding results are illustrated in Fig. 5(c). Most of the deep models achieve around 10% EER and with large standard deviations among all attack types, which indicates the huge challenges in this protocol. Benefited from the large-scale pretraining, ViTranZFAS [38] achieves surprising 6.7% EER, implying the promising usage of transfer learning for unknown attack type detection. Detailed intra-dataset cross-type testing results on SiW-M with the leave-one-type-out setting are shown in Table-A 4 (in Appendix).

Cross-Dataset Cross-Type Protocol. Although the three protocols mentioned above mimic most factors in real-world applications, they do not consider the most challenging case, i.e., cross-dataset cross-type testing. [37] proposes a 'Cross-Dataset Cross-Type Protocol' to measure the FAS model's generalization on both unseen domain and unknown attack types. In this protocol, OULU-NPU and SiW (with 2D attacks) are mixed for training, while HKBU-MARs and 3DMask (with 3D attacks) are used for testing. It can be seen from Fig. 5(d) that recent deep models (DTN [38] and NAS-FAS [37]) hold good generalization for lab-controlled low-fidelity 3D mask detection on HKBU-MARs but still cannot satisfactorily detect unrestricted high fidelity masks on 3DMask.

Besides these four mainstream evaluation protocols, more new trends about practical protocol settings (e.g., semi-/un-supervised, real-world open-set, and dynamic multimodality) will be discussed in Section 5.

3 Deep FAS with Commercial RGB Camera

As commercial RGB camera is widely used in many real-world applicational scenarios (e.g., access control system and mobile device unlocking), in this section, we will review existing commercial RGB camera based FAS methods. Several milestone deep FAS methods are illustrated in Fig. 6.

3.1 Hybrid (Handcraft + Deep Learning) Method

Although deep learning and CNNs have achieved great success in many computer vision tasks (e.g., image classification [104], [105], semantic segmentation [106], and object detection [107]), they suffer from the overfitting problem for the FAS task due to the limited amount and diversity of the training data. As handcrafted features (e.g., LBP [108], HOG [109] descriptors, image quality [110], optical flow motion [111], and rPPG clues [112]) have been proven to be discriminative to distinguish bonafide from PAs, some recent *hybrid* works combine handcrafted features with deep features for FAS. Typical properties of these hybrid methods are summarized in Table-A 5 (in Appendix).

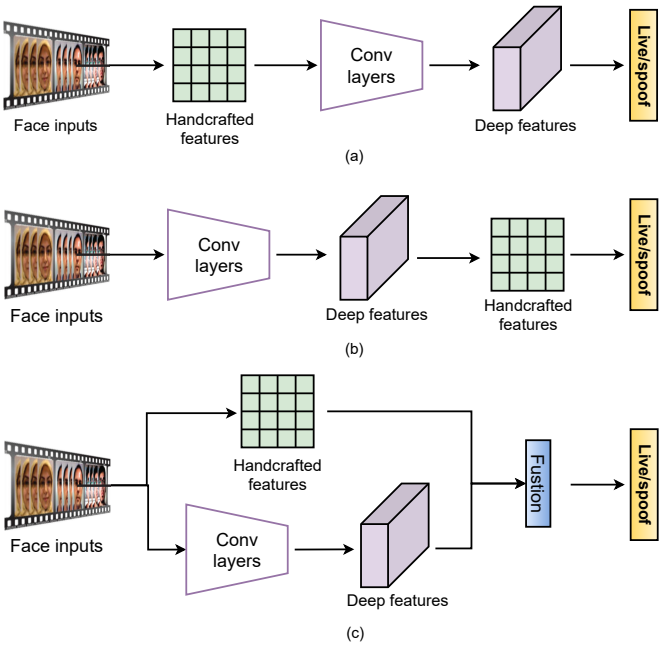


Fig. 7: Hybrid frameworks for FAS. (a) Deep features from handcrafted features. (b) Handcrafted features from deep features. (c) Fused handcrafted and deep features.

Some FAS approaches firstly extract handcrafted features from face inputs, and then employ CNNs for semantic feature representation (see Fig. 7(a) for paradigm). On one hand, color texture based static features are extracted from each frame, and then are feed into the deep model. Based on the rich low-level texture features, deep model is able to mine texture-aware semantic clues. To this end, Cai and Chen [113] adopt multi-scale color LBP features as local texture descriptors, then a random forest is cascaded for semantic representation. Similarly, Khammari [22] extracts LBP and Weber local descriptor encoded CNN features, which are combined to preservate the local intensity and edge orientation information. However, compared with the original face input, local descriptor based features lose pixel-level details thus limiting the model performance. On the other hand, dynamic features (e.g., motion, illumination changes, physiological signals) across temporal frames are also effective CNN inputs. Feng et al. [114] propose to train a multi-layer perceptron from the extracted dense optical flow-based motions, which reveal anomalies in print attacks. Moreover, Yu et al. [115] construct spatio-temporal rPPG maps from face videos, and use a vision transformer to capture the periodic heartbeat liveness features for the bonafide. However, head motions and rPPG signals are easily imitated in the replay attack, making such dynamic clues less reliable. Basing on the fact that replay attacks usually have abnormal reflection changes, Li et al. [116] propose to capture such illumination changes using a 1D CNN with inputs of the intensity difference histograms from reflectance images.

Several other hybrid FAS methods extract handcrafted features from deep convolutional features, which follow the hybrid framework in Fig. 7(b). To reduce the FAS-unrelated redundancy, Li et al. [30] utilize the block principal component analysis (PCA) to filter out the irrelevant deep features from pretrained VGG-face model. Besides PCA-based dimension reduction, Agarwal et al. [117] explicitly extract the color LBP descriptor from the shallow convolutional features, which contains richer low-level statistics. In addition to static spoof patterns, some works also explore handcrafted dynamic temporal clues from well-trained deep models. Asim et al. [20] and Shao et al. [118] extract deep dynamic textures and motion features using LBP-TOP [119] and optical flow from the sequential convolutional features, respectively. One limitation of this hybrid framework is that the handcrafted features are highly dependent on the well-trained convolutional features, and it is still unknown whether shallow or deep convolutional features are more suitable for different kinds of handcrafted features.

As handcrafted and deep convolutional features hold different properties, another favorite hybrid framework (see Fig. 7(c)) fuses them for more generic representation. To make more reliable predictions, Sharifi [120] proposes to fuse the predicted scores from both handcrafted LBP features and deep VGG16 model. However, it is challenging to determine the optimal score weights for these two kinds of features. Besides score-level fusion, Rehmana et al. [21], [121] propose to utilize HOG and LBP maps to perturb and modulate the low-level convolutional features. Despite the fact that local prior knowledge from handcrafted features enhances discriminative capacity, the overall model suffers from semantic representation degradation. In terms of the temporal methods, to leverage the dynamic discrepancy between the bonafide and PAs, Li et al. [122] extract intensity variation features via 1D CNN, which are fused with the motion blur features from motion magnified face videos for replay attack detection.

Overall, benefitted from the explicit expert-designed feature extraction, hybrid methods are able to represent particular non-texture clues (e.g., temporal rPPG and motion blur), which are hard to capture via end-to-end texture-based FAS models. However, the shortcomings are also obvious: 1) handcrafted features highly rely on the expert knowledge and not learnable, which are inefficient once enough training data are available; and 2) there might be feature gaps/incompatibility between handcrafted and deep features, resulting in performance saturation.

3.2 Traditional Deep Learning Method

Benefited from the development of the advanced CNN architectures [105], [123] and regularization [124], [125] techniques as well as the recent released large-scale FAS datasets [44], [59], [77], end-to-end deep learning based methods attract more and more attention, and dominate the field of FAS. Different from the hybrid methods which integrate parts of handcrafted features without learnable parameters, *Traditional* deep learning based methods directly learn the mapping functions from face inputs to spoof detection. Traditional deep learning frameworks usually include: 1) direct supervision with binary cross-entropy loss (see Fig. 8(a)); and 2) pixel-wise supervision with auxiliary tasks (see Fig. 8(b)) or generative models (see Fig. 8(c)).

3.2.1 Direct Supervision with Binary Cross Entropy Loss

As FAS can be intuitively treated as a binary (bonafide vs. PA) classification task, numerous end-to-end deep learning

methods are directly supervised with binary cross-entropy (CE) loss as well as extended losses (e.g., triplet loss [126]), which are summarized in Table-A 6 (in Appendix).

On one side, researchers have proposed various network architectures supervised by binary CE loss for FAS. Yang et al. [29] propose the first end-to-end deep FAS method using 8-layer shallow CNN for feature representation. However, due to the limited scale and diversity of datasets, CNN-based models easily overfit in the FAS task. To alleviate this issue, some works [127], [128], [129] finetune the ImageNet-pretrained models (e.g., VGG16, ResNet18 and vision transformer) for FAS. Towards moblie-level FAS applications, Heusch et al. [55] consider using the lightweight MobileNetV2 [130] for efficient FAS. The above-mentioned generic backbones usually focus on high-level semantic representation while neglect low-level features, which are also important for spoof pattern mining. To better leverage the multi-scale features for FAS, Deb and Jain [131] propose to use a shallow fully convolutional network (FCN) to learn local discriminative cues from face images in a self-supervised manner. Besides the single-frame-based appearance features, several works [25], [132], [133], [134] consider the temporal discrepancy between bonafide and PAs, and cascade multi-frame-based CNN features with LSTM [135] for robust dynamic clues propagation.

On the other side, considering the weak intra- and inter-class constraints from binary CE, a few works modify binary CE loss to provide CNNs more discriminative supervision signals. Instead of binary constraints, Xu et al. [100] rephrase FAS as a fine-grained classification problem, and the type labels (e.g., bonafide, print, and replay) are used for multi-class supervision. In this way, the particular properties (e.g., materials) of PAs could be represented. However, FAS models supervised with multi-class CE loss still have confused live/spoof distributions especially on hard live/spoof samples. For instance, high-fidelity PAs have similar appearance clues as the corresponding bonafide. On one hand, to learn a compact space with small intra-class distances but large inter-class distances, Hao [136] and Almeida et al. [80] introduce contrastive loss and triplet loss, respectively. However, different from vision retrieval tasks, the bonafide and PAs in FAS task usually hold asymmetric intra-distributions (more compact and diverse, respectively). Based on this evidence, Wang et al. [101] propose to supervise the FAS patch models via an asymmetric angular-margin softmax loss to relax the intra-class constraints among PAs. On the other hand, to provide more confident predictions on hard samples, Chen et al. [137] adopt the binary focal loss to guide the model to enlarge the margin between live/spoof samples and achieve strong discrimination for hard samples.

Overall, both binary CE loss and its extended losses are easy and efficient to use, which supervise deep FAS models to fastly converge. However, these supervision signals only provide global (spatial/temporal) constraints for live/spoof embedding learning, which may causes FAS models to easily overfit to unfaithful patterns. Furthermore, FAS models with binary supervision are usually black-box and the characteristic of their learned features are hard to understand.

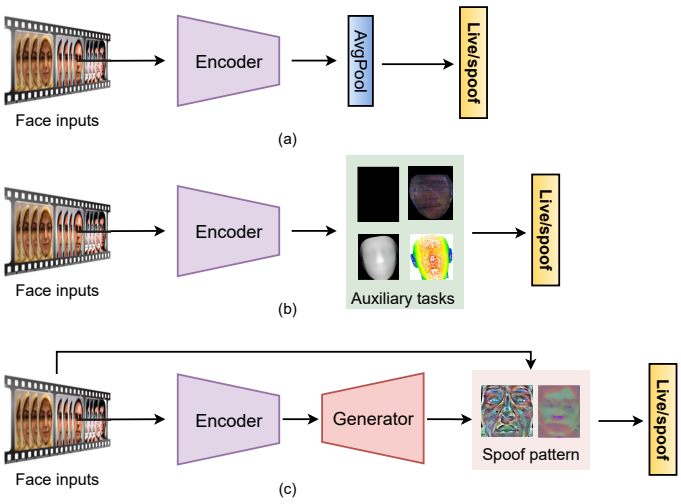


Fig. 8: Traditional end-to-end deep learning frameworks for FAS. (a) Direct supervision with binary cross entropy loss. (b) Pixel-wise supervision with auxiliary tasks. (c) Pixel-wise supervision with generative model for implicit spoof pattern representation.

3.2.2 *Pixel-wise Supervision*

Deep models directly supervised by binary loss might easily learn unfaithful patterns (e.g., screen bezel). In contrast, pixel-wise supervision can provide more fine-grained and contextual task-related clues for better intrinsic feature learning. On one hand, based on the physical clues and discriminative design philosophy, auxiliary supervision signals such as pseudo depth labels [13], [26], binary mask label [32], [38], [39] and reflection maps [24], [36] are developed for local live/spoof clues description. On the other hand, generative models with explicit pixel-wise supervision (e.g., original face input reconstruction [42], [138]) are recently utilized for generic spoof pattern estimation. We summarize the representative pixel-wise supervision methods in Table-A 7 (in Appendix).

Pixel-wise supervision with Auxiliary Task. According to the human prior knowledge of FAS, most PAs (e.g., plain printed paper and electronic screen) merely have no genuine facial depth information, which could be utilized as discriminative supervision signals. As a result, some recent works [23], [26], [96], [139] adopt pixel-wise *pseudo depth* labels to guide the deep models, enforcing them predict the genuine depth for live samples while zero maps for the spoof ones. Atoum et al. [26] first leverage pseudo depth labels to guide the multi-scale FCN (namely 'DepthNet' for simplicity). Thus, the well-trained DepthNet is able to predict holistic depth maps as decision evidence. To further improve the fine-grained intrinsic feature representation capacity, Yu et al. [23] replace vanilla convolution in DepthNet with central difference convolution (CDC) to form the CDCN architecture (see Fig. 9 for detailed structures). In terms of static architectures, DepthNet and CDCN are favorite and widely used in the deep FAS community due to their compactness and excellent performance. Many recent variants [37], [98], [140] are established based on the DepthNet/CDCN. As for the temporal architectures,

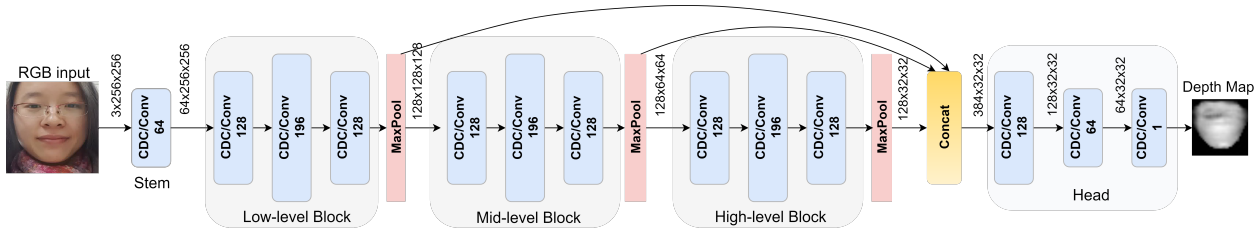


Fig. 9: The shared architecture of CDCN [23]/DepthNet [13]. Inside the blue block are the convolutional filters with 3x3 kernel size and their feature dimensionalities. 'CDC' and 'Conv' suggest central difference convolution adopted in CDCN and vanilla convolution adopted in DepthNet, respectively.

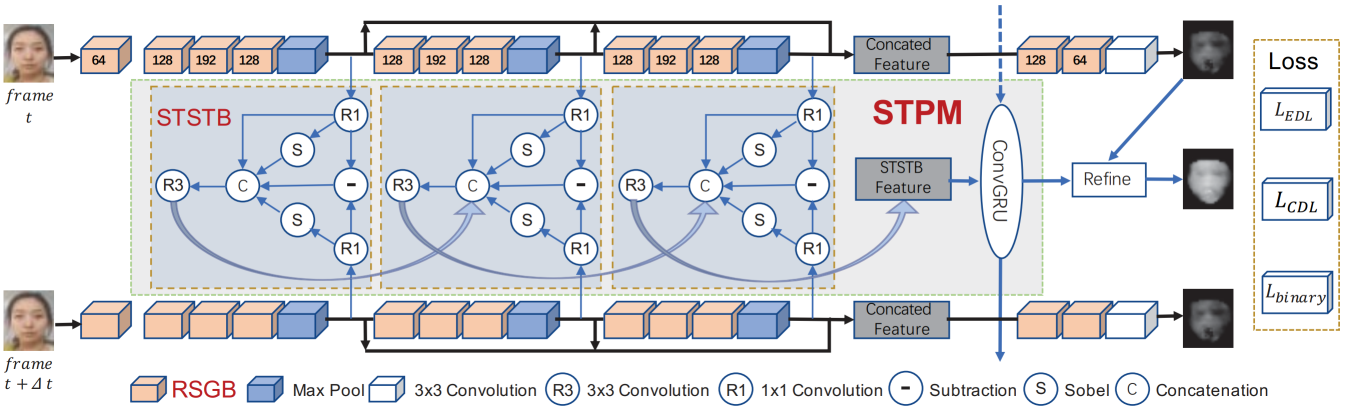


Fig. 10: The network structure of FAS-SGTD [96]. The inputs are consecutive frames with a fixed interval. Each frame is processed by cascaded Residual Spatial Gradient Block (RSGB) with a shared backbone which generates a corresponding coarse depth map. The number in RSGB cubes denotes the output channels. Spatio-Temporal Propagation Module (STPM) is plugged between frames for estimating the temporal depth and refining the corresponding coarse depth map.

