Unified Detection of Digital and Physical Face Attacks

Debayan Deb, Xiaoming Liu, Anil K. Jain
Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI, 48824
{debdebay, liuxm, jain}@cse.msu.edu

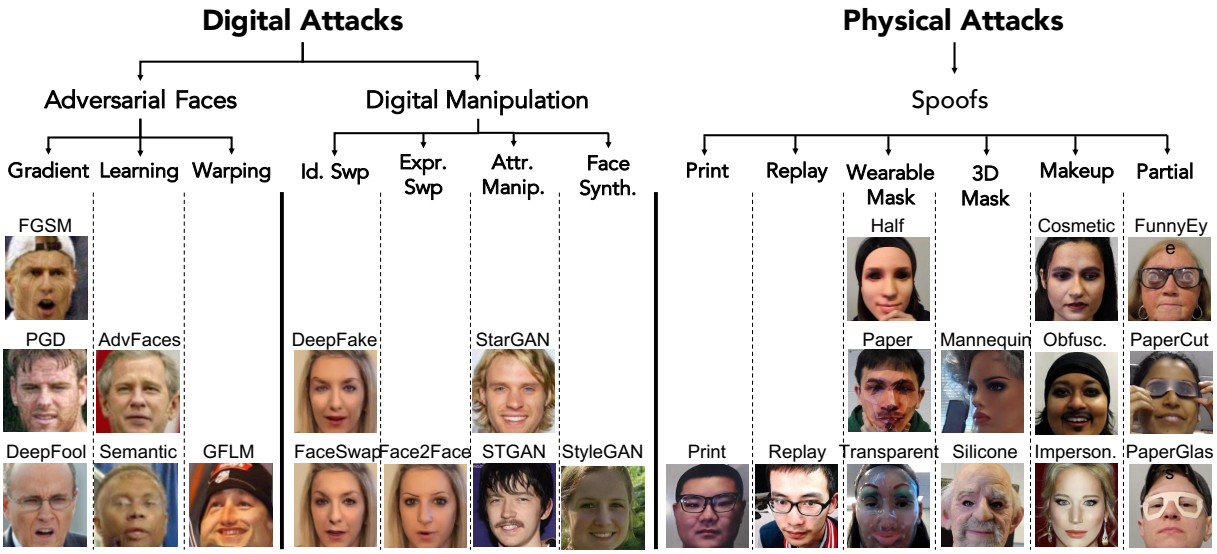

Fig. 1: Face attacks against AFR systems are continuously evolving in both digital and physical spaces. Given the diversity of the face attacks, prevailing methods fall short in detecting attacks across all three categories (*i.e.*, adversarial, digital manipulation, and spoofs). This work is among the first to define the task of face attack detection on the 25 attack types across 3 categories shown here.

***Abstract*— State-of-the-art defense mechanisms against face attacks achieve near perfect accuracies within one of three attack categories, namely adversarial, digital manipulation, or physical spoofs, however, they fail to generalize well when tested across all three categories. Poor generalization can be attributed to learning incoherent attacks jointly. To overcome this shortcoming, we propose a unified attack detection framework, namely UniFAD, that can automatically cluster 25 coherent attack types belonging to the three categories. Using a multi-task learning framework along with k-means clustering, UniFAD learns joint representations for coherent attacks, while uncorrelated attack types are learned separately. Proposed UniFAD outperforms prevailing defense methods and their fusion with an overall TDR = 94.73% @ 0.2% FDR on a large fake face dataset consisting of $341K$ bona fide images and $448K$ attack images of 25 types across all 3 categories. Proposed method can detect an attack within 3 milliseconds on a Nvidia 2080Ti. UniFAD can also identify the attack categories with 97.37% accuracy. Code and dataset will be publicly available.**

I. INTRODUCTION

Automated face recognition (AFR) systems have been projected to grow to USD 3.35B by 2024[1]. It is estimated that over a billion smartphones today unlock via face authentication[2]. However, the foremost challenge facing AFR systems is their vulnerability to *face attacks*. For instance, an attacker can hide his identity by wearing a 3D mask [29], or intruders can assume a victim's identity by digitally swapping their face with the victim's face image [13]. With unrestricted access to the rapid proliferation of face images on social media platforms, launching attacks against AFR systems has become even more accessible. Given the growing dissemination of "fake news" and "deepfakes" [3], the research community and social media platforms alike are pushing towards *generalizable* defense against continuously evolving and sophisticated face attacks.

In literature, face attacks can be broadly classified into three attack categories: (i) Spoof attacks: artifacts in the *physical* domain (*e.g.*, 3D masks, eye glasses, replaying videos) [38], (ii) Adversarial attacks: imperceptible noises added to probes for evading AFR systems [64], and (iii) Digital manipulation attacks: entirely or partially modified photo-realistic faces using generative models [13]. Within each of these categories, there are different attack types. For example, each spoof medium, *e.g.*, 3D mask and makeup, constitutes one attack type, and there are 13 common types of spoof attacks [38]. Likewise, in adversarial and digital manipulation attacks, each attack model, designed by unique objectives and losses, may be considered as one attack type. Thus, the attack categories and types form a 2-layer tree structure encompassing the diverse attacks (see Fig. 1). Such

[1] https://bwnews.pr/2OqY0nD
[2] https://bit.ly/30vYBHg

979-8-3503-4544-5/23/$31.00 ©2023 IEEE

a tree will inevitably grow in the future.

In order to safeguard AFR systems against these attacks, numerous face attack detection approaches have been proposed [13], [14], [17]–[20], [34], [39], [47], [49], [59]. Despite impressive detection rates, prevailing research efforts focus on a few attack types within *one* of the three attack categories. Since the exact type of face attack may not be known *a priori*, a generalizable detector that can defend an AFR system against any of the three attack categories is of utmost importance.

