Unified Physical-Digital Face Attack Detection

Hao Fang*[1], **Ajian Liu***[1], **Haocheng Yuan**[2], **Junze Zheng**[2], **Dingheng Zeng**[3], **Yanhong Liu**[3], **Jiankang Deng**[4], **Sergio Escalera**[5], **Xiaoming Liu**[6], **Jun Wan**†[1,2] and **Zhen Lei**[1,7]
[1]MAIS, Institute of Automation of Chinese Academy of Sciences, Beijing, China
[2]Macau University of Science and Technology (MUST), Macau, China
[3]Mashang Consumer Finance Co., Ltd., Chongqing, China
[4]Imperial College London, London, UK
[5]Computer Vision Center (CVC), Barcelona, Catalonia, Spain
[6]Department of Computer Science and Engineering, Michigan State University
[7]CAIR Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences
fanghao21@mails.ucas.ac.cn, {ajian.liu, jun.wan}@ia.ac.cn, zlei@nlpr.ia.ac.cn

Abstract

Face Recognition (FR) systems can suffer from physical (i.e., print photo) and digital (i.e., DeepFake) attacks. However, previous related work rarely considers both situations at the same time. This implies the deployment of multiple models and thus more computational burden. The main reasons for this lack of an integrated model are caused by two factors: (1) The lack of a dataset including both physical and digital attacks which the same ID covers the real face and all attack types; (2) Given the large intra-class variance between these two attacks, it is difficult to learn a compact feature space to detect both attacks simultaneously. To address these issues, we collect a Unified physical-digital Attack dataset, called **UniAttackData**. The dataset consists of $1,800$ participations of 2 and 12 physical and digital attacks, respectively, resulting in a total of $28,706$ videos. Then, we propose a Unified Attack Detection framework based on Vision-Language Models (VLMs), namely **UniAttackDetection**, which includes three main modules: the Teacher-Student Prompts (TSP) module, focused on acquiring unified and specific knowledge respectively; the Unified Knowledge Mining (UKM) module, designed to capture a comprehensive feature space; and the Sample-Level Prompt Interaction (SLPI) module, aimed at grasping sample-level semantics. These three modules seamlessly form a robust unified attack detection framework. Extensive experiments on UniAttackData and three other datasets demonstrate the superiority of our approach for unified face attack detection. Dataset link: *https://sites.google.com/view/face-anti-spoofing-challenge/dataset-download/uniattackdatacvpr2024*

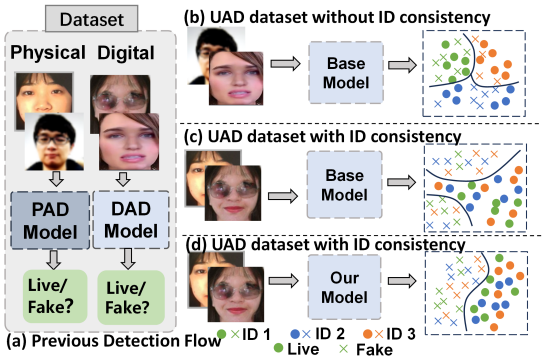

Figure 1: **Paradigm Comparison.** (a) Prior approaches necessitate the separate training and deployment of PAD and DAD models, demanding significant computational resources and inference time. (b) UAD dataset without ID consistency introduces the risk of the algorithm learning noise related to ID. (c) The base model encounters challenges in acquiring a compact feature space when confronted with the UAD dataset. (d) Our algorithm learns compact feature space and clear class boundaries.