FAS-SGTD [96] is classical and well-known for its excellent short- and long-term micro-motion estimation, which can be utilized for accurate facial depth prediction. The detailed architecture of FAS-SGTD is illustrated in Fig. 10, which is later modified and extended in a transformer counterpart [141].

Synthesizing 3D shape labels for every training sample is costly and not accurate enough, and also lacks the reasonability for the PAs with real depth (e.g., 3D mask and Mannequin). In contrast, binary mask label [32], [38], [99], [142], [143] is easier to be generated and more generalizable to all PAs. Specifically, binary supervision would be provided for the deep embedding features in each spatial position. In other words, through the binary mask label, we can find whether PAs occur in the corresponding patches, which is attack-type-agnostic and spatially interpretable. George and Marcel [32] are the first to introduce deep pixel-wise binary supervision to predict the intermediate confidence map for the cascaded final binary classification. With sufficient pixel-wise supervision, the backbone DenseNet121 converges well and is able to provide patch-wise live/spoof predictions. As subtle spoof clues (e.g., moiré pattern) usually exist in different spatial regions with different intensity, vanilla pixel-wise binary supervision treats all patches with equal contributions, which might lead to biased feature representation. To tackle this issue, Hossaind et al. [142] propose to add a learnable attention module for feature refinement before calculating the deep pixel-wise binary loss, which benefits the salient information propagation. Though flexible and easy to use, current binary mask labels usually assume all pixels in the face region have the same live/spoof distributions thus generate all 'one' and 'zero' maps for bonafide and PAs, respectively. However, such labels are inaccurate and noisy to learn when encountering partial attacks (e.g., FunnyEye).

Besides the mainstream depth map and binary mask labels, there are several informative auxiliary supervisions (e.g., pseudo reflection map [24], [36], [44], 3D point cloud map [40], ternary map [39], and Fourier spectra [144]). According to the discrepancy of facial material-related albedo between the live skin and spoof mediums, Kim et al. [36] propose to supervise deep models with both depth and reflection labels. Moreover, to further enhance the type-agnostic generalization, binary mask maps are introduced in [24] to train the bilateral convolutional networks with all these three pixel-wise supervisions simultaneously. Unlike binary mask labels considering all spatial positions including live/spoof-unrelated background, Sun et al. [39] remove the face-unrelated parts and leave the entire face regions as a refined binary mask called 'ternary map', which eliminates the noise outside the face and benefits the facial spoof clue mining. Based on the rich texture and geometry discrepancy between the bonafide and PAs, deep models with other auxiliary supervisions from the Fourier map [33], [144], LBP texture map [97], and sparse 3D point cloud map [40], also show their excellent representation capability.

Overall, pixel-wise auxiliary supervision benefits the physically meaningful and explainable live/spoof feature

learning (e.g., reflection and depth supervisions for material and geometry representation, respectively). Moreover, a reliable and generalized FAS model can be supervised with multiple complementary auxiliary supervisions (e.g., depth, reflection, and albedo) in a multi-task learning fashion [24]. However, two limitations of auxiliary supervision should be mentioned: 1) pixel-wise supervision usually relies on the high-quality (e.g., high-resolution) training data for fine-grained spoof clue mining, and is harder to provide effective supervision signals when training data are too noisy and with low quality; and 2) the pseudo auxiliary labels are either human-designed or generated by other off-the-shelf algorithms, which are not always trustworthy.

Pixel-wise Supervision with Generative Model. Despite the fine-grained supervision signal in the auxiliary task, it is still hard to understand whether the deep black-box models represent intrinsic FAS features. Recently, one hot trend is to mine the visual spoof patterns existing in the spoof samples, aiming to provide a more intuitive interpretation of the sample's spoofness. We summarize such kind of generative models with pixel-wise supervision in the lower part of Table-A 7 (in Appendix). In consideration of the strong physical-inspired constraints of auxiliary pixel-wise supervision, several works relax such explicit supervision signals and provide a broader space for implicit spoof clues mining. Jourabloo et al. [33] rephrase FAS as a spoof noise modeling problem, and design an encoder-decoder architecture to estimate the underlying spoof patterns with relaxed pixel-wise supervisions (e.g., zero-noise map for live faces). With such unilateral constraint on the bonafide, the models are able to mine the spoof clues flexibly for PAs. Similarly, Feng et al. [41] design a spoof cue generator to minimize the spoof cues of live samples while imposes no explicit constraints on those of spoof samples. Unlike above-mentioned works forcing strict constraints on live samples, Mohammadi et al. [138] use the reconstruction-error maps computed from a live-face-pretrained autoencoder for spoofing detection. As such error maps are generated from the residual noises of reconstructed live faces without human-defined elements, they are robust under domain shift with knowledge clue change. However, the low-quality reconstructed faces from autoencoder may lead to noisy residual error maps.

Besides direct spoof pattern generation, Qin et al. [43] propose to automatically generate pixel-wise labels via a meta-teacher framework, which is able to provide better-suited supervision for the student FAS models to learn sufficient and intrinsic spoofing cues. However, only the learnable spoof supervision is generated in [43]. Therefore, how to generate the optimal pixel-wise signals automatically for both live and spoof samples is still worth exploring.

Overall, pixel-wise supervision with generative model usually relaxes the expert-designed hard constraints (e.g., auxiliary tasks), and leaves the decoder to reconstruct more natural spoof-related trace. Thus, the predicted spoof patterns are strongly data-driven and have explainable views. The generated spoof patterns are visually insightful, and are challenging to manually describe with human prior knowledge. However, such soft pixel-wise supervision might easily fall into the local optimum and overfit on unexpected interference (e.g., sensor noise), which would generalize poorly under real-world scenarios. Combining explicit auxiliary supervision with generative model based supervision for jointly training might alleviate this issue.

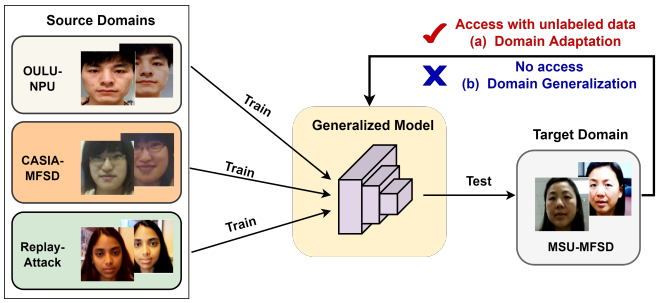


Fig. 11: Framework comparison among domain adaptation (DA) and domain generalization (DG). (a) The DA methods need the (unlabeled) target domain data to learn the model while (b) DG methods learn generalized model without knowledge from the target domain.

3.3 Generalized Deep Learning Method

Traditional end-to-end deep learning based FAS methods might generalize poorly on unseen dominant conditions (e.g., illumination, facial appearance, and camera quality) and unknown attack types (e.g., emerging high fidelity mask made of new materials). Thus, these methods are unreliable to be applied in practical applications with strong security needs. In light of this, more and more researchers focus on enhancing the generalization capacity of the deep FAS models. On one hand, domain adaptation and generalization techniques are leveraged for robust live/spoof classification under unlimited domain variations. On the other hand, zero/few-shot learning as well as anomaly detection frameworks are applied for unknown face PA types detection. In this paper, the unseen domains indicate the spoof-irrelated external changes (e.g., lighting and sensor noise) but actually influence the appearance quality. In contrast, the unknown spoofing attacks usually mean the novel attack types with intrinsic physical properties (e.g., material and geometry) which have not occurred in the training phase. Representative generalized deep FAS methods on unseen domains and unknown attack types are summarized in Tables-A 8 and 9 (in Appendix), respectively.

3.3.1 *Generalization to Unseen Domain*

As shown in Fig. 11, serious domain shifts exist among source domains and target domain, which easily leads to poor performance on biased target dataset (e.g., MSU-MFSD) when training deep models directly on sources datasets (e.g., OULU-NPU, CASIA-MFSD, and Replay-Attack). *Domain adaptation* technique leverages the knowledge from target domain to bridge the gap between source and target domains. In contrast, *domain generalization* helps the FAS model learn generalized feature representation from multiple source domains directly without any access to target data, which is more practical for real-world deployment.

Domain Adaptation. Domain adaptation technique alleviates the discrepancy between source and target domains. The distribution of source and target features are usually

matched in a learned feature space. If the features share similar distributions, a classifier trained on features of the source samples can be used to classify the target samples.

To align the features space between source and target domain data, Li et al. [78] propose the unsupervised domain adaptation to learn a mapping function to align the source-target embedded subspaces via minimizing their Maximum Mean Discrepancy (MMD) [145]. To further enhance the generalization between both domains, UDA-Net [146], [147] is proposed with unsupervised adversarial domain adaptation in order to jointly optimize the source and target domain encoders. The domain-aware common feature space is learned when the features cannot be distinguished from both domains. The domain-invariant features constrained via MMD and adversarial domain adaptation are still with weak discrimination capacity because the label information in the target domain is inaccessible. To alleviate this problem, semi-supervised learning was introduced to domain adaptation by two works [103], [148], where a few labeled data and a large amount unlabeled data in the target domain can be utilized. Jia et al. [103] propose a unified unsupervised and semi-supervised domain adaptation network to represent the domain-invariant feature space, and find that leveraging few labeled target data (three to five) can significantly improve the performance on the target domain. Similarly, Quan et al. [148] propose a semi-supervised learning FAS method using only a few labeled training data for pretraining, and progressively adopt reliable unlabeled data during training to reduce the domain gap. Despite excellent adaptation, such semi-supervised methods heavily rely on class-balanced few-shot labeled data (i.e., with both live and spoof samples simultaneously), and performance degrade obviously when labeled spoof samples are unavailable.

Different from the above-mentioned methods only adapting the final classifier layer, there are a few works designing and adapting the whole FAS networks. As different deep layers share different granularities of domain information, Authors of [149] consider multi-layer distribution adaptation on both the representation layers and the classifier layer with MMD loss. Despite efficient adaptation via multi-level clues, the architecture might be redundant and have limited generalization capacity per se. To obtain more generalized architectures, Mohammadi et al. [150] propose to prune the filters with high feature divergence that do not generalize well from one dataset to another, thus the performance of the network on target domain can be improved. Different from the network pruning in specific filters/layers, Li et al. [151] propose to distill the whole FAS model for the application-specific domain from a well-trained teacher network, which is regularized with feature MMD and pair-wise similarity embedding from both domains. In this way, lightweight yet generalized FAS models could be discovered but with weaker discrimination capacities compared with teacher FAS networks.

Although domain adaptation benefits to minimize the distribution discrepancy between the source and the target domain by utilizing unlabeled target data, in many real-world FAS scenarios, it is difficult and expensive to collect a lot of unlabeled target data (especially the spoofing attacks) for training. In addition, in consideration of the privacy issue, the source face data are usually inaccessible when deploying FAS models on the target domain.

Domain Generalization. Domain generalization assumes that there exists a generalized feature space underlying the seen multiple source domains and the unseen but related target domain, on which the learned model from the seen source domains can generalize well to the unseen domain.

On one hand, a few works adopt domain-aware adversarial constraints to learn discriminative but domain-unrelated features. They assume that such features contain intrinsic clues across all seen domains thus would generalize well on unseen domain. Shao et al. [48] are the first to propose to learn a generalized feature space shared by multiple source domains via a multi-adversarial discriminative domain generalization framework. Meanwhile, a domain generalization benchmark across four FAS datasets is also established in [48]. However, there are two limitations: 1) such domain-independent features might still contain spoof-unrelated clues (e.g., subject ID and sensor noise); and 2) the discrimination of the domain generalized features is still unsatisfactory. To improve the first limitation, Wang et al. [50] propose to disentangle generalized FAS features from subject discriminative and domain-dependent features. As for the second limitation, in consideration of the large distribution discrepancies among spoof faces of different domains, Jia et al. [51] propose to learn a discriminative and generalized feature space, where the feature distribution of the bonafide is compact while that of the PAs is dispersed among domains but is compact within each domain.

On the other hand, several representative works utilize domain-aware meta-learning to learn the domain generalized feature space. Specifically, faces from partial source domains are used as query set while those from remained non-overlap domains as support set. Based on this setting, Shao et al. [49] propose to regularize the FAS model by finding generalized learning directions in the fine-grained domain-aware meta-learning process. To alternatively force the meta-learner to perform well on support sets (domains), the learned models have robust generalization capacity. However, such domain-aware meta learning strictly needs the domain labels of the source data to construct the query and support sets but domain labels are not always available in real-world cases. Without using domain labels, Chen et al. [152] propose to train a generalized FAS model using the domain dynamic adjustment meta-learning, which iteratively divides mixture domains into clusters with pseudo domain labels. However, the spoof-discriminative and domain-aware features are disentangled via a simple channel attention module, making the domain features unreliable for pseudo domain label assignment. From the perspective of feature normalization, based on the evidence that instance normalization is effective to remove domain discrepancy, Liu et al. [153] propose to adaptively aggregate batch and instance normalizations for generalized representation via meta-learning. Note that the adaptive tradeoffs between batch and instance normalizations might weaken the live/spoof discrimination capacity.

Overall, domain generalization for FAS is a new hot spot in recent three years, and some potential and exciting trends such as combining domain generalization with adap-

tation [154], and learning without domain labels [152] are investigated. However, there still lacks of the works lifting the veil about discrimination and generalization capacities. In other words, domain generalization benefits FAS models to perform well in unseen domain, but it is still unknown whether it deteriorates the discrimination capability for spoofing detection under the seen scenarios.

3.3.2 Generalization to Unknown Attack Types

Besides domain shift issues, FAS models are vulnerable to emerging novel PAs in real-world practical applications. Most previous deep learning methods formulate FAS as a close-set problem to detect various pre-defined PAs, which need large-scale training data to cover as many attacks as possible. However, the trained model can easily overfit several common attacks (e.g., print and replay) and is still vulnerable to unknown attack types (e.g., mask and makeup). Recently, many researches focus on developing generalized FAS models for unknown spoofing attack detection. On one side, *zero/few-shot learning* is employed for improving novel spoofing detection with very few or even none samples of target attack types. On the other side, FAS can also be treated as a *one-class classification* task where the bonafide samples are clustered compactly, and *anomaly detection* is used for detecting the out-of-distribution PA samples.

Zero/Few-Shot Learning. One straightforward way for novel attack detection is to finetune the FAS model with sufficient samples of the new attacks. However, collecting labeled data for every new attack is expensive and time-consuming since the spoofing keeps evolving. To overcome this challenge, several works [38], [53], [155] propose to treat FAS as an open-set zero- and few-shot learning problem. *Zero-shot learning* aims to learn generalized and discriminative features from the predefined PAs for unknown novel PA detection. *Few-shot learning* aims to quickly adapt the FAS model to new attacks by learning from both the predefined PAs and the collected very few samples of the new attack.

Without any prior knowledge of the unknown spoof attacks, Liu et al. [38] design a Deep Tree Network (DTN) to learn the semantic attributes of pre-defined attacks and partition the spoof samples into semantic sub-groups in an unsupervised fashion. Based on the similarity of the embedding features, DTN adaptively routes the known or unknown PAs to the corresponding spoof clusters. The live/spoof tree topology is constructed via DTN automatically, which is more semantic and generalized compared with the human-defined category relationship. However, without any prior knowledge of unknown attacks, the zero-shot DTN may fail to detect novel high-fidelity attacks. To alleviate this issue, two works adopt an open-set few-shot framework to introduce partial yet effective unknown attack knowledge for representation learning. Qin et al. [53] unify the zero- and few-shot FAS tasks together by fusion training a meta-learner with an adaptive inner-update learning rate strategy. Training meta-learner on both zero- and few-shot tasks simultaneously enhances the discrimination and generalization capacities of FAS models from pre-defined PAs and few instances of the new PAs. However, directly using few-shot meta learning on novel attacks easily suffers from catastrophic forgetting about the pre-defined PAs. To tackle this issue, Perez-Cabo et al. [155] propose a continual few-shot learning paradigm, which incrementally extends the acquired knowledge from the continuous stream of data, and detects new PAs using a small number of training samples via a meta-learning solution.