Due to the vast diversity in attack characteristics, from glossy 2D printed photographs to imperceptible perturbations in adversarial faces, we find that learning a single *unified* network is inadequate. Even when prevailing state-of-the-art (SOTA) detectors are trained on all 25 attack types, they fail to generalize well during testing. Via ensemble training, we comprehensively evaluate the detection performance on fusing decisions from three SOTA detectors that individually excel at their respective attack categories. However, due to the diversity in attack characteristics, decisions made by each detector may not be complementary and result in poor detection performance across all 3 categories.

This research is among the first to focus on detecting *all* 25 *attack types* known in literature (6 adversarial, 6 digital manipulation, and 13 spoof attacks). Our approach consists of (i) automatically clustering attacks with similar characteristics into distinct groups, and (ii) a multi-task learning framework to learn salient features to distinguish between bona fides and coherent attack types, while early sharing layers learn a joint representation to distinguish bona fides from any generic attack.

This work makes the following contributions:

- Among the first to define the task of face attack detection on 25 attack types across 3 attack categories: adversarial faces, digital face manipulation, and spoofs.
- A novel **uni**fied **f**ace **a**ttack **d**etection framework, namely *UniFAD*, that automatically clusters similar attacks and employs a multi-task learning framework to detect digital and physical attacks.
- Proposed *UniFAD* achieves SOTA detection performance, TDR = 94.73% @ 0.2% FDR on a large fake face dataset, namely *GrandFake*. To the best of our knowledge, *GrandFake* is the largest face attack dataset studied in literature in terms of the number of diverse attack types.
- Proposed *UniFAD* allows for further identification of the attack categories, *i.e.*, whether attacks are adversarial, digitally manipulated, or contains physical spoofing artifacts, with a classification accuracy of 97.37%.

II. RELATED WORK

Individual Attack Detection. Early work on face attack detection primarily focused on one or two attack types in their respective categories. Studies on adversarial face detection [21], [23] primarily involved detecting gradient-based attacks, such as FGSM [22], PGD [42], and DeepFool [48].

	Study	Year	# BonaFides	# Attacks	# Types
Adversarial	UAP-D [1]	2018	9,959	29,877	1
	Goswami *et al.* [23]	2019	16,685	50,055	3
	Agarwal *et al.* [2]	2020	24,042	72,126	3
	Massoli *et al.* [44]	2020	169,396	$1M$	6
	FaceGuard [15]	2020	507,647	$3M$	6
Digital Manip.	Yang *et al.* [62]	2018	$241(I)/49(V)$	$252(I)/49(V)$	1
	DeepFake [32]	2018	–	$620(V)$	1
	FaceForensics++ [32]	2019	$1,000(V)$	$3,000(V)$	3
	FakeSpotter [60]	2019	6,000	5,000	2
	DFFD [13]	2020	58,703	240,336	7
Phys. Spoofs	Replay-Attack [9]	2012	$200(V)$	$1,000(V)$	3
	MSU MFSD [61]	2015	$160(V)$	$280(V)$	3
	OuluNPU [7]	2017	$990(V)$	$3,960(V)$	4
	SiW [37]	2018	$1,320(V)$	$3,158(V)$	6
	SiW-M [38]	2019	$660(V)$	$960(V)$	13
	GrandFake (ours)	**2022**	**341,738**	**447,674**	**25**

TABLE I: Face attack datasets with no. of bona fide images, no. of attack images, and no. of attack types. Here, I denotes images and V refers to videos. *GrandFake* will be publicly available.

DeepFakes were among the first studied digital attack manipulation [32], [62], however, generalizability of the proposed methods to a larger number of digital manipulation attack types is unsatisfactory [33]. Majority of face anti-spoofing methods focus on print and replay attacks [5], [7], [37], [43], [52], [55], [61], [63].

Over the years, a clear trend in the increase of attack types in each category can be observed in Tab. I. Since a community of attackers dedicate their efforts to craft new attacks, it is imperative to comprehensively evaluate existing solutions against a large number of attack types.

Joint Attack Detection. Recent studies have used multiple attack types in order to defend against face attacks. For *e.g.*, FaceGuard [15] proposed a generalizable defense against 6 adversarial attack types. The Diverse Fake Face Dataset (DFFD) [13] includes 7 digital manipulation attack types. In the spoof attack category, recent studies focus on detecting 13 spoof types.

Majority of the works tackling multiple attack types pose detection as a binary classification problem with a single network learning a joint feature space. For simplicity, we refer to such a network architecture as *JointCNN*. For instance, it is common in adversarial face detection to train a JointCNN with bona fide faces and adversarial attacks synthesized on-the-fly by a generative network [15], [28], [36], [50]. On the other hand, majority of the proposed defenses against digital manipulation, fine-tune a pre-trained JointCNN (*e.g.*, Xception) on bona fide faces and all available digital manipulation attacks [4], [8], [13], [53], [60]. Due to the availability of a wide variety of physical spoof artifacts in face anti-spoofing datasets (*e.g.*, eyeglasses, print and replay instruments, masks, *etc*) along with evident cues for detecting them, studies on anti-spoofs are more sophisticated. The associated JointCNN employs either auxiliary cues, such as depth map and heart pulse signals (rPPG) [37], [56], [66], or a "compactness" loss to prevent overfitting [20], [45].