1 Introduction

Face recognition (FR) system [Wan *et al.*, 2023] has been widely used in face unlocking, face payment, and video surveillance. It can face a diverse set of attacks: (1) Physical attacks (PAs), i.e. print-attack, replay-attack and mask-attack [Liu *et al.*, 2022b; Fang *et al.*, 2023]; and (2) Digital attacks (DAs), i.e. Face2Face [Thies *et al.*, 2016], FaceSwap [Ding *et al.*, 2020], Deepfakes, and NeuralTextures [Thies *et al.*, 2019]. Both Physical Attack Detection (PAD) [Liu *et al.*, 2018; Liu *et al.*, 2023b; Liu and Liang, 2022; Liu *et al.*, 2021b; Liu *et al.*, 2022a] and Digital Attack Detection (DAD) [Dang *et al.*, 2020; Zhao *et al.*, 2021] are still being studied by related works as two independent tasks. As shown in Fig. 1(a), this will lead to the training of both physical and digital attack detection models and their deployment, requiring large computing resources.

Two main reasons hinder the unification of detecting

physical-digital attacks: (1) **Lack of the physical-digital dataset.** Although two datasets (namely GrandFake [Deb *et al.*, 2023] and JFSFDB [Yu *et al.*, 2022]) are proposed to tackle this problem, they only simply merge PAs and DAs datasets. They cannot guarantee each ID covers the real face and all attack types, which would lead the model to learn the liveness-irrelevant signals such as face ID, domain-specific information, and background. Therefore, the lack of a physical-digital dataset with ID consistency limits the research on unified detection algorithms. **(2) Distinct intra-class variances.** Although both physical and digital attacks are classified as fake in the solution space of the face anti-spoofing problem, the vast differences between these two types of attacks increase the intra-class distances. Most of the existing attack detection methods [Liu *et al.*, 2018; Yu *et al.*, 2020b] are proposed for a specific attack, thus overfitting to a certain attack and failing to learn a compact feature space to detect both attacks simultaneously.

To solve the first issue, we collect and release a dataset combining physical and digital attacks, named UniAttackData, which contains 2 physical and 12 digital attacks for each of $1,800$ subjects, with a total of $28,706$ videos. Compared to GrandFake and JFSFDB datasets, the UniAttackData provides four advantages: (1) **The most complete attack types for each face ID.** Our dataset constructs physical-digital attacks for each face ID, rather than simply merging existing PAs and DAs datasets. (2) **The most advanced forgery method.** For digital attacks, we select the most advanced digital forgery methods in the past three years. For physical attacks, we consider printing attacks and video replay attacks in various environments and physical media. (3) **The most amount of images.** The proposed dataset is the largest one in terms of the number of images, which is more than $3.2\times$ boosted compared to the previous face anti-spoofing dataset like GrandFake. (4) **The most diverse evaluation protocols.** Beyond the within-modal evaluation protocols, we also provide the cross-attack evaluation protocols in our dataset, in which algorithms trained in one attack category are evaluated in other attack categories. For the second issue, we propose a unified detection framework, named UniAttackDetection, which has three main modules: **(1) Introducing textual information to depict visual concepts.** Learning joint and category-specific knowledge features by constructing teacher prompts and student prompts, respectively. **(2) Enhanced learning of complete feature space.** We propose unified knowledge mining loss to learn complete and tight feature spaces using text-to-text optimization. **(3) Sample-level visual text interaction.** We map learnable text tokens to the visual embedding space and understand sample-level semantics through multi-modal prompt learning. To sum up, the main contributions of this paper are summarized as follows:

- We propose a dataset that combines physical and digital attacks called UniAttackData. To the best of our knowledge, this is the first unified attack dataset with guaranteed ID consistency.
- We propose a unified attack detection framework based on the vision language model, named UniAttackDetection. It adaptively learns a tight and complete feature space with the help of extensive visual concepts and rich semantic information in text prompts.
- We conducted extensive experiments on UniAttackData and three existing attack datasets. The results show the superiority of our approach in the task of unified face attack detection.