Although few-shot learning benefits the FAS models for unknown attack detection, the performance drops obviously when the data of the target attack types are unavailable for adaptation (i.e., zero-shot case). We observe that the failed detection usually occurs in the challenging attack types (e.g., transparent mask, funny eye, and makeup), which share similar appearance distribution with the bonafide.

Anomaly Detection. Anomaly detection based FAS assumes that the live samples are in a normal class as they share more similar and compact feature representation while features from the spoof samples have large distribution discrepancies in the anomalous sample space due to the high variance of attack types and materials. Based on the assumption, anomaly detection usually firstly trains a reliable one-class classifier to accurately cluster the live samples. Then any samples (e.g., unknown attacks) outside the margin of the live sample cluster would be detected as attacks.

Arashloo et al. [52] is the first to evaluate one-class anomaly detection and traditional binary classification FAS systems on cross-type testing protocols. They find that anomaly-based methods using one-class SVM are not inferior compared to binary classification approaches using two-class SVM. To better represent the probability distribution of bonafide samples, Nikisins et al. [156] propose to replace traditional one-class SVM with Gaussian Mixture Model (GMM) as the anomaly detector. Besides one-class SVM and GMM, Xiong and AbdAlmageed [157] also consider the autoencoder based outliers detector with LBP feature extractor for open-set unknown PAD. The above-mentioned works separate the feature extraction with the one-class classifier, which makes the bonafide representation learning challenging and sub-optimal. In contrast, Baweja et al. [158] present an end-to-end anomaly detection approach to train the one-class classifier and feature representations together. Moreover, to learn robust bonafide representation against out-of-distribution perturbations, they generate pseudo negative features to mimic the PA class and force the one-class classifier to be discriminative for PAD. However, the generated pseudo PA features cannot represent diverse real-world features, making the one-class anomaly detection system less reliable for real-world deployment.

Though reasonable, utilizing only live faces to train the classifier usually limits the anomaly detection model's generalization on new PA types. Instead of using only live samples, some works train the generalized anomaly detection systems with both live and spoof samples via metric learning. Pérez-Cabo et al. [159] propose to regularize the FAS model by a triplet focal loss to learn discriminative feature representation, and then introduce a few-shot posteriori probability estimation as anomaly detector for unknown PA detection. Similarly, George and Marcel [160] design a pair-wise one-class contrastive loss (OCCL) to force the network to learn a compact embedding for bonafide class while being far from the representation of attacks. Then

an one-class GMM is cascaded for unknown PA detection. Although discriminative embedding could be learned via triplet or contrastive loss, the works [159], [160] still need extra anomaly detectors (e.g., one-class GMM) cascaded after embedding features, which influences the end-to-end representation learning. In contrast, Li et al. [161] propose to supervise deep FAS models with a novel hypersphere loss to keep the intra-class live compactness as well as inter-class live-spoof separation. The unknown attacks could be directly detected on learned feature space with no need of additional anomaly detection classifiers. One limitation is that the predicted live/spoof score is calculated from the square of L2-norm of the embedding features, which is hard to select a suitable predefined threshold for detecting different kinds of attack types.

Despite satisfactory generalization capacity for unknown attack detection, anomaly detection based FAS methods would suffer from discrimination degradation compared with conventional live/spoof classification in the real-world open-set scenarios (i.e., both known and unknown attacks).

4 Deep FAS with Advanced Sensors

Commercial RGB camera-based FAS would be an excellent tradeoff solution in terms of security and hardware cost in daily face recognition applications. However, some high-security scenarios (face payment and vault entrance guard) require very low false acceptance errors. Recently, advanced sensors with various modalities are developed to facilitate the ultra-secure FAS. Merits and demerits of various sensors and hardware modules for FAS in terms of environmental conditions (lighting and distance) and attack types (print, replay, and 3D mask) are listed in Table 2.

Compared with monocular visible RGB camera (VIS), stereo cameras (VIS-Stereo) [162] benefits the 3D geometry information reconstruction for 2D spoofing detection. When assembling with dynamic flash light on the presentation face, VIS-Flash [163] is able to capture intrinsic reflection-based material clues to detect all three attack types.

Besides visible RGB modality, depth and NIR modalities are also widely used in practical FAS deployment with acceptable costs. Two kinds of depth sensors including Time of Flight (TOF) [164] and 3D Structured Light (SL) [165] have been embedded in mainstream mobile phone platforms (e.g., Iphone, Sumsung, OPPO, and Huawei). They provide the accurate 3D depth distribution of the captured face for 2D spoofing detection. Compared with SL, TOF is more robust to environmental conditions such as lighting and distance. In contrast, NIR [166] modality is a complementary spectrum (900 to 1800nm) besides VIS, which effectively exploits reflection differences between live and spoof faces but is with poor imaging quality in long distance. In addition, the VIS-NIR integration hardware module is with a high performance-price ratio for many access control systems.

Meanwhile, several niche but effective sensors are introduced in FAS. Shortwave infrared (SWIR) [55] with the wavelengths of 940nm and 1450nm bands discriminates live skin material from non-skin pixels in face images via measuring water absorption, which is reliable for generic spoofing attacks detection. A thermal camera [167] is an alternative sensor for efficient FAS via face temperature estimation. However, it performs poorly when subjects wear transparent masks. Expensive Light Field camera [87] and four-directional Polarization sensor [47] are also explored for FAS according to their excellent representation for facial depth and reflection/refraction light, respectively.

TABLE 2: Comparison with sensor/hardware for FAS under 2 environments (lighting condition and distance) and 3 attack types (print, replay and 3D mask). 'TOF', 'SL', 'C', 'M', 'E', 'P', 'G', 'VG' are short for 'Time of Flight', ' Structured Light', 'Cheap', 'Medium', 'Expensive', 'Poor', 'Good', 'Very Good', respectively.

Sensor	Cost	Environment		Attack Type		
		Lighting	Distance	Print	Replay	Mask
VIS	C	M	M	M	M	M
VIS-Stereo	M	M	M	VG	VG	M
VIS-Flash	C	M	M	G	G	M
Depth(TOF)	M	M	G	VG	VG	P
Depth(SL)	C	P	P	VG	VG	P
NIR	C	G	P	G	VG	M
VIS-NIR	M	G	M	G	VG	G
SWIR	E	G	M	VG	VG	G
Thermal	E	G	M	G	VG	M
Light Field	E	P	M	VG	VG	M
Polarization	E	G	M	VG	VG	G

4.1 Uni-Modal Deep Learning upon Specialized Sensor

Based on the specialized sensor/hardware for distinct imaging, researchers have developed sensor-aware deep learning methods for efficient FAS, which are summarized in Table-A 10 (in Appendix). Seo and Chung [167] propose a lightweight Thermal Face-CNN to estimate the facial temperature from the thermal image, and detect the spoofing with abnormal temperature (e.g., out of scope from 36 to 37 degrees). They find that the thermal image is more suitable than the RGB image for replay attack detection. However, such thermal-based method is vulnerable to the transparent mask attack. In terms of stereo-based FAS, several works [162], [168], [169] prove that leveraging the estimated disparity or depth/normal maps from the stereo inputs (from stereo and dual pixel (DP) sensors) via CNN could achieve remarkable performance on 2D print and replay attack detection. However, it usually performs poorly on the 3D mask attack with similar geometric distribution of live faces. To further capture detailed 3D local patterns, Liu et al. [87] propose to extract the ray difference and microlens images from a single-shot light field camera, and then a shallow CNN is used for face PAD. Due to the rich 3D information in light field imaging, the method is potential to classify fine-grained spoofing types. Towards real-time and mobile-level deployment, Tian et al. [47] propose to use lightweight MobileNetV2 to extract efficient DOLP features from an on-chip integrated polarization imaging sensor. The above-mentioned methods aim at tackling specific PA types (e.g., replay and print), which cannot generalize well across all PA types. In contrast, Heusch et al. [55] propose to use a multi-channel CNN for deep material-related feature extraction from the selected SWIR-difference inputs, which is able to almost perfectly detect all impersonation attacks while ensuring low bonafide classification errors.

Apart from using specialized hardware such as infrared dot projectors and dedicated cameras, some deep FAS meth-

ods are developed based on visible cameras with extra environmental flash. In [163] and [170], dynamic flash from the smartphone screen is utilized to illuminate a user's face from multiple directions, which enables the recovery of the face surface normals via photometric stereo. Such dynamic normal cues are then fed into CNN to predict facial depth and light CAPTCHA for PA detection. Similarly, Ebihara et al. [171] design a novel descriptor to represent the specular and diffuse reflections leveraging the difference cues with and without flash, which outperforms the end-to-end ResNet with concatenated flash inputs. These methods are easy to deploy without extra hardware integration, and have been used in mobile verification and payment systems such as Alipay and WeChat Pay. However, dynamic flash is sensitive under outdoor environments and is not user-friendly due to the long temporal activation time.

4.2 Multi-Modal Deep Learning

Meanwhile, multi-modal learning based methods become hot and active in the FAS research community. Representative multi-modal fusion and cross-modal translation approaches for FAS are collected in Table-A 11 (in Appendix).

Multi-Modal Fusion. With multi-modal inputs, mainstream FAS methods extract complementary multi-modal features using feature-level fusion strategies. As there are redundancy across multi-modal features, direct feature concatenation easily results in high-dimensional features and overiftting. To alleviate this issue, Zhang et al. [28] propose the SD-Net using a feature re-weighting mechanism to select the informative and discard the redundant channel features among RGB, depth, and NIR modalities. However, the re-weighting fusion in SD-Net is only conducted on the high-level features but neglecting the multi-modal low-level clues. To further boost the multi-modal feature interaction at different levels, authors from [172] and [173] introduce a multi-modal multi-layer fusion branch to enhance the contextual clues among modalities. Despite advanced fusion strategies, multi-modal fusion is easily dominated by partial modalities (e.g., depth) thus performs poorly when these modalities are noisy or missing. To tackle this issue, Shen et al. [174] design a Modal Feature Erasing operation to randomly dropout partial-modal features to prevent modality-aware overfitting. In addition, George and Marcel [175] present a cross-modal focal loss to modulate the loss contribution of each modality, which benefits the model to learn complementary information among modalities. Overall, feature-level fusion is flexible and effective for multi-modal clue aggregation. However, modality features are usually extracted from separate branches with high computational cost.

Besides feature-level fusion, there are a few works that consider input-level and decision-level fusions. Input-level fusion assumes that multi-modal inputs are already aligned spatially, and can be fused in the channel dimension directly. In [176], the composite image is fused from gray-scale, depth, and NIR modalities by stacking the normalized images, and then fed to deep PA detectors. Similarly, Liu et al. [177] composite VIS-NIR inputs via different fusion operators (i.e., stack, summation, and difference), and all fused face images are forwarded by a multi-modal FAS network for live/spoof prediction. These input-level fusion methods are efficient and with a little extra computational cost (mostly on fusion operator and the first network layer). However, fusion in too early stage easily vanishes multi-modal clues in the subsequent mid- and high-level spaces. In contrast, to tradeoff the individual modality bias and make reliable binary decision, some works adopt decision-level fusion based on the predicted score from each modality branch. On one hand, Yu et al. [27] directly average the predicted binary scores of individual models from RGB, depth, and NIR modalities, which outperforms the input- and feature-level fusions on CeFA [91] dataset. On the other hand, Zhang et al. [178] design a decision-level fusion strategy to firstly aggregate scores from several models using depth modality, and then cascaded with the score from the IR model for final live/spoof classification. Despite reliable prediction, decision-level fusion is inefficient as it needs separate well-trained models for particular modalities.

Cross-Modal Translation. Multi-modal FAS system needs additional sensors for imaging face inputs with different modalities. However, in some conventional scenarios, only partial modalities (e.g., RGB) can be available. To tackle this modality missing issues at the inference stage, a few works adopt the cross-modal translation technique to generate the missing modal data for multi-modal FAS. To generate the corresponding NIR images from RGB face images, Jiang et al. [179] first propose a novel multiple categories (live/spoof, genuine/synthetic) image translation cycle-GAN. Based on the generated NIR and original RGB inputs, the method is able to extract more robust fused features compared with using only the RGB images. However, the generated NIR images from raw cycle-GAN are with low quality, which limits the performance of the fused features. To generate high-fidelity target NIR modality, Liu et al. [180] design a novel subspace-based modality regularization in the cross-modal translation framework. Besides generating the NIR images, Mallat and Dugelay [181] propose a visible-to-thermal conversion scheme to synthesize thermal attacks from RGB face images using a cascaded refinement network. Though effectiveness on intra-dataset testings, one main concern of these methods is that the domain shifts and unknown attacks might significantly influence the generated modality's quality, and the fused features would be unreliable using paired noisy modality data.

Despite a rising trend since 2019, the progress of sensor-based multi-modal FAS is still slow compared with RGB based unimodal methods. It is worth noting that multi-modal approaches also exist in deep FAS with commercial RGB camera. For instance, decision-level fusion of two RGB video based modalities (i.e., remote physiological signals and face visual image) has been explored in [14]. Therefore, to effectively fuse such natural modalities from commercial RGB camera with those from advanced sensors will be an interesting and valuable direction. Meanwhile, some advanced sensors (e.g., SWIR, light field, and polarization) are expensive and non-portable for real-world deployment. More efficient FAS-dedicated sensors as well as multi-modal approaches should be explored.

5 Discussion and Future Directions

Thanks to the recent advances in deep learning, FAS has achieved rapid improvement over the past few years. As can be seen from Fig. 5, recent deep FAS methods refresh the state of the arts and obtain satisfied performance (e.g., <5% ACER, <15% HTER, <10% EER, and <20% HTER) on four evaluation protocols, respectively. On one hand, advanced architectures (e.g., NAS-FAS [37] and FAS-SGTD [96]) and pixel-wise supervision (e.g., pseudo depth and reflection maps) benefit the 2D attack detection as well as the fine-grained spoof material perception (e.g., silicone and transparent 3D masks). On the other hand, domain and attack generalization based methods (e.g., SSDG [51], FGHV [182], and SSAN [102]) mine the intrinsic live/spoof clues across multiple source domains and attack types, which can generalize well even on unseen domains and unknown attacks. These generalized deep learning based methods usually detect different kinds of attacks (2D & 3D) under diverse scenarios more stably (with lower standard deviation errors) under leave-one-out cross-testing protocols. Furthermore, some insightful conclusions could be drawn from Tables-A 2, 3, and 4 (in Appendix): 1) Advanced architectures (e.g., DC-CDN [98]) with elaborate supervisions (e.g., pseudo depth supervision) dominate the testing performance when training on single source domain. In contrast, when training on multiple (three) domains, generalized learning strategies play more important roles. 2) Transfer learning from large-scale pre-trained models (e.g., SSAN [102] and ViTranZ-FAS [38] using ResNet18 and vision transformer pretrained from ImageNet1K and ImageNet21K, respectively) alleviates the overfitting issue caused by limited-scale live/spoof data, thus improves the generalization capacity and benefits cross-dataset and cross-type testings.

However, FAS is still an unsolved problem due to the challenges such as subtle spoof pattern representation, complex real-world domain gaps, and rapidly iterative novel attacks. We conclude the limitations of the current development as follows: 1) *Limited live/spoof representation capacity with sub-optimal deep architectures, supervisions, and learning strategies*. Learning discriminative and generalized live/spoof features is vital for deep FAS. Until now, it is still hard to find the best-suited architectures as well as the supervisions across all different evaluation benchmarks. For example, CDCN with pixel-wise supervision achieves excellent and poor performance on intra-dataset and multi-source-domain cross-dataset testings, respectively, while ResNet with binary cross-entropy loss performs inversely. 2) *Evaluation under saturating and unpractical testing benchmarks and protocols*. For example, for intra-testing on the OULU-NPU dataset, ACER of 0.4% and 0.8% might make slight difference and indicate the performance saturation on such small-scale and monotonous test set. And the cross-domain testings are still far from real-world scenarios as only limited sorts of attack types are considered. 3) *Isolating the anti-spoofing task on only the face area and physical attacks*. Besides physical presentation attacks in the face area, spoofing in more general applications (e.g., commodity and document) and even digital attacks via stronger and stronger face swapping and generative models should be considered. These tasks might share partial intrinsic knowledge and benefit the representation learning. 4) *Insufficient consideration about the interpretability and privacy issues*. Most existing FAS researches devote to developing novel algorithms against state-of-the-art performance but rarely think about the interpretability behind. Such black-box methods are hard to make reliable decisions in real-world cases. In addition, most existing works train and adapt deep FAS models with huge stored source face data, and neglect the privacy and biometric sensitivity issue. According to the discussion above, we summarize some solutions and potential research directions in the following subsections.