While jointly detecting multiple attack types is promising, detecting attack types *across* different categories is of the utmost importance. An early attempt proposed a defense

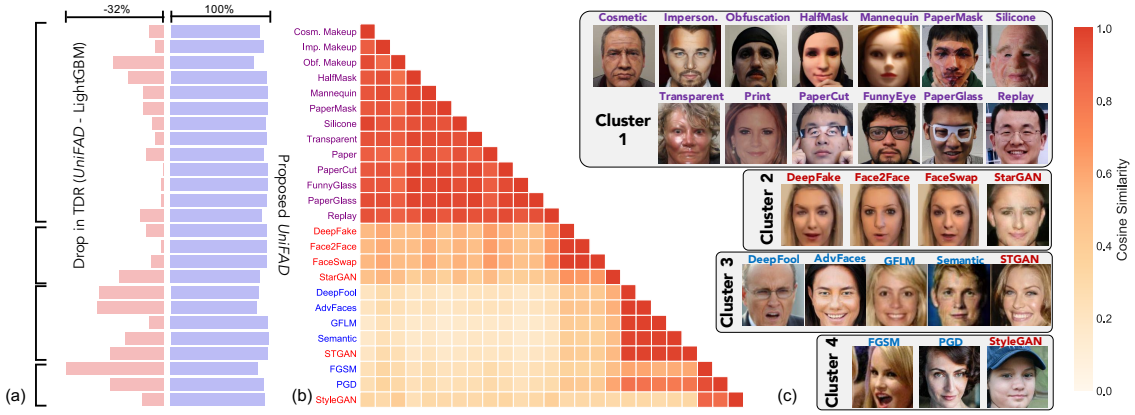

Fig. 2: (a) Detection performance (TDR @ 0.2% FDR) in detecting each attack type by the proposed *UniFAD* (purple) and the difference in TDR from the best fusion scheme, LightGBM [31] (pink). (b) Cosine similarity between mean features for 25 attack types extracted by *JointCNN*. (c) Examples of attack types from 4 different clusters via k-means clustering on JointCNN features. Attack types in purple, blue, and red denote spoofs, adversarial, and digital manipulation attacks, respectively.

against 4 attack types on 2 categories, 3 spoofs and 1 digital manipulation [45]. On the other hand, Ibsen *et al.* [27] proposes a JointCNN trained on face identity embeddings for 2 categories, 3 spoofs and 3 digital manipulation attacks, to achieve the same. To the best of our knowledge, we are the first to attempt detecting **25** attack types across **3** categories.

Multi-task Learning. In multi-task learning (MTL), a task, $\mathcal{T}_i$ is usually accompanied by a training dataset, $\mathcal{D}_{tr}$ consisting of N_t training samples, *i.e.*, $D_{tr} = \{\mathbf{x}_i^{tr}, y_i^{tr}\}_{i=1}^{N_{tr}}$, where $\mathbf{x}_i^{tr} \in \mathbb{R}$ is the ith training sample in $\mathcal{T}_i$ and y_i^t is its label. Most MTL methods rely on well-defined tasks [41], [46]. Crawshaw *et al.* [11] summarize various works on MLT with CNNs. In this work, we propose a MTL framework in an extreme situation where only a single task is available (bona fide vs. 25 attack types) and utilize k-means clustering to construct new auxiliary tasks from $\mathcal{D}_{tr}$. A recent study also utilized k-means for constructing new tasks, however, their approach utilizes a meta-learning framework where the groups themselves can alter throughout training [24]. Instead, we propose a new unified attack detection framework that first utilizes k-means to partition the 25 attacks types, and then learns shared and attack-specific representations to distinguish them from bona fides.

III. Dissecting Prevailing Defense Systems

A. Datasets

In order to detect 25 attack types (6 adversarial, 6 digital manipulation, and 13 spoofs), we propose the *GrandFake* dataset, an amalgamation of multiple face attack datasets from each category. We provide additional details of *GrandFake* in Sec. V-A.

B. Drawback of JointCNN

Consider the diversity in the available attacks: from imperceptible adversarial perturbations to digital manipulation attacks, both of which are entirely different from physical print attacks (hard surface, glossy, 2D). Even within the spoof category, characteristics of mask attacks are quite different from replay attacks. In addition, discriminative cues for some attack types may be observed in high-frequency domain (*e.g.*, defocused blurriness, chromatic moment), while others exhibit low-frequency cues (*e.g.*, color diversity and specular reflection). For these reasons, learning a common feature space to discriminate all attack types from bona fides is challenging and a JointCNN may fail to generalize well even on attack types seen during training.

We demonstrate this by first training a JointCNN on the 25 attack types in *GrandFake* dataset. We then compute an *attack similarity matrix* between the 25 types (see Fig. 2(b)). The mean feature for each attack type is first computed on a validation set composed of $1,000$ images per attack. We then compute the pairwise cosine similarity between mean features from all attack pairs. From Fig. 2, we note that physical attacks have little correlation with adversarial attacks and therefore, learning them jointly within a common feature space may degrade detection performance.

Although prevailing JointCNN-based defense achieve near perfect detection when trained and evaluated on the respective attack types in isolation, we observe a significantly degraded performance when trained and tested on all 3 attack categories together (see Tab. II). In other words, even when a prevailing SOTA defense system is trained on all 3 categories, it may lead to degraded performance on testing.

C. Unifying Multiple JointCNNs

Another possible approach is to consider ensemble techniques; instead of using a single JointCNN detector, we can fuse decisions from multiple individual detectors that are already experts in distinguishing between bona fides and attacks from their respective attack category. Given three SOTA detectors, one per attack category, we perform a comprehensive evaluation on parallel and sequential score-level fusion schemes.

In our experiments, we find that, indeed, fusing score-level decisions from single-category detectors outperforms a single

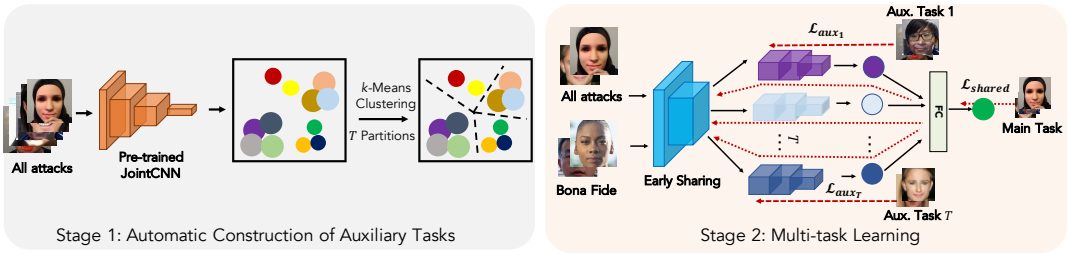

Fig. 3: An overview of training *UniFAD* in two stages. Stage 1 automatically clusters coherent attack types into T groups. Stage 2 consists of a MTL framework where early layers learn generic attack features while T branches learn to distinguish bona fides from coherent attacks.