5.1 Architecture, Supervision and Interpretability

As can be seen from Sections 3 and 4, most of the researchers choose the off-the-shelf network architectures as well as handcrafted supervision signals for deep FAS, which might be sub-optimal and hard to leverage the large-scale training data adequately. Although several recent works have applied AutoML in FAS for searching well-suited architecture [23], [37], loss function [54], and auxiliary supervision [43], they focus on uni-modality and single-frame configuration while neglecting the temporal or multi-modal situation. Hence, one promising direction is to automatically search and find the best-suited temporal architectures especially for multi-modal usage. In this way, more reasonable fusion strategies would be discovered among modalities instead of coarse handcrafted design. In addition, rich temporal context should be considered in dynamic supervision design instead of static binary or pixel-wise supervision.

On the other hand, to design *efficient* network architecture is vital for real-time FAS in mobile devices. Over the past years, most research focuses on tackling the accuracy and generalization issues in FAS while only a few works consider lightweight [143] or distilled [151] CNNs for efficient deployment. Besides CNN with strong inductive bias, researchers should also rethink the usage of some flexible architectures (e.g., vision transformer [115], [129]) in terms of efficiency and computational cost.

Recently, great efforts have been achieved on *interpretable* FAS [183]. Some methods try to localize the spoof regions according to the feature activation using visual interpretability tools (e.g., Grad-CAM [184]) or soft-gating strategy [131]. In addition, auxiliary supervised [13], [24] and generative [33], [42] FAS models devote to estimating the underlying spoof maps. Besides the visual activation maps, natural language [185] has been introduced for explaining the FAS predictions with meaningful sentence-level descriptions. All these trials help researchers understand and localize the spoof patterns, and convince the FAS decision. However, due to the lack of precious pixel-level spoof annotation, the estimated spoof maps are still coarse and easily influenced by unfaithful clues (e.g., hands). More advanced feature visualization manners and fine-grained pixel-wise spoof segmentation should be developed for interpretable FAS.

5.2 Representation Learning

Learning discriminative and intrinsic feature representation is the key to reliable FAS. A handful of previous researches have proven the effectiveness of transfer learning [127], [172] and disentangled learning [42], [97] for FAS. The

former leverages the pre-trained semantic features from other large-scale datasets to alleviate the overfitting issue, while the latter aims to disentangle the intrinsic spoofing clues from the noisy representation. To learn discriminative embedding spaces with compact distributions among live faces and distinguishable distances between live/spoof faces, deep metric learning is used for training FAS models. However, the uncertainty of the model prediction is still high in the extreme/noisy scenario (e.g., presenting with very high-quality spoof and low-quality live samples). More advanced metric learning techniques (e.g., on hyperbolic manifold space) could be explored in the future for mining subtle spoof patterns. Moreover, rephrasing FAS as a fine-grained recognition [24], [101] problem to learn type-discriminative representation is worth exploring, which is inspired by the fact that humans could detect spoofing via recognizing the specific attack types.

Researchers should also get hung up on fully exploiting the live/spoof training data with or without labels for representation enhancement. On one side, self-supervised on large-scale combined datasets might reduce the risk of overfitting, and actively mine the intrinsic knowledge (e.g., high similarity among intra face patches). On the other side, in real-world scenarios, daily unlabeled face data are collected from various face recognition terminals continuously, which could be utilized for semi-supervised learning [148]. One challenge is how to make full use of the unlabeled imbalanced (i.e., live $\gg$ spoof) data, avoiding unexpected performance drop. In addition, suitable data augmentation strategies [98] for FAS are rarely investigated. Adversarial learning might be a good choice for adaptive data augmentation in more diverse domains.

5.3 Real-World Open-Set FAS

As discussed in Section 2.4, traditional FAS evaluation protocols usually consider intra-domain [77], cross-domain [48], and cross-type [38] testings within one or several small-scale datasets. The state-of-the-art methods in such protocols cannot guarantee consistently good performance in practical scenarios because 1) the data amount (especially testing set) is relatively small thus the high performance is not very convincing; and 2) the protocols focus on a single factor (e.g., seen/unseen domains or known/unknown attack types), which cannot satisfy the need of complex real-world scenarios. Recently, more practical protocols such as GrandTest [155] and open-set [42], [59] are proposed. GrandTest contains large-scale mixed-domain data, while open-set testing considers models' discrimination and generalization capacities on both known and unknown attack types. However, real-world open-set situations with simultaneous domains and attack types are still neglected. More comprehensive protocols (e.g., domain- and type-aware open-set) should be explored for fair and practical evaluation to bridge the gap between academia and industry.

As for the multi-modal protocols, training data with multiple modalities are assumed available, and two testing settings are widely used: 1) with corresponding multiple modalities [186]; and 2) only single modality [175], [180] (usually RGB). However, there are various kinds of modality combinations [187] (e.g., RGB-NIR, RGB-D, NIR-D, and RGB-D-NIR) in real-world deployment according to different user terminal devices. Therefore, it is pretty costly and inefficient to train individual models for each multi-modal combination. Although pseudo modalities could be generated via cross-modality translation [179], [180], their fidelity and stability are still weaker compared with modalities from real-world sensors. To design a dynamic multi-modal framework to propagate the learned multi-modal knowledge to various modality combinations might be a possible direction for unlimited multi-modal deployment.

Fig. 12: Illustration of the physical adversarial faces generated by Adv-glasses [189], Adv-hat [190], Adv-makeup [191], and Adv-sticker [192].

5.4 Generic and Unified PA Detection

Understanding the intrinsic property of face PAD with other related tasks (e.g., generic PAD, and digital face attack detection) is important for explainable FAS. On one hand, 'generic' assumes that both face and other object presentation attacks might have independent content but share intrinsic spoofing patterns [188]. For instance, replay attacks about different objects (e.g., a face and a football) are made of the same glass material [24], and with abnormal reflection clues. Thus, generic PAD and material recognition datasets could be introduced in face PAD for common live/spoof feature representation in a multi-task learning fashion.

Apart from common PAs, two kinds of physical adversarial attacks (AFR-aware and FAS-aware) should be considered for generic PAD. As illustrated in Fig. 12, physical eyeglass [189] and hat [190] achieved from adversarial generators, or special stickers [192] containing feature patterns proved to be effective against deep learning based AFR systems can be printed out and wore by attackers to spoof such systems. Moreover, imperceptible makeup [191] nearby the eye regions have been verified for attacking commercial AFR systems. Besides AFR-aware adversarial attacks, adversarial print/replay attacks [193] with perturbation before physical broadcast are developed to fool the FAS system. Therefore, it is expected and necessary to establish large-scale FAS datasets with diverse physical adversarial attacks as well as annotated attack localization labels.

On the other hand, besides physical face presentation attacks, there are many vicious digital manipulation attacks (e.g., Deepfake [194]) and morphing attacks (e.g., via generative model StyleGAN [195]) on face videos. As generative models become stronger and stronger, these direct digital attacks from generative models become bigger threats. Despite different generation manners with diverse attack traces and visual qualities, parts of these attacks might still have coherent properties. In [**?**], [196], a unified digital and physical face attack detection framework is proposed to learn joint representations for coherent attacks. However, there are serious imbalanced numbers among digital and physical attack types due to data collection costs. In other

words, large-scale digital attacks are easier to generate compared with high-cost presentation attacks. Such imbalanced distribution might harm the intrinsic representation during the multi-task learning, which needs to think about in the future.

5.5 Privacy-Preserved Training

Leveraging large-scale live/spoof face data, deep learning based FAS has achieved huge breakthroughs. However, the legal and privacy issues of the face data attract more and more attention. For example, the GDPR (General Data Protection Regulation) [197], came into effect in May 2018, brings the importance of preserving the privacy of personal information (e.g., face images) to the forefront. Therefore, a noteworthy direction is to alleviate the privacy issue (i.e., storing/sharing large-scale users' face data) but maintaining satisfied performance for deep FAS models.

On one hand, the live/spoof face training data are usually not directly shared between data owners (domains). To tackle this challenge, federate learning [198], a distributed and privacy-preserving machine learning technique, is introduced in FAS to simultaneously take advantage of rich live/spoof information available at different data owners while maintaining data privacy. To be specific, each data center/owner locally trains its own FAS model. Then a server learns a global FAS model by iteratively aggregating model updates from all data centers without accessing original private data in each of them. Finally, the converged global FAS model would be utilized for inference. To enhance the generalization ability of the server model, in [199], a federated domain disentanglement strategy is introduced, which treats each data center as one domain and decomposes the FAS model into domain-invariant and domain-specific parts in each data center. Overall, the existing federated learning based FAS usually focuses on the privacy problem of *data* sets but neglects the privacy issues in the *model* level. Thus, the training of the global model needs multiple teams to share their own local models, which might harm the commercial competition.

On the other hand, due to privacy and security concerns of human faces, source data are usually inaccessible during adaptation for practical deployment. Specifically, in a source-free [200] setting, a FAS model is first pre-trained on the (large-scale) source data and is released for deployment. In the deployment phase, the source data cannot be shared for adapting the pre-trained model to the target data, as they contain sensitive biometric information. Lv et al. [201] benchmark the source-free setting for FAS via directly applying a self-training approach, which easily obtains noisy target pseudo labels due to the challenges in the FAS task (e.g., the intra-class distance between live faces of different identities probably exceeds the inter-class distance between live and spoof faces of the same identity). Thus, the performance gain (1.9% HTER reduction on average) by adaptation is quite limited. To efficiently and accurately adapt the source knowledge without accessing source data is worth exploring in the future.

6 Conclusion

This paper has presented a contemporary survey of the deep learning based methods, datasets as well as protocols for face anti-spoofing (FAS). A comprehensive taxonomy of these methods have been presented. Merits and demerits of various methods and sensors for FAS are also covered, with potential research directions being listed.

Acknowledgments This work was supported by the Academy of Finland (Academy Professor project EmotionAI with grant numbers 336116 and 345122, and ICT2023 project with grant number 345948), the National Natural Science Foundation of China (No. 61876178, 61976229, and 62106264), and Beijing Academy of Artificial Intelligence (BAAI).

References

[1] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," in *CVPR*, 2020.
[2] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *ICCV*, 2007.
[3] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *ICPR*, 2016.
[4] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp- top based countermeasure against face spoofing attacks," in *ACCV*, 2012.
[5] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face anti-spoofing," in *BTAS*, 2013.
[6] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *TIFS*, 2016.
[7] H.-K. Jee, S.-U. Jung, and J.-H. Yoo, "Liveness detection for embedded face recognition system," *International Journal of Biological and Medical Sciences*, 2006.
[8] J.-W. Li, "Eye blink detection based on multiple gabor response waves," in *ICMLC*, vol. 5. IEEE, 2008, pp. 2852–2856.
[9] L. Wang, X. Ding, and C. Fang, "Face live detection method based on physiological motion analysis," *Tsinghua Science & Technology*, vol. 14, no. 6, pp. 685–690, 2009.
[10] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *ICASSP*, 2009.
[11] J. Bigun, H. Fronthaler, and K. Kollreider, "Assuring liveness in biometric identity authentication by real-time face tracking," in *CIHSPS*. IEEE, 2004.
[12] A. Ali, F. Deravi, and S. Hoque, "Liveness detection using gaze collinearity," in *ICEST*. IEEE, 2012.
[13] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *CVPR*, 2018.
[14] B. Lin, X. Li, Z. Yu, and G. Zhao, "Face liveness detection by rppg features and contextual patch-based cnn," in *ICBEA*. ACM, 2019.
[15] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement," in *ICCV*, 2019.
[16] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *ICIP*, 2015.
[17] Boulkenafet, Zinelabidine and Komulainen, Jukka and Hadid, Abdenour, "Face antispoofing using speeded-up robust features and fisher vector encoding," *SPL*, vol. 24, no. 2, pp. 141–145, 2016.
[18] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *ECCV*. Springer, 2010.
[19] X. Song, X. Zhao, L. Fang, and T. Lin, "Discriminative representation combinations for accurate face spoofing detection," *Pattern Recognition*, 2019.
[20] M. Asim, Z. Ming, and M. Y. Javed, "Cnn based spatio-temporal feature extraction for face anti-spoofing," in *ICIVC*. IEEE, 2017.
[21] Y. A. U. Rehman, L.-M. Po, and J. Komulainen, "Enhancing deep discriminative feature maps via perturbation for face presentation attack detection," *Image and Vision Computing*, vol. 94, p. 103858, 2020.
[22] M. Khammari, "Robust face anti-spoofing using cnn with lbp and wld," *IET Image Processing*, 2019.
[23] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *CVPR*, 2020.

[24] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *ECCV*, 2020.

[25] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *CVPR*, 2019.

[26] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *IJCB*, 2017.

[27] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *CVPRW*, 2020.

[28] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, "Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing," *TBIOM*, vol. 2, no. 2, pp. 182–193, 2020.

[29] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:1408.5601*, 2014.

[30] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *IPTA*, 2016.

[31] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *CCBR*, 2016.

[32] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *ICB*, no. CONF, 2019.

[33] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *ECCV*, 2018.

[34] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3d face anti-spoofing with factorized bilinear coding," *arXiv preprint arXiv:2005.06514*, 2020.

[35] L. Li, Z. Xia, X. Jiang, F. Roli, and X. Feng, "Compactnet: learning a compact space for face presentation attack detection," *Neurocomputing*, 2020.

[36] T. Kim, Y. Kim, I. Kim, and D. Kim, "Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing," in *ICCVW*, 2019.

[37] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "Nas-fas: Static-dynamic central difference network search for face anti-spoofing," *IEEE TPAMI*, pp. 1–1, 2020.

[38] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *CVPR*, 2019.

[39] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot, "Face spoofing detection based on local ternary label supervision in fully convolutional networks," *TIFS*, 2020.

[40] X. Li, J. Wan, Y. Jin, A. Liu, G. Guo, and S. Z. Li, "3dpc-net: 3d point cloud network for face anti-spoofing," 2020.

[41] H. Feng, Z. Hong, H. Yue, Y. Chen, K. Wang, J. Han, J. Liu, and E. Ding, "Learning generalized spoof cues for face anti-spoofing," *arXiv preprint arXiv:2005.03922*, 2020.

[42] Y. Liu and X. Liu, "Physics-guided spoof trace disentanglement for generic face anti-spoofing," *arXiv preprint arXiv:2012.05185*, 2020.

[43] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, "Meta-teacher for face anti-spoofing," *TPAMI*, 2021.

[44] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *ECCV*. Springer, 2020.

[45] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *TIFS*, 2019.

[46] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral swir imaging," in *ICB*. IEEE, 2016.

[47] Y. Tian, K. Zhang, L. Wang, and Z. Sun, "Face anti-spoofing by learning polarization cues in a real-world scenario," *arXiv preprint arXiv:2003.08024*, 2020.

[48] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *CVPR*, 2019.

[49] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *AAAI*, 2020.

[50] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *CVPR*, 2020.

[51] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *CVPR*, 2020.

[52] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE access*, 2017.

[53] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi, and Z. Lei, "Learning meta model for zero-and few-shot face anti-spoofing," in *AAAI*, 2020.

[54] Y. Qin, W. Zhang, J. Shi, Z. Wang, and L. Yan, "One-class adaptation face anti-spoofing with loss function search," *Neurocomputing*, vol. 417, pp. 384–395, 2020.

[55] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," *TBIOM*, 2020.

[56] L. A. Pereira, A. Pinto, F. A. Andaló, A. M. Ferreira, B. Lavi, A. Soriano-Vargas, M. V. Cirne, and A. Rocha, "The rise of data-driven models in presentation attack detection," in *Deep Biometrics*. Springer, 2020, pp. 289–311.

[57] S. Jia, G. Guo, and Z. Xu, "A survey on 3d mask presentation attack detection and countermeasures," *Pattern Recognition*, 2020.

[58] Y. S. El-Din, M. N. Moustafa, and H. Mahdi, "Deep convolutional neural networks for face and iris presentation attack detection: survey and case study," *IET Biometrics*, 2020.

[59] A. Liu, C. Zhao, Z. Yu, J. Wan, A. Su, X. Liu, Z. Tan, S. Escalera, J. Xing, Y. Liang *et al.*, "Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection," *arXiv preprint arXiv:2104.06148*, 2021.

[60] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, 2020.

[61] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa, "Detecting and mitigating adversarial perturbations for robust face recognition," *International Journal of Computer Vision*, 2019.

[62] A. Liu, X. Li, J. Wan, Y. Liang, S. Escalera, H. J. Escalante, M. Madadi, Y. Jin, Z. Wu, X. Yu *et al.*, "Cross-ethnicity face anti-spoofing recognition challenge: A review," *IET Biometrics*, 2021.

[63] J. Hernandez-Ortega, J. Fierrez, A. Morales, and J. Galbally, "Introduction to face presentation attack detection," in *Handbook of Biometric Anti-Spoofing*. Springer, 2019, pp. 187–206.

[64] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?" in *ICB*. IEEE, 2013, pp. 1–8.

[65] L. Li, P. L. Correia, and A. Hadid, "Face recognition under spoofing attacks: countermeasures and research directions," *IET Biometrics*, vol. 7, no. 1, pp. 3–14, 2018.

[66] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019.

[67] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

[68] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *ICIP*. IEEE, 2011.

[69] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *ICB*, 2012.

[70] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Biometrics Special Interest Group*, 2012.

[71] N. Kose and J.-L. Dugelay, "Shape and texture based countermeasure to protect face recognition systems against mask attacks," in *CVPRW*, 2013.

[72] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *TIFS*, 2015.

[73] A. Pinto, W. R. Schwartz, H. Pedrini, and A. de Rezende Rocha, "Using visual rhythms for detecting video-based facial spoof attacks," *TIFS*, 2015.

[74] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *BIOSIG*. IEEE, 2016.

[75] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3d mask face anti-spoofing with remote photoplethysmography," in *ECCV*. Springer, 2016.

[76] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *TIFS*, 2017.

[77] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *FGR*, 2017.

[78] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *TIFS*, 2018.

[79] R. H. Vareto, A. M. Saldanha, and W. R. Schwartz, "The swax benchmark: Attacking biometric systems with wax figures," in *ICASSP*, 2020.

[80] W. R. Almeida, F. A. Andaló, R. Padilha, G. Bertocco, W. Dias, R. d. S. Torres, J. Wainer, and A. Rocha, "Detecting face presentation attacks in mobile devices with a patch-based cnn and a sensor-aware loss function," *PloS one*, 2020.

[81] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3d masks," *TIFS*, 2014.

[82] R. Raghavendra, K. B. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *TIP*, vol. 24, no. 3, pp. 1060–1075, 2015.

[83] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, 2016.

[84] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, "Face recognition systems under spoofing attacks," in *Face Recognition Across the Imaging Spectrum*. Springer, 2016, pp. 165–194.

[85] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *CVPRW*, 2017.

[86] S. Bhattacharjee and S. Marcel, "What you can't see can help you-extended-range imaging for 3d-mask presentation attack detection," in *BIOSIG*. IEEE, 2017.

[87] M. Liu, H. Fu, Y. Wei, Y. A. U. Rehman, L.-m. Po, and W. L. Lo, "Light field-based face liveness detection with convolutional neural networks," *Journal of Electronic Imaging*, 2019.

[88] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *BTAS*, 2018.

[89] J. Xiao, Y. Tang, J. Guo, Y. Yang, X. Zhu, Z. Lei, and S. Z. Li, "3dma: A multi-modality 3d mask face anti-spoofing database," in *AVSS*. IEEE, 2019.

[90] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *CVPR*, 2019.

[91] A. Li, Z. Tan, X. Li, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," *WACV*, 2021.

[92] M. Rostami, L. Spinoulas, M. Hussein, J. Mathai, and W. Abd-Almageed, "Detection and continual learning of novel face presentation attacks," in *ICCV*, 2021.

[93] J. Galbally, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "A high performance fingerprint liveness detection method based on quality related features," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 311–321, 2012.

[94] I. Chingovska, A. R. Dos Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE TIFS*, 2014.

[95] I. J. S. Biometrics., "Information technology–biometric presentation attack detection–part 3: testing and reporting," 2017.

[96] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *CVPR*, 2020.

[97] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *ECCV*. Springer, 2020.

[98] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, "Dual-cross central difference network for face anti-spoofing," in *IJCAI*, 2021.

[99] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *ECCV*. Springer, 2020.

[100] X. Xu, Y. Xiong, and W. Xia, "On improving temporal consistency for online face liveness detection," in *ICCVW*, 2021.

[101] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. Lai, "Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition," in *CVPR*, 2022.

[102] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, S. Li, and Z. Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *CVPR*, 2022.

[103] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing," *Pattern Recognition*, 2021.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.

[105] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[106] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[107] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2016.

[108] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *TPAMI*, no. 12, pp. 2037–2041, 2006.

[109] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005.

[110] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *ICPR*, 2014.

[111] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *TPAMI*, vol. 33, no. 3, pp. 500–513, 2010.

[112] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," in *ECCV*. Springer, 2020.

[113] R. Cai and C. Chen, "Learning deep forest with multi-scale local binary pattern features for face anti-spoofing," *arXiv preprint arXiv:1910.03850*, 2019.

[114] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *Journal of Visual Communication and Image Representation*, 2016.

[115] Z. Yu, X. Li, P. Wang, and G. Zhao, "Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection," *IEEE SPL*, 2021.

[116] L. Li, Z. Xia, X. Jiang, Y. Ma, F. Roli, and X. Feng, "3d face mask presentation attack detection based on intrinsic image analysis," *IET Biometrics*, 2020.

[117] A. Agarwal, M. Vatsa, and R. Singh, "Chif: Convoluted histogram image features for detecting silicone mask based face presentation attack," in *BTAS*, 2019.

[118] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing," *TIFS*, 2018.

[119] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE TPAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[120] O. Sharifi, "Score-level-based face anti-spoofing system using handcrafted and deep learned characteristics," *International Journal of Image, Graphics and Signal Processing*, 2019.

[121] Y. A. U. Rehman, L.-M. Po, M. Liu, Z. Zou, and W. Ou, "Perturbing convolutional feature maps with histogram of oriented gradients for face liveness detection," in *CISIS and ICEUTE 2019*. Springer, 2019.

[122] L. Li, Z. Xia, A. Hadid, X. Jiang, H. Zhang, and X. Feng, "Replayed video attack detection based on motion blur analysis," *TIFS*, 2019.

[123] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[124] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015.

[125] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, 2014.

[126] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[127] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *ICIAR*, 2017.

[128] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *TIFS*, 2019.

[129] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," *arXiv preprint arXiv:2011.08019*, 2020.

[130] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[131] D. Deb and A. K. Jain, "Look locally infer globally: A generalizable face anti-spoofing approach," *TIFS*, 2020.

[132] Z. Xu, S. Li, and W. Deng, "Learning temporal features using lstm-cnn architecture for face anti-spoofing," in *ACPR*, 2015.

[133] U. Muhammad, T. Holmberg, W. C. de Melo, and A. Hadid, "Face anti-spoofing via sample learning based recurrent neural network (rnn)." in *BMVC*, 2019.

[134] H. Ge, X. Tu, W. Ai, Y. Luo, Z. Ma, and M. Xie, "Face anti-spoofing by the enhancement of temporal motion," in *CTISC*. IEEE, 2020.

[135] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[136] H. Hao, M. Pei, and M. Zhao, "Face liveness detection based on client identity using siamese network," in *PRCV*. Springer, 2019.

[137] B. Chen, W. Yang, H. Li, S. Wang, and S. Kwong, "Camera invariant feature learning for generalized face anti-spoofing," *TIFS*, 2021.

[138] A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Improving cross-dataset performance of face presentation attack detection systems using face recognition datasets," in *ICASSP*, 2020.

[139] D. Peng, J. Xiao, R. Zhu, and G. Gao, "Ts-fen: Probing feature selection strategy for face anti-spoofing," in *ICASS*. IEEE, 2020.

[140] H. Wu, D. Zeng, Y. Hu, H. Shi, and T. Mei, "Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning," *TCSVT*, 2021.

[141] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Learning multi-granularity temporal characteristics for face anti-spoofing," *TIFS*, 2022.

[142] M. S. Hossain, L. Rupty, K. Roy, M. Hasan, S. Sengupta, and N. Mohammed, "A-deeppixbis: Attentional angular margin for face anti-spoofing," 2020.

[143] Z. Yu, Y. Qin, X. Xu, C. Zhao, Z. Wang, Z. Lei, and G. Zhao, "Auto-fas: Searching lightweight networks for face anti-spoofing," in *ICASSP*, 2020.

[144] K. Roy, M. Hasan, L. Rupty, M. Hossain, S. Sengupta, S. N. Taus, N. Mohammed *et al.*, "Bi-fpnfas: Bi-directional feature pyramid network for pixel-wise face anti-spoofing by leveraging fourier spectra," *Sensors*, vol. 21, no. 8, p. 2799, 2021.

[145] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, 2012.

[146] G. Wang, H. Han, S. Shan, and X. Chen, "Improving cross-database face presentation attack detection via adversarial domain adaptation," in *ICB*. IEEE, 2019.

[147] Wang, Guoqing and Han, Hu and Shan, Shiguang and Chen, Xilin, "Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection," *TIFS*, 2020.

[148] R. Quan, Y. Wu, X. Yu, and Y. Yang, "Progressive transfer learning for face anti-spoofing," *TIP*, vol. 30, pp. 3946–3955, 2021.

[149] F. Zhou, C. Gao, F. Chen, C. Li, X. Li, F. Yang, and Y. Zhao, "Face anti-spoofing based on multi-layer domain adaptation," in *ICMEW*. IEEE, 2019.

[150] A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Domain adaptation for generalization of face presentation attack detection in mobile settengs with minimal information," in *ICASSP*, 2020.

[151] H. Li, S. Wang, P. He, and A. Rocha, "Face anti-spoofing with deep neural network distillation," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[152] Z. Chen, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, F. Huang, and X. Jin, "Generalizable representation learning for mixture domain face anti-spoofing," in *AAAI*, 2021.

[153] S. Liu, K.-Y. Zhang, T. Yao, M. Bi, S. Ding, J. Li, F. Huang, and L. Ma, "Adaptive normalized representation learning for generalizable face anti-spoofing," in *ACM MM*, 2021.

[154] J. Wang, J. Zhang, Y. Bian, Y. Cai, C. Wang, and S. Pu, "Self-domain adaptation for face anti-spoofing," in *AAAI*, 2021.

[155] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Learning to learn face-pad: a lifelong learning approach," in *IJCB*. IEEE, 2020.

[156] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *ICB*. IEEE, 2018.

[157] F. Xiong and W. AbdAlmageed, "Unknown presentation attack detection with face rgb images," in *BTAS*, 2018.

[158] Y. Baweja, P. Oza, P. Perera, and V. M. Patel, "Anomaly detection-based unknown face presentation attack detection," *arXiv preprint arXiv:2007.05856*, 2020.

[159] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Deep anomaly detection for generalized face anti-spoofing," in *CVPRW*, 2019.

[160] A. George and S. Marcel, "Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks," *TIFS*, 2020.

[161] Z. Li, H. Li, K.-Y. Lam, and A. C. Kot, "Unseen face presentation attack detection with hypersphere loss," in *ICASSP*, 2020.

[162] Y. A. U. Rehman, L.-M. Po, and M. Liu, "Slnet: Stereo face liveness detection via dynamic disparity-maps and convolutional neural network," *Expert Systems with Applications*, 2020.

[163] W. Hu, G. Te, J. He, D. Chen, and Z. Guo, "Aurora guard: Real-time face anti-spoofing via light reflection," *arXiv preprint arXiv: 1902.10311*, 2019.

[164] B. Wu, M. Pan, and Y. Zhang, "A review of face anti-spoofing and its applications in china," in *International Conference on Harmony Search Algorithm*. Springer, 2019, pp. 35–43.

[165] J. Connell, N. Ratha, J. Gentile, and R. Bolle, "Fake iris detection using structured light," in *ICASSP*, 2013.

[166] X. Sun, L. Huang, and C. Liu, "Context based face spoofing detection using active near-infrared images," in *ICPR*, 2016.

[167] J. Seo and I.-J. Chung, "Face liveness detection using thermal face-cnn with external knowledge," *Symmetry*, 2019.

[168] M. Kang, J. Choe, H. Ha, H.-G. Jeon, S. Im, and I. S. Kweon, "Facial depth and normal estimation using single dual-pixel camera," *arXiv preprint arXiv:2111.12928*, 2021.

[169] X. Wu, J. Zhou, J. Liu, F. Ni, and H. Fan, "Single-shot face anti-spoofing for dual pixel camera," *TIFS*, 2020.

[170] H. Farrukh, R. M. Aburas, S. Cao, and H. Wang, "Facerevelio: a face liveness detection system for smartphones with a single front camera," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[171] A. F. Ebihara, K. Sakurai, and H. Imaoka, "Specular-and diffuse-reflection-based face spoofing detection for mobile devices," in *IJCB*. IEEE, 2020.

[172] A. Parkin and O. Grinchuk, "Recognizing multi-modal face spoofing with face recognition networks," in *CVPRW*, 2019.

[173] H. Kuang, R. Ji, H. Liu, S. Zhang, X. Sun, F. Huang, and B. Zhang, "Multi-modal multi-layer fusion network with average binary center loss for face anti-spoofing," in *ACM MM*, 2019.

[174] T. Shen, Y. Huang, and Z. Tong, "Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing," in *CVPRW*, 2019.

[175] A. George and S. Marcel, "Cross modal focal loss for rgbd face anti-spoofing," in *CVPR*, 2021.

[176] O. Nikisins, A. George, and S. Marcel, "Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing," in *ICB*. IEEE, 2019.

[177] W. Liu, X. Wei, T. Lei, X. Wang, H. Meng, and A. K. Nandi, "Data fusion based two-stage cascade framework for multi-modality face anti-spoofing," *TCDS*, 2021.

[178] P. Zhang, F. Zou, Z. Wu, N. Dai, S. Mark, M. Fu, J. Zhao, and K. Li, "Feathernets: Convolutional neural networks as light as feather for face anti-spoofing," in *CVPRW*, 2019.

[179] F. Jiang, P. Liu, X. Shao, and X. Zhou, "Face anti-spoofing with generated near-infrared images," *Multimedia Tools and Applications*, vol. 79, no. 29, pp. 21 299–21 323, 2020.

[180] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *TIFS*, 2021.

[181] K. Mallat and J.-L. Dugelay, "Indirect synthetic attack on thermal face biometric systems via visible-to-thermal spectrum conversion," in *CVPRW*, 2021.

[182] S. Liu, S. Lu, H. Xu, J. Yang, S. Ding, and L. Ma, "Feature generation and hypothesis verification for reliable face anti-spoofing," in *AAAI*, 2022.

[183] A. F. Sequeira, T. Gonçalves, W. Silva, J. R. Pinto, and J. S. Cardoso, "An exploratory study of interpretability for face presentation attack detection," *IET Biometrics*, 2021.

[184] 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 *ICCV*, 2017.

[185] H. Mirzaalian, M. E. Hussein, L. Spinoulas, J. May, and W. Abd-Almageed, "Explaining face presentation attack detection using natural language," in *FG*. IEEE, 2021.

[186] A. Liu, X. Li, J. Wan, Y. Liang, S. Escalera, H. J. Escalante, M. Madadi, Y. Jin, Z. Wu, X. Yu *et al.*, "Cross-ethnicity face anti-spoofing recognition challenge: A review," *IET Biometrics*, vol. 10, no. 1, pp. 24–43, 2021.

[187] Z. Yu, C. Zhao, K. H. Cheng, X. Cheng, and G. Zhao, "Flexible-modal face anti-spoofing: A benchmark," *arXiv preprint arXiv:2202.08192*, 2022.

[188] J. Stehouwer, A. Jourabloo, Y. Liu, and X. Liu, "Noise modeling, synthesis and classification for generic object anti-spoofing," in *CVPR*, 2020.

[189] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Transactions on Privacy and Security (TOPS)*, 2019.

[190] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *ICPR*, 2021.

[191] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," in *IJCAI*, 2021.

[192] Y. Guo, X. Wei, G. Wang, and B. Zhang, "Meaningful adversarial stickers for face recognition in physical world," *arXiv preprint arXiv:2104.06728*, 2021.

[193] B. Zhang, B. Tondi, and M. Barni, "Attacking cnn-based anti-spoofing face authentication in the physical domain," *arXiv preprint arXiv:1910.00327*, 2019.

[194] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *TPAMI*, 2020.

[195] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Are gan-based morphs threatening face recognition?" in *ICASSP*, 2022.

[196] D. Deb, X. Liu, and A. K. Jain, "Unified detection of digital and physical face attacks," *arXiv preprint arXiv:2104.02156*, 2021.

[197] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[198] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017.

[199] R. Shao, B. Zhang, P. C. Yuen, and V. M. Patel, "Federated test-time adaptive face presentation attack detection with dual-phase privacy preservation," in *FG*. IEEE, 2021.

[200] J. N. Kundu, N. Venkat, R. V. Babu *et al.*, "Universal source-free domain adaptation," in *CVPR*, 2020.

[201] L. Lv, Y. Xiang, X. Li, H. Huang, R. Ruan, X. Xu, and Y. Fu, "Combining dynamic image and prediction ensemble for cross-domain face anti-spoofing," in *ICASSP*, 2021.

[202] O. Kähm and N. Damer, "2d face liveness detection: An overview," in *BIOSIG*. IEEE, 2012.