SOTA defense system trained on all attack types. Note that efforts in utilizing prevailing defense systems rely on the assumption that attack categories are independent of each other. However, Fig. 2 shows that some digital manipulation attacks, such as STGAN and StyleGAN, are *more closely related* to some of the adversarial attacks (*e.g.*, AdvFaces, GFLM, and Semantic) than other digital manipulation types. This is likely because all five methods utilize a GAN to synthesize their attacks and may share similar attack characteristics. Therefore, a SOTA adversarial detector and a SOTA digital manipulation detector may individually excel at their respective categories, but may not provide complementary decisions when fused. Instead of training detectors on groups with manually assigned semantics (*e.g.*, adversarial, digital manipulation, spoofs), it is better to train JointCNNs on coherent attacks. In addition, utilizing decisions from pre-trained JointCNNs may tend to overfit to the attack categories used for training.

IV. Proposed Method: UniFAD

We propose a new multi-task learning framework for Unified Attack Detection, namely *UniFAD*, by training an end-to-end network for improved physical and digital face attack detection. In particular, a k-means augmentation module is utilized to automatically construct auxiliary tasks to enhance single task learning (such as a JointCNN). Then, a joint model is decomposed into a feature extractor (shared layers) $\mathcal{F}$ that is shared across all tasks, and task-specific branches for each auxiliary task. Fig. 3 illustrates the auxiliary task creation and the training process of *UniFAD*.

A. Problem Definition

Let the "main task" be defined as the overall objective of a unified attack detector: given an input image, $\mathbf{x}$, assign a score close to 0 if $\mathbf{x}$ is bona fide or close to 1 if $\mathbf{x}$ is any of the available face attack types. We are also given a labeled training set, D_{tr}. Prevailing defenses follow a single task learning approach where the main task is adopted to be the ultimate training objective. In order to avoid the shortcomings of a JointCNN and unification of multiple JointCNNs, we first use D_{tr} to automatically construct multiple auxiliary tasks $\{\mathcal{T}_t\}_{t=1}^{T}$, where T_i is the ith cluster of coherent attack types. If the auxiliary tasks are appropriately constructed, jointly learning these tasks along with the main task should improve unified attack detection compared to a single task learning approach.

B. Automatic Construction of Auxiliary Tasks

One way to construct auxiliary tasks is to train a separate binary JointCNN on each attack type. Such partitioning massively increases computational burden (*e.g.*, training and testing 25 JointCNNs). Other simple partitioning methods, such as randomly partition are likely to cluster uncorrelated attacks. On the other hand, clustering in the pixel-space is also unappealing due to poor correlation between the distances in the pixel-space, and clustering in the high-dimensional space is challenging [25]. Therefore, we require a reasonable alternative to manual inspection of the attack similarity matrix in Fig. 2 to partition the attack types into appropriate clusters.

Fortunately, we already have a JointCNN trained via a single task learning framework that can extract salient representations. Thus, we can map the data $\{\mathbf{x}\}$ into JointCNN's embedding space $\mathcal{Z}$, producing $\{\mathbf{z}\}$. We can then utilize a traditional clustering algorithm, k-means, which takes a set of feature vectors as input, and clusters them into k distinct groups based on a geometric constraint. Specifically, for each attack type, we first compute the mean feature. We then utilize k-means clustering to partition the L features into $T(\leq L)$ sets, $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_T\}$ such that within-cluster sum of squares (WCSS) is minimized,

$$\arg\min_{\mathcal{P}} \sum_{i=1}^{T} \sum_{\bar{\mathbf{z}} \in \mathcal{P}_i} ||\bar{\mathbf{z}} - \mu_i||^2, \tag{1}$$

where, $\bar{\mathbf{z}}$ represents a mean feature for an attack type and μ_i is the mean of the features in $\mathcal{P}_i$. Fig. 2(c) shows an example on clustering the 25 attack types of *GrandFake*.

C. Multi-Task Learning with Constructed Tasks

With a multi-task learning framework, we learn coherent attack types jointly, while uncorrelated attacks are learned in their own feature spaces. We construct T "branches" where each branch is a neural network trained on a binary classification problem (*i.e.*, aux. task). The learning objective of each branch, $\mathcal{B}_t$, is to minimize,

$$\mathcal{L}_{aux_t} = \mathbb{E}_{\mathbf{x}}\left[\log \mathcal{B}_t(\mathbf{x}_{bf})\right] + \mathbb{E}_{\mathbf{x}}\left[\log\left(1 - \mathcal{B}_t(\mathbf{x}_{fake}^{\mathcal{P}_t})\right)\right]. \tag{2}$$

where $\mathbf{x}_{bf}$ denotes bona fide images and $\mathbf{x}_{fake}^{\mathcal{P}_t}$ is face attacks corresponding to the attack types in the partition $\mathcal{P}_t$.

D. Parameter Sharing

Early Sharing. We adopt a hard parameter sharing module which learns a common feature representation for distinguishing between bona fides and attacks prior to aux. task learning branches. Baxter [6] demonstrated the shared parameters have a lower risk of overfitting than the task-specific parameters.

Therefore, adopting early convolutional layers as a preprocessing step prior to branching can help *UniFAD* in its generalization to all 3 categories. We construct hidden layers between the input and the branches to obtain shared features, $\mathbf{h} = \mathcal{F}(\mathbf{x})$, while the auxiliary learning branches output $\mathcal{B}_t(\mathbf{h})$.