[203] A. Hadid, "Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions," in *CVPRW*, 2014.

[204] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, 2014.

[205] S. Kumar, S. Singh, and J. Kumar, "A comparative study on face spoofing attacks," in *ICCCA*. IEEE, 2017.

[206] D. R. Kisku and R. D. Rakshit, "Face spoofing and counter-spoofing: A survey of state-of-the-art algorithms," *Transactions on Machine Learning and Artificial Intelligence*, vol. 5, no. 2, pp. 31–31, 2017.

[207] L. Souza, L. Oliveira, M. Pamplona, and J. Papa, "How far did we get in face spoofing detection?" *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 368–381, 2018.

[208] E. A. Raheem, S. M. S. Ahmad, and W. A. W. Adnan, "Insight on face liveness detection: A systematic literature review," *International Journal of Electrical and Computer Engineering*, 2019.

[209] Z. Ming, M. Visani, M. M. Luqman, and J.-C. Burie, "A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices," *Journal of Imaging*, 2020.

[210] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing," *TIFS*, vol. 16, pp. 937–951, 2020.

[211] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE TBIOM*, 2021.

[212] T. Kim and Y. Kim, "Suppressing spoof-irrelevant factors for domain-agnostic face anti-spoofing," *arXiv preprint arXiv:2012.01271*, 2020.

[213] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[214] L. Li and X. Feng, "Face anti-spoofing via deep local binary pattern," in *Deep Learning in Object Detection and Recognition*. Springer, 2019, pp. 91–111.

[215] H. Chen, Y. Chen, X. Tian, and R. Jiang, "A cascade face spoofing detector based on face anti-spoofing r-cnn and improved retinex lbp," *IEEE Access*, 2019.

[216] P. K. Das, B. Hu, C. Liu, K. Cui, P. Ranjan, and G. Xiong, "A new approach for face anti-spoofing using handcrafted and deep network features," in *SOLI*. IEEE, 2019.

[217] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *TIFS*, 2015.

[218] L. Li, Z. Xia, L. Li, X. Jiang, X. Feng, and F. Roli, "Face anti-spoofing via hybrid convolutional neural network," in *FADS*. IEEE, 2017.

[219] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *IJCNN*. IEEE, 2019.

[220] X. Tu and Y. Fang, "Ultra-deep neural network for face anti-spoofing," in *International Conference on Neural Information Processing*. Springer, 2017.

[221] Y. A. U. Rehman, L. M. Po, and M. Liu, "Deep learning for face anti-spoofing: An end-to-end approach," in *SPA*. IEEE, 2017, pp. 195–200.

[222] C. Lin, Z. Liao, P. Zhou, J. Hu, and B. Ni, "Live face verification with multiple instantialized local homographic parameterization." in *IJCAI*, 2018.

[223] G. B. de Souza, J. P. Papa, and A. N. Marana, "On the learning of deep local features for robust face spoofing detection," in *SIBGRAPI*. IEEE, 2018.

[224] Y. A. U. Rehman, L. M. Po, and M. Liu, "Livenet: Improving features generalization for face liveness detection using convolution neural networks," *Expert Systems with Applications*, vol. 108, pp. 159–169, 2018.

[225] S. Luo, M. Kan, S. Wu, X. Chen, and S. Shan, "Face anti-spoofing with multi-scale information," in *ICPR*, 2018.

[226] K. Larbi, W. Ouarda, H. Drira, B. B. Amor, and C. B. Amar, "Deep-colorfasd: Face anti spoofing solution using a multi channeled color spaces cnn," in *SMC*. IEEE, 2018.

[227] R. Bresan, A. Pinto, A. Rocha, C. Beluzo, and T. Carvalho, "Facespoof buster: a presentation attack detector based on intrinsic image properties and deep learning," *arXiv preprint arXiv:1902.02845*, 2019.

[228] Y. A. U. Rehman, L.-M. Po, M. Liu, Z. Zou, W. Ou, and Y. Zhao, "Face liveness detection using convolutional-features fusion of real and deep network generated face images," *Journal of Visual Communication and Image Representation*, 2019.

[229] R. Laurensi, A. Israel, L. T. Menon, N. Penna, O. Manoel Camillo, A. L. Koerich, and A. S. Britto Jr, "Style transfer applied to face liveness detection with user-centered models," *arXiv*, pp. arXiv–1907, 2019.

[230] X. Tu, Z. Ma, J. Zhao, G. Du, M. Xie, and J. Feng, "Learning generalizable and identity-discriminative representations for face anti-spoofing," *ACM TIST*, vol. 11, no. 5, pp. 1–19, 2020.

[231] J. Guo, X. Zhu, J. Xiao, Z. Lei, G. Wan, and S. Z. Li, "Improving face anti-spoofing by 3d virtual synthesis," in *ICB*, 2019, pp. 1–8.

[232] A. Pinto, S. Goldenstein, A. Ferreira, T. Carvalho, H. Pedrini, and A. Rocha, "Leveraging shape, reflectance and albedo from shading for face presentation attack detection," *TIFS*, 2020.

[233] Y. Zuo, W. Gao, and J. Wang, "Face liveness detection algorithm based on livenesslight network," in *HPBD&IS*. IEEE, 2020.

[234] B. Chen, W. Yang, and S. Wang, "Face anti-spoofing by fusing high and low frequency features for advanced generalization capability," in *MIPR*. IEEE, 2020, pp. 199–204.

[235] A. Parkin and O. Grinchuk, "Creating artificial modalities to solve rgb liveness," *arXiv preprint arXiv:2006.16028*, 2020.

[236] Y. Huang, W. Zhang, and J. Wang, "Deep frequent spatial temporal learning for face anti-spoofing," *arXiv preprint arXiv:2002.03723*, 2020.

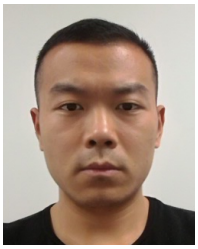
[237] Y. Ma, L. Wu, Z. Li *et al.*, "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," *PR Letters*, vol. 131, pp. 261–267, 2020.

[238] W. Sun, Y. Song, H. Zhao, and Z. Jin, "A face spoofing detection method based on domain adaptation and lossless size adaptation," *IEEE Access*, 2020.

[239] W. Zheng, M. Yue, S. Zhao, and S. Liu, "Attention-based spatial-temporal multi-scale network for face anti-spoofing," *TBIOM*, 2021.

[240] Y. Chen, T. Wang, J. Wang, P. Shi, and H. Snoussi, "Towards good practices in face anti-spoofing: An image reconstruction based method," in *CAC*. IEEE, 2019.
[241] X. Tu, H. Zhang, M. Xie, Y. Luo, Y. Zhang, and Z. Ma, "Deep transfer across domains for face antispoofing," *Journal of Electronic Imaging*, 2019.
[242] S. Saha, W. Xu, M. Kanakis, S. Georgoulis, Y. Chen, D. Pani Paudel, and L. Van Gool, "Domain agnostic feature learning for image and video based face anti-spoofing," in *CVPRW*, 2020.
[243] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Spoofing attack detection by anomaly detection," in *ICASSP*, 2019.
[244] S. Fatemifar, M. Awais, A. Akbari, and J. Kittler, "A stacking ensemble for anomaly based client-specific face spoofing detection," in *ICIP*. IEEE, 2020.
[245] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Client-specific anomaly detection for face presentation attack detection," *Pattern Recognition*, 2020.
[246] M. Kowalski, "A study on presentation attack detection in thermal infrared," *Sensors*, vol. 20, no. 14, p. 3988, 2020.
[247] G. Wang, C. Lan, H. Han, S. Shan, and X. Chen, "Multi-modal face presentation attack detection via spatial and channel attentions," in *CVPRW*, 2019.
[248] L. Li, Z. Gao, L. Huang, H. Zhang, and M. Lin, "A dual-modal face anti-spoofing method via light-weight networks," in *ASID*. IEEE, 2019.
[249] X. Li, W. Wu, T. Li, Y. Su, and L. Yang, "Face liveness detection based on parallel cnn," in *Journal of Physics: Conference Series*. IOP Publishing, 2020.
[250] A. George and S. Marcel, "Can your face detector do anti-spoofing? face presentation attack detection with a multi-channel face detector," *arXiv preprint arXiv:2006.16836*, 2020.
[251] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, "Pipenet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing," in *CVPRW*, 2020.
[252] G. Te, W. Hu, and Z. Guo, "Exploring hypergraph representation on face anti-spoofing beyond 2d attacks," in *ICME*. IEEE, 2020.

Zitong Yu received the M.S. degree in multimedia from University of Nantes, France, in 2016, and he received the Ph.D. degree in computer science from University of Oulu, Finland, in 2022. His research interests include face anti-spoofing, remote physiological measurement and video understanding. He led the team and won the 1st Place in the ChaLearn multi-modal face anti-spoofing attack detection challenge with CVPR 2020.



Yunxiao Qin received the M.S. degree in control theory and control engineering from the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China, in 2015, and he received the Ph.D. degree in control science and engineering from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2021. His current research interests include meta-learning, face anti-spoofing, and deep reinforcement learning.

Xiaobai Li received her B.Sc. degree in Psychology from Peking University, M.Sc. degree in Biophysics from the Chinese Academy of Science, and Ph.D. degree in Computer Science from University of Oulu. She is currently an assistant professor in the Center for Machine Vision and Signal Analysis of University of Oulu. Her research of interests includes facial expression recognition, micro-expression analysis, remote physiological signal measurement from facial videos, and related applications in affective computing, healthcare and biometrics. She is an associate editor for IEEE TCSVT, Frontiers in Psychology, and Image and Vision Computing. Dr. Li was a co-chair of several international workshops in CVPR, ICCV, FG and ACM Multimedia.

Chenxu Zhao received M.S. Degree from Beihang University, Beijing, China, in 2016, and was in the joint programme with National Laboratory of Pattern Recognition (NLPR) Laboratory of Institute of Automation, Chinese Academy of Sciences, from 2014 to 2016. He is currently served as a Co-Founder in SailYond Technology, Beijing, China. He served as a Research Director in MiningLamp Technology, Beijing, China. His major research areas include face analysis, anomaly detection and meta-learning.

Zhen Lei received the B.S. degree in automation from the University of Science and Technology of China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently a Professor. He is IAPR Fellow and AAIA Fellow. He has published over 200 papers in international journals and conferences with 21000+ citations in Google Scholar and h-index 71. He was competition co-chair of IJCB2022 and has served as area chairs for server conferences and is associate editor for IEEE Trans. on Information Forensics and Security, Pattern Recognition, Neurocomputing and IET Computer Vision journals. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He is the winner of 2019 IAPR Young Biometrics Investigator Award.

Guoying Zhao (IEEE Fellow 2022) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently an Academy Professor and full Professor (tenured in 2017) with University of Oulu. She is also a visiting professor with Aalto University. She is a member of Finnish Academy of Sciences and Letters, IAPR Fellow and AAIA Fellow. She has authored or co-authored more than 280 papers in journals and conferences with 20100+ citations in Google Scholar and h-index 66. She is panel chair for FG 2023, was co-program chair for ACM International Conference on Multimodal Interaction (ICMI 2021), co-publicity chair for FG2018, and has served as area chairs for several conferences and was/is associate editor for IEEE Trans. on Multimedia, Pattern Recognition, IEEE Trans. on Circuits and Systems for Video Technology, and Image and Vision Computing Journals. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics. Her research has been reported by Finnish TV programs, newspapers and MIT Technology Review.

TABLE 3: A summary of existing surveys in FAS. Most of them focus on handcrafted feature based methods under single RGB modality, and investigate limited number of public datasets as well as evaluation protocols. 'DL' and 'M&H' is short for 'deep learning' and 'modality & hardware', respectively. 'VIS', 'NIR', 'SWIR','LF', and 'Polarized' denotes using commercial visible RGB, near infrared, short-wave infrared, light field, and four-directional polarized camera, respectively.

Title & Reference	Year	DL	M&H	Datasets	Protocol
2D Face Liveness Detection: an Overview [202]	2014	No	VIS	2	Intra-dataset intra-type
Face Biometrics under Spoofing Attacks: Vulnerabilities, Countermeasures, Open Issues and Research Directions [203]	2014	No	VIS	4	Intra-dataset intra-type
Biometric Anti-spoofing Methods: A Survey in Face Recognition [204]	2015	No	VIS	6	Intra-dataset intra-type
A Comparative Study on Face Spoofing Attacks [205]	2017	No	VIS	9	Intra-dataset intra-type
Face Spoofing and Counter-Spoofing: A Survey of State-of-the-art Algorithms [206]	2017	No	VIS	6	Intra-dataset intra-type, Cross-dataset intra-type
Presentation Attack Detection Methods for Face Recognition Systems: A Comprehensive Survey [67]	2017	No	VIS	11	Intra-dataset intra-type
How Far Did We Get in Face Spoofing Detection? [207]	2018	Yes (Few, <10)	VIS	9	Intra-dataset intra-type, Cross-dataset intra-type
Insight on Face Liveness Detection: A Systematic Literature Review [208]	2019	No	VIS	14	Intra-dataset intra-type
The Rise of Data-driven Models in Presentation Attack Detection [56]	2019	Yes (Few, <10)	VIS	7	Intra-dataset intra-type, Cross-dataset intra-type
A Survey on 3D Mask Presentation Attack Detection and Countermeasures [57]	2020	Yes (Few, <10)	VIS	10	Intra-dataset intra-type
Deep Convolutional Neural Networks for Face and Iris Presentation Attack Detection: Survey and Case Study [58]	2020	Yes (Few, <30)	VIS	8	Intra-dataset intra-type, Cross-dataset intra-type
A Survey On Anti-Spoofing Methods For Face Recognition with RGB Cameras of Generic Consumer Devices [209]	2020	Yes (Few, <50)	VIS	12	Intra-dataset intra-type, Cross-dataset intra-type
Deep Learning for Face Anti-spoofing: A Survey (Ours)	2022	Yes (Full, >100)	VIS, Flash, NIR, Thermal, Depth, SWIR, LF, Polarized	36	Intra-dataset intra-type, Cross-dataset intra-type, Intra-dataset cross-type, Cross-dataset cross-type

TABLE 4: ACER (%) results of the intra-dataset intra-type testings on OULU-NPU (4 sub-protocols) and SiW (3 sub-protocols) datasets for common deep learning methods with binary cross-entropy supervision and pixel-wise supervision.

	Method	Venue	OULU-NPU				SiW		
			Prot. 1	Prot. 2	Prot. 3	Prot. 4	Prot. 1	Prot. 2	Prot. 3
Binary Cross-entropy Supervision	STASN [25]	CVPR'19	1.9	2.2	2.8±1.6	7.5±4.7	1.00	0.28±0.05	12.10± 1.50
	TSCNN [128]	TIFS'19	5.9	4.9	5.6±1.6	9.8±4.2	-	-	-
	CIFL [137]	TIFS'21	3.4	2.4	2.5±0.8	6.1±4.1	-	-	-
	DRL-FAS [210]	TIFS'20	4.7	1.9	3.0±1.5	7.2±3.9	**0.00**	**0.00±0.00**	4.51± 0.00
	SSR-FCN [131]	TIFS'20	4.6	3.4	2.8 ±2.2	10:8± 5:1	-	-	-
	FasTCo [100]	ICCVW'21	0.8	**1.1**	**1.1±0.8**	**1.5±1.2**	0.0003	0.01±0.01	2.00±0.56
	PatchNet [101]	CVPR'22	**0.0**	1.2	1.18±1.26	2.9±3.0	**0.00**	**0.00±0.00**	2.45± 0.45
Pixel-wise Supervision	Auxiliary [13]	CVPR'18	1.6	2.7	2.9±1.5	9.5±6.0	3.58	0.57±0.69	8.31±3.81
	PixBiS [32]	IJCB'19	0.4	6.0	11.1±9.4	25.0±12.7	-	-	-
	FAS-SGTD [96]	CVPR'20	1.0	1.9	2.7±0.6	5.0±2.2	0.40	0.02± 0.04	2.78± 3.57
	De-Spoof [33]	ECCV'20	1.5	4.3	3.6±1.6	5.6±5.7	-	-	-
	Disentangled [97]	ECCV'20	1.3	2.4	2.2± 2.2	4.4± 3.0	0.28	0.10± 0.04	5.59± 4.37
	STDN [99]	ECCV'20	1.1	1.9	2.8± 3.3	3.8± 4.2	**0.00**	**0.00± 0.00**	7.9± 3.3
	BCN [24]	ECCV'20	0.8	1.7	2.5± 1.1	5.2± 3.7	0.36	0.11±0.08	2.45± 0.68
	CDCN [23]	CVPR'20	1.0	1.5	2.3± 1.4	6.9± 2.9	0.12	0.06± 0.04	1.71± 0.11
	DC-CDN [98]	IJCAI'21	0.4	1.3	1.9± 1.1	4.0± 3.1	-	-	-
	NAS-FAS [37]	PAMI'21	0.2	1.2	1.7±0.6	2.9±2.8	0.12	0.04±0.05	**1.52±0.13**
	MT-FAS [43]	PAMI'21	0.4	1.4	2.1± 1.7	3.7± 2.9	-	-	-

TABLE 5: HTER (%) results of the cross-dataset intra-type testings among OULU-NPU (O), CASIA-MFSD (C), Replay-Attack (I), and MSU-MFSD (M) datasets with different numbers of source domains for training. For example, 'C to I' means training on CASIA-MFSD and then testing on Replay-Attack.

	Method	Venue	1 source domain		2 source domains		3 source domains			
			C to I	I to C	M&I to C	M&I to O	O&C&I to M	O&M&I to C	O&C&M to I	I&C&M to O
Traditional Deep Learning	Auxiliary [13]	CVPR'18	27.6	28.4	-	-	-	28.4	27.6	-
	CDCN [23]	CVPR'20	15.5	32.6	-	-	-	-	-	-
	FAS-SGTD [96]	CVPR'20	17.0	**22.8**	-	-	-	-	-	-
	BCN [24]	ECCV'20	16.6	36.4	-	-	19.81	25.12	22.75	21.24
	NAS-FAS [37]	PAMI'21	-	-	-	-	16.85	15.21	11.63	13.16
	MT-FAS [43]	PAMI'21	-	-	-	-	11.67	18.44	11.93	16.23
	PS [211]	TBIOM'21	13.8	31.3	-	-	20.42	18.25	19.55	15.76
	DC-CDN [98]	IJCAI'21	**6.0**	30.1	-	-	25.31	15.00	15.88	18.82
Generalized Deep Learning	MADDG [48]	CVPR'19	-	-	41.02	39.35	17.69	24.50	22.19	27.98
	PAD-GAN [50]	CVPR'20	-	-	31.67	34.02	17.02	19.68	20.87	25.02
	RF-Meta [49]	AAAI'20	-	-	-	-	13.89	20.27	17.30	16.45
	SSDG [51]	CVPR'20	-	-	31.89	36.01	7.38	10.44	11.71	15.61
	SDA [154]	AAAI'21	-	-	-	-	15.40	24.50	15.60	23.10
	D^2AM [152]	AAAI'21	-	-	-	-	12.70	20.98	15.43	15.27
	DASN [212]	Access'21	-	-	-	-	8.33	12.04	13.38	**11.77**
	DRDG [**?**]	IJCAI'21	-	-	31.28	33.35	12.43	19.05	15.56	15.63
	ANRL [153]	MM'21	-	-	31.06	30.73	10.83	17.83	16.03	15.67
	FGHV [182]	AAAI'22	-	-	-	-	9.17	12.47	16.29	13.58
	SSAN [102]	CVPR'22	-	-	**30.00**	**29.44**	**6.67**	**10.00**	**8.88**	13.72

TABLE 6: EER (%) results of the Intra-dataset cross-type testings on SiW-M with the leave-one-type-out setting.

	Method	Venue	Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			Average
					Half	Silicone	Trans.	Paper	Manne.	Obfusc.	Im.	Cos.	Fun.	Glasses	Partial	
Traditional Deep Learning	Auxiliary [13]	CVPR'18	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	**9.4**	21.4	18.6	4.0	17.0±17.7
	CDCN [23]	CVPR'20	8.2	7.8	8.3	**7.4**	20.5	5.9	5.0	47.8	1.6	14.0	24.5	18.3	1.1	13.1± 12.6
	STDN [99]	ECCV'20	7.6	3.8	8.4	13.8	14.5	**5.3**	4.4	35.4	**0.0**	19.3	21.0	20.8	1.6	12.0± 10.0
	BCN [24]	ECCV'20	13.4	5.2	8.3	9.7	13.6	5.8	2.5	33.8	**0.0**	14.0	23.3	16.6	1.2	11.3± 9.5
	SSR-FCN [131]	TIFS'20	**6.8**	11.2	**2.8**	6.3	28.5	0.4	3.3	17.8	3.9	11.7	21.6	13.5	3.6	10.1± 8.4
	PS [211]	TBIOM'21	10.3	7.8	8.3	7.4	10.2	5.9	5.0	43.4	**0.0**	12.0	23.9	15.9	**0.0**	11.5±11.4
	NAS-FAS [211]	PAMI'21	10.3	7.8	8.3	7.4	10.2	5.9	5.0	43.4	**0.0**	12.0	23.9	15.9	**0.0**	11.5±11.4
	DC-CDN [98]	IJCAI'21	10.3	8.7	11.1	7.4	12.5	5.9	**0.0**	39.1	**0.0**	12.0	18.9	13.5	1.2	10.8±10.1
	MT-FAS [43]	PAMI'21	7.8	4.4	11.2	5.8	11.2	2.8	2.7	38.9	0.2	10.1	20.5	18.9	1.3	10.4±10.2
	ViTranZFAS [129]	IJCB'21	15.2	5.8	5.8	**4.9**	**5.9**	**0.1**	3.2	**9.8**	0.4	10.7	20.1	**2.9**	1.9	**6.7±5.6**
Generalized Learning	DTN [38]	CVPR'19	10.0	**2.1**	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	16.1± 12.2
	Hypersphere [161]	ICASSP'20	13.2	14.0	18.1	24.0	12.4	3.1	6.2	34.8	3.1	16.3	21.4	21.7	9.3	15.2± 9.0
	FGHV [102]	AAAI'22	9.0	8.0	5.9	9.9	14.3	3.7	4.8	19.3	2.0	9.2	18.9	8.5	4.7	9.1± 5.4

TABLE 7: Summary of the **hybrid (handcraft+deep learning)** FAS methods with **binary cross-entropy supervision**. 'S/D', 'CE', 'OF', 'OFM', 'NN', 'HOG', 'LBP' are short for 'Static/Dynamic', 'cross-entropy', 'optical flow', 'optical flow magnitude', 'nearest neighbor', 'histogram of oriented gradients [109]' and 'local binary pattern [213]', respectively.

Method	Year	Backbone	Loss	Input	S/D	Description
DPCNN [30]	2016	VGG-Face	Trained with SVM	RGB	S	deep partial features with Blocks PCA
Multi-cues+NN [114]	2016	MLP	Binary CE loss	RGB+OFM	D	fused features from image quality cues and motion cues
CNN LBP-TOP [20]	2017	5-layer CNN	Binary CE loss SVM	RGB	D	cascading LBP-TOP with CNN to extract discriminative spatio-temporal features
DF-MSLBP [113]	2018	Deep forest	Binary CE loss	HSV+YCbCr	S	multi-scale LBP based Tree-Ensembled features
SPMT+SSD [19]	2018	VGG16	Binary CE loss SVM bbox regression	RGB Landmarks	S	hand-crafted texture&depth features cascaded with fast deep face spoofing detector
CHIF [117]	2019	VGG-Face	Trained with SVM	RGB	S	convoluted histogram image features for fine-grained mask texture representation
DeepLBP [214]	2019	VGG-Face	Binary CE loss SVM	RGB,HSV, YCbCr	S	extracted the handcrafted features from the convolutional responses of the fine-tuned CNN model
CNN+LBP +WLD [22]	2019	CaffeNet	Binary CE loss	RGB	S	combined CNN features with LBP/WLD for preserving both semantic feature and local information
Intrinsic [116]	2019	1D-CNN	Trained with SVM	Reflection	D	deep temporal cues from reflection intensity histogram
FARCNN [215]	2019	Multi-scale attentional CNN	Regression loss Crystal loss Center loss	RGB	S	cascade detector features with improved Retinex based LBP
CNN-LSP [122]	2019	1D-CNN	Trained with SVM	RGB	D	joint learned temporal features with attentional spatial regions and channels from magnified videos
DT-Mask [118]	2019	VGG16	Binary CE loss Channel&Spatial-discriminability	RGB+OF	D	joint learned discriminative features with attentional spatial regions and channels
VGG+LBP [216]	2019	VGG16	Binary CE loss	RGB	S	combining deep CNN features, and LBP features from brightness and chrominance channels
CNN+OVLBP [120]	2019	VGG16	Binary CE loss NN classifier	RGB	S	hybrid decisions using majority vote of CNN, over-lapped histograms of LBP and their fused vector
HOG-Pert. [121]	2019	Multi-scale CNN	Binary CE loss	RGB+HOG	S	hybrid convolutional features and HOG features
LBP-Pert. [21]	2020	Multi-scale CNN	Binary CE loss	RGB+LBP	S	discriminative features enhanced by LBP perturbation
TransRPPG [115]	2021	Vision Transformer	Binary CE loss	rPPG map	D	intrinsic liveness features via fully attentional transformer

TABLE 8: Summary of the representative **traditional** deep learning based FAS methods with **binary cross-entropy supervision**. 'S/D' and 'CE' are short for 'Static/Dynamic' and 'cross-entropy', respectively. 'Reflect.', 'OF', and 'RP' denote the generated reflection map, optical flow, and rank pooling, respectively.

Method	Year	Backbone	Loss	Input	S/D	Description
CNN1 [29]	2014	8-layer CNN	Trained with SVM	RGB	S	deep features from different spatial scales
LSTM-CNN [132]	2015	CNN+LSTM	Binary CE loss	RGB	D	long-range local and dense features from sequence
SpoofNet [217]	2015	2-layer CNN	Binary CE loss	RGB	S	deep representation with architecture optimization
HybridCNN [218]	2017	VGG-Face	Trained with SVM	RGB	S	hybrid CNN for both global face and facial patches
CNN2 [219]	2017	VGG11	Binary CE loss	RGB	S	model trained with continuous data-randomization
Ultra-Deep [220]	2017	ResNet50+LSTM	Binary CE loss	RGB	D	ultra-deep features with rich long-range temporal context
FASNet [127]	2017	VGG16	Binary CE loss	RGB	S	transfer learned features based on a pre-trained CNN
CNN3 [221]	2018	Inception, ResNet	Binary CE loss	RGB	S	transferred deep feature
MILHP [222]	2018	ResNet+STN	Multiple Instances CE loss	RGB	D	underlying subtle motion features
LSCNN [223]	2018	9 PatchNets	Binary CE loss	RGB	S	global feature via aggregating 9 deep local features
LiveNet [224]	2018	VGG11	Binary CE loss	RGB	S	model trained with continuous data-randomization
MS-FANS [225]	2018	AlexNet+LSTM	Binary CE loss	RGB	S	multi-scale deep feature with rich spatial context
DeepColorFAS [226]	2018	5-layer CNN	Binary CE loss	RGB, HSV, YCbCr	S	investigates the effect of multi-channel space colors on CNN architectures and proposes a fusion based voting method for FAS
Siamese [136]	2019	AlexNet	Contrastive loss	RGB	S	deep features guided by client identity information
FSBuster [227]	2019	ResNet50	Trained with SVM	RGB	S	fused deep features from Intrinsic Image Properties
FuseDNG [228]	2019	7-layer CNN	Binary CE loss Reconstruction loss	RGB	S	adaptive fusion of deep features learned from real-world face and deep autoencoder generated face
STASN [25]	2019	ResNet50+LSTM	Binary CE loss	RGB	D	deep spatio-temporal feature from local salient regions
TSCNN [128]	2019	ResNet18	Binary CE loss	RGB MSR	S	attentional illumination-invariant features with discriminative high-frequency information
FAS-UCM [229]	2019	MobileNetV2 VGG19	Binary CE loss Style loss	RGB	S	deep features trained from generated style transferred images
SLRNN [133]	2019	ResNet50+LSTM	Binary CE loss	RGB	D	augmented temporal features via sparse filtering
GFA-CNN [230]	2019	VGG16	Binary CE loss	RGB	S	generalizable features via multitask and metric learning
3DSynthesis [231]	2019	ResNet15	Binary CE loss	RGB	S	trained on synthesized virtual data of print attacks
CompactNet [35]	2020	VGG19	Points-to-Center triplet loss	RGB	S	deep features on the learned color-liked compact space
SSR-FCN [131]	2020	FCN with 6 layers	Binary CE loss	RGB	S	local discriminative features from Self-Regional Supervision
DRL-FAS [210]	2020	ResNet18+GRU	Binary CE loss	RGB	S	fused local(sub-patches) & global(entire face) features
SfSNet [232]	2020	6-layer CNN	Binary CE loss	Albedo, Depth, Reflect.	S	intrinsic features from shape-from-shading generated pseduo albedo, depth, and reflectance maps
LivenesSlight [233]	2020	6-layer CNN	Binary CE loss	RGB	S	lightweight model and takes less training time
Motion-Enhancement [134]	2020	VGGface+LSTM	Binary CE loss	RGB	D	deep temporal dynamics features with eulerian motion magnification and temporal attention mechanism
CFSA-FAS [234]	2020	ResNet18	Binary CE loss	RGB	S	fuse high and low frequency information with cross-frequency spatial and self-channel attention modules
MC-FBC [34]	2020	VGG16 ResNet50	Binary CE loss	RGB	S	fine-grained features via factorizing bilinear coding of multiple color channels
SimpleNet [235]	2020	Multi-stream 5-layer CNN	Binary CE loss	RGB, OF, RP	D	using intermediate representations from RankPooling and optical flow to increase model's robustness
PatchCNN [80]	2020	SqueezeNet v1.1	Binary CE loss Triplet loss	RGB	S	trained with multi-resolution patches and a multi-objective loss function
FreqSpatial-TempNet [236]	2020	ResNet18	Binary CE loss	RGB, HSV, Spectral	D	discriminative fused features of frequent, spatial and temporal information
ViTranZFAS [129]	2021	Vision Transformer	Binary CE loss	RGB	S	transfer learning from the pre-trained vision transformer model
CIFL [137]	2021	ResNet18	Binary focal loss camear type loss	RGB	S	camera-invariant spoofing features in the high-frequency domain and enhanced image
FasTCo [100]	2021	ResNet50 MobileNetV2	Multi-class CE loss Temporal Consistency loss Class Consistency loss	RGB	D	temporal consistent features as well as temporal smooothed predictions
PatchNet [101]	2022	ResNet18	Asymmetric AM-Softmax loss self-supervised similarity loss	RGB patch	S	fine-grained patch-type live/spoof recognition with strong patch embedding space regularization

TABLE 9: Summary of the representative **traditional deep learning** based FAS methods with **pixel-wise supervision**. Most methods (in the upper part) are supervised with *auxiliary* tasks while the methods in the last eight rows are based on the *generative* models. 'S/D' is short for Static/Dynamic. 'NAS' denotes neural searched architecture. 'TSM' and 'FPN' denote temporal shift module and feature pyramid network, respectively. 'Info-VAE' means information maximizing variational autoencoder. Note that some methods also consider classification loss (e.g., binary cross entropy loss, triplet loss, and adversarial loss), which are not listed in the 'Supervision' column.