Late Sharing. Each branch $\mathcal{B}_t$ is trained to output a decision score where scores closer to 0 indicate that the input image is a bona fide, whereas, scores closer to 1 correspond to attack types pertaining to the branch's partition. The scores from all T branches are then concatenated and passed to a final decision layer. For simplicity, we define the final decision output as, $FC(\mathbf{x}) := FC(\mathcal{B}_1(\mathbf{h}), \mathcal{B}_2(\mathbf{h}), \ldots, \mathcal{B}_T(\mathbf{h}))$.

The early shared layers and the final decision layer are learned via a binary cross-entropy loss,

$$\mathcal{L}_{shared} = \mathbb{E}_{\mathbf{x}}\left[\log FC(\mathbf{x}_{bf})\right] + \mathbb{E}_{\mathbf{x}}\left[\log\left(1 - FC(\mathbf{x}_{fake})\right)\right], \tag{3}$$

between bona fides and all available attack types.

E. Training and Testing

The entire network is trained in an end-to-end manner by minimizing the following composite loss,

$$\mathcal{L}_{UniFAD} = \mathcal{L}_{shared} + \sum_{t=1}^{T} \mathcal{L}_{aux_t}. \tag{4}$$

The $\mathcal{L}_{shared}$ loss is backpropagated throughout *UniFAD*, while $\mathcal{L}_{aux_t}$ is only responsible for updating the weights of the branch, $\mathcal{B}_t$, and the final classification layer. For the forward and backward passes of $\mathcal{L}_{shared}$, an equal number of bona fide and attack samples are used for training. On the other hand, for training each branch, $\mathcal{B}_t$, we sample the equal number of bona fides and equal number of attack images from the attack partition $\mathcal{P}_t$.

Attack Detection. During testing, an image passes through the shared layers and then each branch of *UniFAD* outputs a decision whether the image is bona fide (values close to 0) or an attack (close to 1). The final decision layer outputs the final decision score. Unless stated otherwise, we use the final decision scores to report performance.

Attack Classification. Once an attack is detected, *UniFAD* can automatically classify the attack type and category. For all L attack types in the training set, we extract intermediate 128-dim feature vectors from T branches. The features are then concatenated and the mean feature across all L attack types is computed, such that, we have L feature vectors of size $T \times 128$. For a detected attack, Cosine similarity is computed between the testing sample's feature vector and the mean training features for L types. The predicted attack type is the one with the highest similarity score.

V. EXPERIMENTAL RESULTS

A. Experimental Settings

Dataset. *GrandFake* consists of 25 face attacks from 3 attack categories. Both bona fide and fake faces are of varying quality due to different capture conditions.

Bona Fide Faces. We utilize faces from CASIA-WebFace [65], LFW [26], CelebA [40], SiW-M [38], and FFHQ [30] datasets since the faces therein cover a broad variation in race, age, gender, pose, illumination, expression, resolution, and acquisition conditions.

Adversarial Faces. We craft 6 SOTA adversarial faces from CASIA-WebFace [65] and LFW [26]: FGSM [22], PGD [42], DeepFool [48], AdvFaces [16], GFLM [12], and SemanticAdv [51].

Digital Manipulation. There are four broad types of digital face manipulation: identity swap, expression swap, attribute manipulation, and entirely synthesized faces [13]. We use all clips from FaceForensics++ [53], including identity swap by FaceSwap and DeepFake, and expression swap by Face2Face [57]. We utilize two SOTA models, StarGAN [10] and STGAN [35], to generate attribute manipulated faces in Celeba [40] and FFHQ [30]. We use the pretrained StyleGAN2 model[3] to synthesize 100K fake faces.

Physical Spoofs. We utilize the publicly available *SiW-M* dataset [38], comprising 13 spoof types. Compared with other spoof datasets (Tab. I), SiW-M is diverse in spoof attack types, environmental conditions, and face poses.

Protocol. As is common practice in face recognition literature, bona fides and attacks from CASIA-WebFace [65] are used for training, while bona fides and attacks for LFW [26] are sequestered for testing.

Implementation. *UniFAD* is trained with a constant learning rate of $1e^{-3}$ and batch size of 180. $\mathcal{L}_{UniFAD}$, is minimized using Adam optimizer for $100K$ iterations (see Supp.).

Metrics. Studies on different attack categories provide their own metrics. Following the recommendation from IARPA ODIN program, we report the TDR @ 0.2% FDR[4] and the overall detection accuracy (in Supp.).

B. Comparison with Individual SOTA Detectors

We compare the proposed *UniFAD* to detectors via publicly available codes provided by the authors (see Supp.).

Without Re-training. In Tab. II, we first report the performance of 4 pre-trained SOTA detectors. These baselines were chosen since they report the best detection performance in datasets with largest numbers of attack types in their respective categories (see Tab. I). We find that pre-trained methods indeed excel in their specific attack categories, however, generalization performance across all 3 categories deteriorates catastrophically.

With Re-training. After re-training the 4 SOTA detectors on all 25 attack types, we find that they generalize better

[3] https://github.com/NVlabs/stylegan2

[4] https://www.iarpa.gov/index.php/research-programs/odin

	TDR (%) @ 0.2% FDR	Year	Proposed For	Adv.	Dig. Man.	Phys.	Overall	Time (ms)
w/o Retrain	FaceGuard [15]	2020	Adversarial	**99.91**	22.28	00.58	29.64	01.41
	FFD [13]	2020	Digital Manipulation	09.49	**94.57**	01.25	34.55	11.57
	SSRFCN [14]	2020	Spoofs	00.25	00.76	93.19	22.71	02.22
	MixNet [54]	2020	Spoofs	00.36	09.83	78.21	21.12	12.47
Baselines	FaceGuard [15]	2020	Adversarial	**99.86**	41.56	04.35	56.69	01.41
	FFD [13]	2020	Digital Manipulation	76.06	91.32	87.43	68.25	11.57
	SSRFCN [14]	2020	Spoofs	08.23	27.67	89.19	43.26	02.22
	One-class [20]	2020	Spoofs	04.81	45.96	79.32	39.40	07.92
	MixNet-*UniFAD*	2022	All	82.33	91.59	**94.60**	**90.07**	12.47
Fusion Schemes	Cascade [58]	–	–	88.39	81.98	69.19	77.46	05.16
	Min-score	–	–	03.65	11.08	00.43	07.22	16.14
	Median-score	–	–	10.87	42.33	47.19	39.48	16.12
	Mean-score	–	–	14.53	47.18	61.32	38.23	16.12
	Max-score	–	–	85.32	61.93	56.87	73.89	16.13
	Sum-score	–	–	74.93	58.01	50.34	69.21	16.11
	LightGBM [31]	–	–	76.25	81.28	88.52	85.97	17.92
	Proposed UniFAD	2022	All	92.56	**97.21**	**98.76**	**94.73**	02.59

TABLE II: Detection accuracy (TDR (%) @ 0.2% FDR) on *GrandFake* dataset. Results on fusing FaceGuard [15], FFD [13], and SSRFCN [14] are also reported. We report time taken to detect a single image (on a Nvidia 2080Ti GPU). [Keys: **Best**, **Second best**]

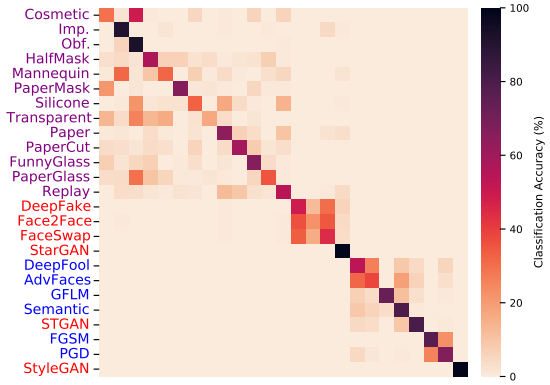

Fig. 4: Confusion matrix representing the classification accuract of *UniFAD* in identifying the 25 attack types. Majority of misclassifications occur within the attack category. Darker values indicate higher accuracy. Overall, *UniFAD* achieves 75.81% and 97.37% classification accuracy in identifying attack types and categories, respectively. Purple, blue, and red denote spoofs, adversarial, and digital manipulation attacks, respectively.

across categories. FaceGuard [15], FFD [13], SSRFCN [14], and One-Class [20] employ a JointCNN for detecting attacks. Unsurprisingly, these defenses perform well on some attack categories, while failing on others. We also modify MixNet, namely *MixNet-UniFAD* such that clusters are assigned via k-means with 4 branches. In contrast to *MixNet-UniFAD*, *UniFAD* (i) employs early shared layers for generic attack cues, and (ii) each branch learns to distinguish between bona fides and specific attack types. MixNet, on the other hand, assigns a bona fide label (0) to attack types outside a respective branch's partition. This negatively impacts network convergence. Overall, we find that *UniFAD* outperforms *MixNet-UniFAD* with TDR 90.07% $\rightarrow$ 94.73% @ 0.2% FDR.

C. Comparison with Fused SOTA Detectors

We also comprehensively evaluate detection performance on fusing SOTA detectors. We utilize three best performing detectors from each attack category, namely FaceGuard [15], FFD [13], and SSRFCN [14]. Inspired by the Viola-Jones object detection [58], we adopt a sequential ensemble technique, namely Cascade [58], where an input probe is passed through each detector sequentially. We also evaluate 5 parallel score fusion rules (min, mean, median, max, and sum) and a SOTA ensemble technique, namely LightGBM [31]. More details are provided in Supp. Indeed, we observe an overhead in detection speed compared to the individual detectors in isolation, however, cascade, max-score fusion and LightGBM [31] can enhance the overall detection performance compared to the individual detectors at the cost of slower inference speed. Since the individual detectors still train with incoherent attack types, we find that proposed *UniFAD* outperforms all the considered fusion schemes.

In Fig. 2(a), we show the performance degradation of LightGBM [31], the best fusing baseline, w.r.t. *UniFAD*. We observe that among 4 clusters, the last 2 have the overall largest degradation. Interestingly, these 2 clusters are the only ones including attack types across different attack categories, learned via our k-mean clustering. In other words, the cross-category attacks types within a branch benefit each other, leading to the largest performance gain over [31]. This further demonstrates the necessity and importance of a unified detection scheme — the more attack types the detector sees, the more likely it would nourish among each other and be able to generalize.

D. Attack Classification

We classify the exact attack type and categories using the method described in Sec. IV-E. In Fig. 4, we find that *UniFAD* can predict the attack type with 75.81% classification accuracy. While predicting the exact type may be challenging, we highlight that majority of the misclassifications occurs within attack's category. Without human intervention, once *UniFAD* is deployed in AFR pipelines, it can predict whether an input image is adversarial, digitally manipulated, or contains spoof artifacts with 97.37% accuracy.

E. Analysis of UniFAD

Ratio of Shared Layers. Our backbone network consists of a 4-layer CNN. In Fig. 5a, we report the detection

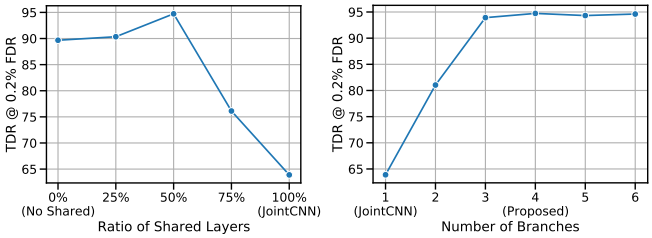

Fig. 5: Detection performance with respect to varying ratio of shared layers (left) and number of branches (right). Our proposed architecture uses 50% shared layers with 4 branches.