Method	Year	Supervision	Backbone	Input	S/D	Description
Depth&Patch [26]	2017	Depth	PatchNet DepthNet	YCbCr HSV	S	local patch features and holistic depth maps extracted by two-stream CNNs
Auxiliary [13]	2018	Depth rPPG spectrum	DepthNet	RGB HSV	D	local temporal features learned from CNN-RNN model with pixel-wise depth and sequence-wise rPPG supervision
BASN [36]	2019	Depth Reflection	DepthNet Enrichment	RGB HSV	S	generalizable features via bipartite auxiliary supervision
DTN [38]	2019	BinaryMask	Tree Network	RGB HSV	S	partition the spoof samples into semantic sub-groups in an unsupervised fashion
PixBiS [32]	2019	BinaryMask	DenseNet161	RGB	S	deep pixel-wise binary supervision without trivial depth synthesis
A-PixBiS [142]	2020	BinaryMask	DenseNet161	RGB	S	incorporate a variant of binary cross entropy that enforces a margin in angular space for attentive pixel wise supervision
Auto-FAS [143]	2020	BinaryMask	NAS	RGB	S	well-suitable lightweight networks searched for mobile-level FAS
MRCNN [237]	2020	BinaryMask	Shallow CNN	RGB	S	introducing local losses to patches, and constraints the entire face region to avoid over-emphasizing certain local areas
FCN-LSA [238]	2020	BinaryMask	DepthNet	RGB	S	high frequent spoof cues from lossless size adaptation module
CDCN [23]	2020	Depth	DepthNet	RGB	S	intrinsic detailed patterns via aggregating both intensity and gradient information from stacked central difference convolutions.
FAS-SGTD [96]	2020	Depth	DepthNet STPM	RGB	D	detailed discriminative dynamics cues from stacked Residual Spatial Gradient Block and Spatio-Temporal Propagation Module
TS-FEN [139]	2020	Depth	ResNet34 FCN	RGB YCbCr HSV	S	discriminative fused features from depth-stream and chroma-stream networks
SAPLC [39]	2020	TernaryMap	DepthNet	RGB HSV	S	accurate image-level decision via spatial aggregation of pixel-level local classifiers even with insufficient training samples
BCN [24]	2020	BinaryMask Depth Reflection	DepthNet	RGB	S	intrinsic material-based patterns captured via aggregating multi-level bilateral macro- and micro- information
Disentangled [97]	2020	Depth TextureMap	DepthNet	RGB	S	liveness and content features via disentangled representation learning
AENet [44]	2020	Depth Reflection	ResNet18	RGB	S	rich semantic features using Auxiliary Information, Embedding Network with multi-task learning framework
3DPC-Net [40]	2020	3D Point Cloud	ResNet18	RGB	S	discriminative features via fine-grained 3D Point Cloud supervision
PS [211]	2020	BinaryMask Depth	ResNet50 CDCN	RGB	S	pyramid supervision guides models to learn both local details and global semantics from multi-scale spatial context
NAS-FAS [37]	2020	BinaryMask Depth	NAS	RGB	D	leveraging cross-domain/type knowledge and static-dynamic representation for central difference network search
DAM [239]	2021	Depth	VGG16 TSM	RGB	D	attentional fused depth and multi-scale temporal clues using a two-stream network as well as a self-supervised symmetry loss
Bi-FPNFAS [144]	2021	Fourier spectra	EfficientNetB0 FPN	RGB	S	multiscale bidirectional propagated features with self-generated frequency spectra supervision
DC-CDN [98]	2021	Depth	CDCN	RGB	S	efficient feature learning on dual-cross central difference network with Cross Feature Interaction Modules
De-Spoof [33]	2018	Depth BinaryMask FourierMap	DSNet DepthNet	RGB HSV	S	inversely decomposing a spoof face into a spoof noise and a live face, and estimating subtle spoof noise with proper supervisions
Reconstruction [240]	2019	RGB Input (live) ZeroMap (spoof)	U-Net	RGB	S	multi-level semantic features from autoencoder
LGSC [41]	2020	ZeroMap (live)	U-Net ResNet18	RGB	S	discriminative live-spoof differences learned within a residual- learning framework with the perspective of anomaly detection
TAE [138]	2020	Binary CE loss Reconstruction loss	Info-VAE+ DenseNet161	RGB	S	self-pretrained autoencoder in large-scale face recognition datasets to obtain the reconstruction-error images for FAS
STDN [99]	2020	BinaryMask RGB Input (live)	U-Net PatchGAN	RGB	S	disentangled spoof trace via adversarial learning and hierarchical combination of patterns at multiple scales
GOGen [188]	2020	RGB input	DepthNet	RGB+one- hot vector	S	GAN-based architecture to synthesize and identify the spoof noise patterns from medium/sensor combinations
PhySTD [42]	2021	Depth RGB Input (live)	U-Net PatchGAN	Frequency Trace	S	disentangling spoof faces into the spoof traces and live counterparts guided by physical properties
MT-FAS [43]	2021	ZeroMap (live) LearnableMap (Spoof)	DepthNet	RGB	S	train a meta-teacher to generate optimal pixel-wise signals for supervising the spoofing detector

TABLE 10: Summary of the representative **generalized deep learning** FAS methods to **unseen domain (domain adaptation and domain generalization)**. 'MMD' is short for 'Maximum Mean Discrepancy'.

	Method	Year	Backbone	Loss	S/D	Description
Domain Adaptation	OR-DA [78]	2018	AlexNet	Binary CE loss MMD loss	S	learned classifier for target domain, and embedding space with similar distribution for source and target domains
	DTCNN [241]	2019	AlexNet	Binary CE loss MMD loss	S	domain invariant features using a few labeled samples from the target domain
	Adversarial [146]	2019	ResNet18	Triplet loss Adversarial loss	S	learn a shared embedding space by both source and target domain models via adversarial domain adaptation
	ML-MMD [149]	2019	Multi-scale FCN	CE loss MMD loss	S	adapt in both representation and classifier layers to bridge for the domain discrepancy
	OCA-FAS [54]	2020	DepthNet	Binary CE loss Pixel-wise binary loss	S	train a meta-learner with loss function search on one-class adaptation FAS tasks with only live samples
	DR-UDA [147]	2020	ResNet18	Center&Triplet loss Adversarial loss Disentangled loss	S	disentangles the features irrelevant to specific domains, and learn a shared embedding space by both source and target domains
	DGP [150]	2020	DenseNet161	Feature divergence measure BinaryMask	S	prune the filters specific to the source dataset for performance improvement on target dataset
	Distillation [151]	2020	AlexNet	Binary CE loss MMD loss Paired Similarity	S	spoofing-specific information captured by distilled deep network on the application-specific domain
	S-CNN +PL+TC [148]	2021	ResNet18	CE Loss in labeled and unlabeled sets	D	semi-supervised learning framework with only a few labeled training data, and progressively adopt the unlabeled data with reliable pseudo labels.
	USDAN [103]	2021	ResNet18	Adaptive binary CE loss Entropy loss Adversarial loss	S	design different distribution alignment operations to enhance generalization for un- & semi-supervised domain adaptation to address cross-scenario problem
Domain Generalization	MADDG [48]	2019	DepthNet	Binary CE & Depth loss Multi-adversarial loss Dual-force Triplet loss	S	leverage the large variability present in FR datasets to induce invariance to factors that cause domain-shift
	PAD-GAN [50]	2020	ResNet18	Binary CE & GAN loss Reconstruction loss	S	disentangled and domain-independent features rather than subject discriminative and domain related features
	SSDG [51]	2020	ResNet18	Binary CE loss Single-Side adversarial loss Asymmetric Triplet loss	S	learn a generalized space where the feature distribution of real faces is compact while that of fake ones is dispersed among domains but compact within each domain
	RF-Meta [49]	2020	DepthNet	Binary CE loss Depth loss	S	meta-learned generalized features across multiple source domains with auxiliary regularization
	CCDD [242]	2020	ResNet50 +LSTM	Binary CE loss Class-conditional loss	D	learn discriminative but domain-robust features with class-conditional domain discriminator module and GRL
	DASN [212]	2021	ResNet18	Binary CE & Spoof-irrelevant factor loss	S	adopt doubly adversarial learning to suppress the spoof-irrelevant factors, and intensify spoof factors.
	SDA [154]	2021	DepthNet	Binary CE & Depth loss Reconstruction loss Orthogonality regularization	S	use meta-learning based adaptor learning for better adaptor initialization, and an unsupervised adaptor loss for appropriate adaptor optimization
	D^2AM [152]	2021	DepthNet	Binary CE loss Depth loss MMD loss	S	iteratively divide mixture domains via discriminative domain representation and train generalizable models with meta-learning without using domain labels
	DRDG [**?**]	2021	DepthNet	Binary CE & Depth loss domain loss	S	iteratively reweight the relative importance between samples and features to extract domain-irrelevant features
	ANRL [153]	2021	DepthNet	Binary CE & Depth loss inter-domain compatible loss inter-class separable loss	S	adaptively select feature normalization methods to learn domain-agnostic and discriminative representation
	FGHV [102]	2022	DepthNet	Variance, relative correlation distribution discrimination constraints	S	feature generation networks generate hypotheses of real faces and known attacks, and two hypothesis verification modules are applied to judge real/generative distributions
	SSAN [102]	2022	DepthNet ResNet	Binary CE loss domain adversarial loss contrastive loss	S	extract and reassemble different content and style features for a stylized feature space, and emphasize liveness-related style information while suppress the domain-specific one

TABLE 11: Summary of the **generalized deep learning** FAS methods to **unknown attack types (zero/few-shot learning and anomaly detection)**. 'OCSVM', 'MD', 'GMM', and 'OCCL' are short for 'One-Class Support Vector Machine', 'Mahalanobis-distance', 'Gaussian Mixture Model', and 'One-Class Contrastive Loss', respectively.

	Method	Year	Backbone	Loss	Input	Description
Zero/Few-Shot	DTN [38]	2019	Deep Tree Network	Binary CE loss Pixel-wise binary loss Unsupervised Tree loss	RGB HSV	adaptively routing the attacks to the most similar spoof cluster, and makes the binary decision
	AIM-FAS [53]	2020	DepthNet	Depth loss Contrastive Depth loss	RGB	adaptive inner-updated meta features generalized to unseen spoof types from predefined PAs
	CM-PAD [155]	2021	DepthNet ResNet	Binary CE loss Depth loss Gradient alignment	RGB	continual meta-learning PAD solution that can be trained on unseen attack scenarios catastrophic seen attack forgetting
Anomaly-Detection	AE+LBP [157]	2018	AutoEncoder	Reconstruction loss	RGB	embedding features (cascaded with LBP) from outlier detection based neural network autoencoder
	Anomaly [159]	2019	ResNet50	Triplet focal loss Metric-Softmax loss	RGB	deep anomaly detection via introducing a few-shot posteriori probability estimation
	Anomaly2 [243]	2019	GoogLeNet ResNet50	MD	RGB	subject specific anomaly detector is trained on genuine accesses only using one-class classifiers
	Hypersphere [161]	2020	ResNet18	Hypersphere loss	RGB HSV	deep anomaly detection supervised by hypersphere loss, and detects PAs directly on learned feature space
	Ensemble-Anomaly [244]	2020	GoogLeNet ResNet50	GMM (not end-to-end)	RGB patches	ensemble of one-class classifiers from different facial regions, CNNs, and anomaly detectors
	MCCNN [160]	2020	LightCNN	Binary CE loss Contrastive loss	Grayscale, IR, Depth, Thermal	learn a compact embedding for bonafide while being far from the representation of PAs via OCCL, and cascaded with a one-class GMM
	End2End-Anomaly [158]	2020	VGG-Face	Binary CE loss Pairwise confusion	RGB	both classifier and representations are learned end-to-end with pseudo negative class
	ClientAnomaly [245]	2020	ResNet50 GoogLeNet VGG16	OCSVM GMM MD	RGB	client-specific knowledge are leveraged for anomaly-based spoofing detectors as well as determination thresholds

TABLE 12: Summary of the representative **deep learning** FAS methods with **specialized sensor/hardware inputs**. 'S/D', 'SD', 'AD', 'FM', 'APD', 'LFC', 'DP' and 'DOLP' are short for 'Static/Dynamic', 'Square Disparity', 'Absolute Disparity', 'Feature Multiplication', 'Approximate Disparity', 'Light Field Camera', 'Dual Pixel' and 'Degree of Linear Polarization', respectively.

Method	Year	Backbone	Loss	Input	S/D	Description
Thermal-FaceCNN [167]	2019	AlexNet	Regression loss	Thermal infrared face image	S	temperature related features based on the fact that real face temperature is 36∼37 degrees on average
SLNet [162]	2019	17-layer CNN	Binary CE loss	Stereo (left&right) face images	S	disparities between deep features are learned using SD, AD, FM, and APD operations
Aurora Guard [163]	2019	U-Net	Binary CE loss Depth regression Light Regression	Casted face with dynamic changing light specified by random light CAPTCHA	D	based on the normal cues extracted from the light reflection, multi-tasck CNN recovers both subjects' depth maps and light CAPTCHA
LFC [87]	2019	AlexNet	Binary CE loss	Ray difference/microlens images from LFC	S	meaningful features extracted from single-shot LFC images with rich depth information of objects
PAAS [47]	2020	MobileNetV2	Contrastive loss SVM	Four-directional polarized face image	S	learned discriminative and robust features from DOLP as polarization reveals the intrinsic attributes
Face-Revelio [170]	2020	Siamese-AlexNet	L1 distance	Four flash lights displayed on four quarters of a screen	D	varying illumination enables the recovery of the face surface normals via photometric stereo
SpecDiff [171]	2020	ResNet4	Binary CE loss	Concatenated face images w/ and w/o flash	S	a novel descriptor based on specular and diffuse reflections, with a flash-based deep FAS baseline
MC-PixBiS [55]	2020	DenseNet161	Binary mask loss	SWIR images differences	S	discriminative features for Impersonation attacks as water is very absorbing in some SWIR wavelengths
Thermal-ization [246]	2020	YOLO V3+ GoogLeNet	Binary CE loss	Thermal infrared face image	S	learned specific physical features from Thermal infrared imaging of PAs
DP Bin--Cls-Net [169]	2021	Shallow U-Net + Xception	Transformation consistency Relative disparity loss Binary CE loss	DP image pair	S	reconstructed depth based on the DP pair with self-supervised loss for planar attack detection

TABLE 13: Summary of the **multi-modal deep learning** FAS methods. 'MFEM', 'SPM', 'LFV' and 'MLP' are short for 'Modal Feature Erasing Module', 'Selective Modal Pipeline', 'Limited Frame Vote' and 'Multilayer Perceptron', respectively.

Method	Year	Backbone	Loss	Input	Fusion	Description
FaceBagNet [174]	2019	Multi-stream CNN	Binary CE loss	RGB, Depth, NIR face patches	Feature-level	spoof-specific features from patch CNN, and MFEM to prevent overfitting and better fusion
FeatherNets [178]	2019	Ensemble-FeatherNet	Binary CE loss	Depth, NIR	Decision-level	single compact FeatherNet trained by depth image, then fused with "ensemble + cascade" structure
Attention [247]	2019	ResNet18	Binary CE loss Center loss	RGB, Depth, NIR	Feature-level	using channel and spatial attention module to refine the multimodal features
mmfCNN [173]	2019	ResNet34	Binary CE loss Binary Center Loss	RGB, NIR, Depth, HSV, YCbCr	Feature-level	fuses multi-level features among modalities in a unified framework with weight-adaptation
MM-FAS [172]	2019	ResNet18/50	Binary CE loss	RGB, NIR, Depth	Feature-level	leverages multimodal data and aggregates intra-channel features at multiple network layers
AEs+MLP [176]	2019	Autoencoder MLP	Binary CE loss Reconstruction loss	Grayscale-Depth-Infrared composition	Input-level	trasfer learning within facial patches from the facial RGB appearance to multi-channel modalities
SD-Net [28]	2019	ResNet18	Binary CE loss	RGB, NIR, Depth	Feature-level	multimodal fusion via feature re-weighting to select more informative channels for modalities
Dual-modal [248]	2019	MoblienetV3	Binary CE loss	RGB, IR	Feature-level	light-weight networks to extract and merge embedding features from NIR-VIS image pairs
Parallel-CNN [249]	2020	Attentional-CNN	Binary CE loss	Depth, NIR	Feature-level	fused deep depth and IR features from paralleled attentional CNN with spatial pyramid pooling
Multi-Channel Detector [250]	2020	RetinaNet (FPN+ResNet18)	Landmark regression Focal loss	Grayscale-Depth-Infrared composition	Input-level	learned joint face detection-based and PAD-based representation from fused 3 channel images
PSMM-Net [91]	2020	ResNet18	Binary CE loss for each stream	RGB, Depth, NIR	Feature-level	static-dynamic fusion mechanism with partially shared fusion strategy is proposed
PipeNet [251]	2020	SENet154	Binary CE loss	RGB, Depth, NIR face patches	Feature-level	SMP takes full advantage of multi-modal data. LFV ensures stable video-level prediction
MM-CDCN [27]	2020	CDCN	Pixel-wise binary loss, Contrastive depth loss	RGB, Depth, NIR	Feature& Decision level	capture central-difference-based intrinsic spoofing patterns among three modalities
HGCNN [252]	2020	Hypergraph-CNN + MLP	Binary CE loss	RGB, Depth	Feature-level	auxiliary depth fused with texture in the feature domain from hypergraph convolution
MCT-GAN [179]	2020	CycleGAN ResNet50	GAN loss Binary CE loss	RGB, NIR	Input-level	generate NIR counterpart for VIS inputs via GAN, and learn fusing features
D-M-Net [177]	2021	ResNeXt	Binary CE loss	Multi-preprocessed Depth, RGB-NIR composition	Input& Feature-level	two-stage cascade architecture to fuse depth features with multi-scale RGB-NIR composite features
CMFL [175]	2021	DenseNet161	Cross modal focal loss Binary CE loss	RGB, Depth	Feature-level	modulate the loss contribution and complementary information from the two modalities
MA-Net [180]	2021	CycleGAN ResNet18	GAN loss Binary CE loss	RGB, NIR	Feature-level	translate the visible inputs into NIR images, and then extract VIS-NIR features
FlexModal -FAS [187]	2022	CDCN ResNet50, ViT	Binary CE loss Pixel-wise binary loss	RGB, Depth, NIR	Feature-level	cross-attention fusion to efficiently mine cross-modal clues for flexible-modal deployment