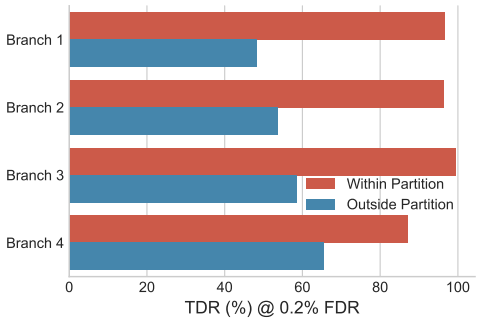

Fig. 6: Detection performance on attack types within and outside a branch's partition. Performance drops on attacks outside partition as they may not have any correlation with within-partition attack types.

performance when we incorporate 0, 1 (25%), 2 (50%), 3 (75%), and 4 (100%) layers for early sharing. We observe a trade-off between detection performance and the number of early layers: too many reduces the effects of learning task-specific features via branching, whereas, less number of shared layers inhibits the network from learning generic features that distinguish any attack from bona fides. We find that an even split results in superior detection performance.

Number of Branches. In Fig. 5b, we vary the number of branches (aux. tasks constructed via kMeans) and report the detection performance. Indeed, increasing the number of branches via additional clusters enhances detection performance. However, the performance saturates after 4 branches. Thus, we choose $T = 4$ due to lower network complexity.

Branch Generalizability. In Fig. 6, scores from the 4 branches are used to compute the detection performance on attack types within respective partitions and those outside a branch's partition (see Fig. 2(b)). Since attack types outside a branch's partition are purportedly incoherent, we see a drop in performance; validating the drawback of JointCNN. We find that the lowest performance branch, Branch 4, also exhibits the best generalization performance across other attack types. This is likely because learning to distinguish bona fides from imperceptible perturbations from FGSM, PGD, and minute synthetic noises from StyleGAN yields a tighter decision boundary which may contribute to better generalization across digital attacks. Anti-spoofing (Branch 1) itself does not directly aid in detecting digital attacks.

Ablation Study. In Tab. III, we conduct a component-wise ablation study over *UniFAD*. We study different partitioning techniques to group the 25 attack types. We employ semantic partitioning, $\mathcal{B}_{Semantic}$ where attack types are clustered

Model	Modules			Overall
	Shared Layers	Branching	kMeans	TDR (%) @ 0.2% FDR
JointCNN	✓			63.89
$\mathcal{B}_{Semantic}$		✓		86.17
$\mathcal{B}_{Random}$		✓		53.95 ± 08.02
$\mathcal{B}_{kMeans}$		✓	✓	89.67
SharedSemantic	✓	✓		92.44
Proposed	✓	✓	✓	**94.73**

TABLE III: Ablation study over components of *UniFAD*. Branching via "$\mathcal{B}_{Semantic}$", "$\mathcal{B}_{Random}$", and "$\mathcal{B}_{kMeans}$" refer to partitioning attack types by their semantic categories, randomly, and kMeans. "SharedSemantic" includes shared layers prior to branching.

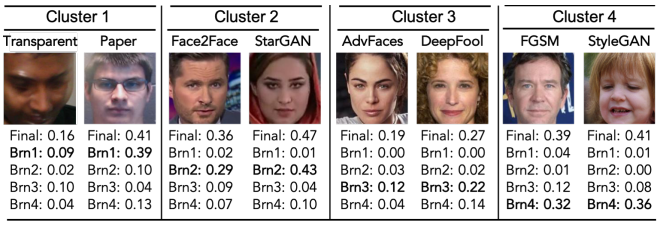

Fig. 7: Example cases where *UniFAD* fails to detect face attacks. Final detection scores along with scores from each of the four branches ($\in [0, 1]$) are given below each image. Scores closer 0 indicate bona fide. Branches responsible for the respective cluster are highlighted in bold.

into the 3 categories. Another technique is to split the 25 attack types into 4 clusters randomly, $\mathcal{B}_{Random}$. We report the mean and standard deviation across 3 trials of random splitting. We also report the performance of clustering via kMeans. We find that both $\mathcal{B}_{Semantic}$ and $\mathcal{B}_{kMeans}$ outperforms JointCNN. Thus, learning separate feature spaces via MTL for disjoint attack types can improve overall detection compared to a JointCNN. We als find that incorporating early shared layers into $\mathcal{B}_{Semantic}$, namely $\mathcal{B}_{SharedSemantic}$, can further improve detection from 86.17% $\rightarrow$ 92.44% TDR @ 0.2% FDR. However, as we observed in Fig. 2, even within semantic categories, some attack types may be incoherent. By automatic construction of auxiliary tasks with k-means clustering and shared representation (Proposed), we can further enhance the detection performance to TDR = 94.73% @ 0.2% FDR.

Failure Cases. Fig. 7 shows a few failure cases. Majority of the failure cases for digital attacks are due to imperceptible perturbations. In contrast, failure to detect spoofs can likely be attributed to the subtle nature of transparent masks, blurring, and illumination changes.

VI. CONCLUSIONS

With new and sophisticated attacks being crafted against AFR systems in both digital and physical spaces, detectors need to be robust across all 3 categories. Poor generalization can be predominantly attributed towards learning incoherent attacks jointly. With a new multi-task learning framework along with k-means augmentation, the proposed *UniFAD* achieved SOTA detection performance (TDR = 94.73% @ 0.2% FDR) on 25 face attacks across 3 categories. *UniFAD* can further identify categories with a 97.37% accuracy. We are exploring whether an attention module can further improve detection.

VII. ACKNOWLEDGMENTS

This work was partially supported by Facebook AI.

REFERENCES

[1] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *BTAS*, 2018.

[2] A. Agarwal, R. Singh, M. Vatsa, and N. K. Ratha. Image transformation based defense against adversarial perturbation on deep learning models. *IEEE TDSC*, 2020.

[3] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, 2019.

[4] V. Asnani, X. Yin, T. Hassner, S. Liu, and X. Liu. Proactive image manipulation detection. In *CVPR*, 2022.

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

[6] J. Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

[7] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *IEEE FG*, pages 612–618, 2017.

[8] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang. Self-supervised learning of adversarial example. In *CVPR*, 2022.

[9] I. Chingovska, A. Anjos, and S. Marcel. On the Effectiveness of Local Binary Patterns in Face Anti-spoofing. In *IEEE BIOSIG*, 2012.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[11] M. Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[12] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi. Fast geometrically-perturbed adversarial faces. In *WACV*, 2019.

[13] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In *CVPR*, 2020.

[14] D. Deb and A. K. Jain. Look locally infer globally: A generalizable face anti-spoofing approach. *IEEE TIFS*, 16:1143–1157, 2020.

[15] D. Deb, X. Liu, and A. K. Jain. Faceguard: A self-supervised defense against adversarial face images. *arXiv:2011.14218*, 2020.

[16] D. Deb, J. Zhang, and A. K. Jain. Advfaces: Adversarial face synthesis. In *IJCB*. IEEE, 2020.

[17] Y. Du, T. Qiao, M. Xu, and N. Zheng. Towards face presentation attack detection based on residual color texture representation. *Security and Communication Networks*, 2021.

[18] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *WACV*, 2022.

[19] 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.

[20] A. George and S. Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE TIFS*, 16:361–375, 2020.

[21] Z. Gong, W. Wang, and W.-S. Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[23] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *ICCV*, 127(6-7):719–742, 2019.

[24] T. Gui, L. Qing, Q. Zhang, J. Ye, H. Yan, Z. Fei, and X. Huang. Constructing multiple tasks for augmentation: Improving neural image classification with k-means features. In *AAAI*, 2020.

[25] K. Hsu, S. Levine, and C. Finn. Unsupervised learning via meta-learning. *ICLR*, 2018.

[26] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild. Technical report, UMass, October 2007.

[27] M. Ibsen, L. J. González-Soler, C. Rathgeb, P. Drozdowski, M. Gomez-Barrero, and C. Busch. Differential anomaly detection for facial images. In *WIFS*. IEEE, 2021.

[28] Y. Jang, T. Zhao, S. Hong, and H. Lee. Adversarial defense via learning to generate diverse attacks. In *ICCV*, 2019.

[29] S. Jia, G. Guo, and Z. Xu. A survey on 3d mask presentation attack detection and countermeasures. *Pattern Recognition*, 98:107032, 2020.

[30] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[31] G. Ke, Q. Meng, T. Finley, W. Chen, W. Ma, Q. Ye, and T. Liu. Highly efficient gradient boosting decision tree. *NeurIPS*, 2017.

[32] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv:1812.08685*, 2018.

[33] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020.

[34] Z. Li, R. Cai, H. Li, K.-Y. Lam, Y. Hu, and A. C. Kot. One-class knowledge distillation face presentation attack detection. *TIFS*, 2022.

[35] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, 2019.

[36] X. Liu and C.-J. Hsieh. Rob-gan: Generator, discriminator, and adversarial attacker. In *CVPR*, 2019.

[37] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, June 2018.

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

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

[40] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015.

[41] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser. Multi-task sequence to sequence learning. *arXiv:1511.06114*, 2015.

[42] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[43] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE TIFS*, 12(7):1713–1723, 2017.

[44] F. V. Massoli, F. Carrara, G. Amato, and F. Falchi. Detection of face recognition adversarial attacks. *CVIU*, page 103103, 2020.

[45] S. Mehta, A. Uberoi, A. Agarwal, M. Vatsa, and R. Singh. Crafting a panoptic face presentation attack detector. In *ICB*. IEEE, 2019.

[46] E. Meyerson and R. Miikkulainen. Pseudo-task augmentation: From deep multitask learning to intratask sharing. In *ICML*, 2018.

[47] 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.

[48] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[49] U. Muhammad, Z. Yu, and J. Komulainen. Self-supervised 2d face presentation attack detection via temporal sequence sampling. *Pattern Recognition Letters*, 2022.

[50] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli. A self-supervised approach for adversarial robustness. In *CVPR*, 2020.

[51] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv preprint arXiv:1906.07927*, 2019.

[52] Y. A. U. Rehman, L. M. Po, and M. Liu. Deep learning for face anti-spoofing: an end-to-end approach. In *IEEE SPA*, 2017.

[53] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Learning to detect manipulated facial images. In *ICCV*, 2019.

[54] N. Sanghvi, S. K. Singh, A. Agarwal, M. Vatsa, and R. Singh. Mixnet for generalized face presentation attack detection. *arXiv:2010.13246*, 2020.

[55] R. Shao, X. Lan, and P. C. Yuen. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing. In *IEEE IJCB*, pages 748–755, 2017.

[56] 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. *IEEE TIFS*, 15:3181–3196, 2020.

[57] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[58] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[59] J. Wang, Z. Zhao, W. Jin, X. Duan, Z. Lei, B. Huai, Y. Wu, and X. He. Cross-domain face presentation attack detection with vocabulary separation and adaptation. In *ACM MM*, 2021.

[60] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019.

[61] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE TIFS*, 10(4):746–761, 2015.

[62] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*. IEEE, 2019.

[63] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu. Model matters, so does data. In *CVPR*, 2019.

[64] X. Yang, D. Yang, Y. Dong, W. Yu, H. Su, and J. Zhu. Delving into the adversarial robustness on face recognition. *arXiv:2007.04118*, 2020.

[65] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014.

[66] 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. *arXiv preprint arXiv:2003.04092*, 2020.