2 Related Work

2.1 Physical and Digital Attack Detection

PAD technology aims to identify whether the face collected by sensors comes from a live face or is a presentation attack. The most advanced algorithms [Liu *et al.*, 2018; Yu *et al.*, 2020a; Liu *et al.*, 2023a] are based on facial depth to determine authenticity. However, their performance severely deteriorates in unknown domains. At present, the generalization of FAS algorithm is increasingly becoming an important evaluation indicator. Digital attack detection aims to distinguish authentic facial images from digitally manipulated facial artifacts. In initial studies [Rossler *et al.*, 2019], image classification backbones were employed to extract features from isolated facial images, facilitating binary classification. With the increasing visual realism of forged faces, recent efforts [Zhao *et al.*, 2021] focus on identifying more reliable forgery patterns, including noise statistics, local textures, and frequency information. Several other works [Zi *et al.*, 2020] have integrated attention mechanisms to bolster the differentiation between live and forged images. Similarly, these methods still have the problem of insufficient generalization ability, that is, their performance on cross-data is weak.

2.2 Physical-Digital Attack Detection

In a recently published paper, Yu et al [Yu *et al.*, 2022] established the first joint benchmark for face fraud and forgery detection and combined visual appearance and physiological rPPG signals to alleviate the generalization problem. Debayan Deb [Deb *et al.*, 2023] et al. classified all 25 attack types documented in the literature and proposed a method to distinguish between real identities and various attacks using a multi-task learning framework and k-means enhancement techniques. However, none of these works have investigated unified attack detection based on ID consistency. In this paper, we present the first unified attack dataset with ID consistency and a robust unified attack detection algorithm.

3 UniAttackData Dataset

3.1 Acquisition Detail

As depicted in Tab. 1, our UniAttackData is an extension of CASIA-SURF CeFA [Liu *et al.*, 2021a] through digital forgery, which includes $1,800$ subjects from 3 ethnicities (e.g., African, East Asian and Central Asian), and 2 types of physical attacks (e.g., Print and Replay). For each subject, we forge 12 types of digital attacks as follows: (1) Pairing each subject video with a video of others as the reference video. (2) For digital forging, we employ the latest 6 digital editing algorithms and 6 adversarial algorithms on each live video, as detailed in Tab. 1. (3) To selectively forge only the facial

Dataset	Attack Type (each ID)	# Datasets / Data	# ID	Physical Attacks		Digital Attacks		
				Dataset Name	No.	# Categories	Methods	No.
GrandFake	Incomplete	6 sets: 789412 (I) (Live: 341738, Fake: 447674)	96817	SiW-M [Liu *et al.*, 2019b]	128112 (I)	Adv (6)	FGSM [Goodfellow *et al.*, 2014]	19739 (I)
							PGD [Madry *et al.*, 2019]	19739 (I)
							DeepFool [Moosavi-Dezfooli *et al.*, 2016]	19739 (I)
							AdvFaces [Deb *et al.*, 2019]	19739 (I)
							GFLM [Dabouei *et al.*, 2019]	17946 (I)
							SemanticAdv [Qiu *et al.*, 2020]	19739 (I)
						DeepFake (6)	FaceSwap [Kowalski, 2018]	14492 (I)
							Deepfake [Korshunov and Marcel, 2018]	18165 (I)
							Face2Face [Thies *et al.*, 2016]	18204 (I)
							StarGAN [Choi *et al.*, 2018]	45473 (I)
							STGAN [Liu *et al.*, 2019a]	29983 (I)
							StyleGAN2 [Karras *et al.*, 2020]	76604 (I)
JFSFDB	Incomplete	9 sets: 27172 (V) (Live: 5650, Fake: 21522)	356	SiW [Liu *et al.*, 2018]	3173 (V)	DeepFake (4)	Face2Face [Thies *et al.*, 2016]	1000 (V)
				3DMAD	85 (V)		FaceSwap [Kowalski, 2018]	1000 (V)
				HKBU [Liu *et al.*, 2016]	588 (V)		NeuralTextures [Thies *et al.*, 2019]	1000 (V)
				MSU [Wen *et al.*, 2015]	210 (V)			
				3DMask [Yu *et al.*, 2020a]	864 (V)		Deepfake [Korshunov and Marcel, 2018]	10752 (V)
				ROSE [Li *et al.*, 2018]	2850 (V)			
UniAttackData (Ours)	**Complete**	**1 set:** **28706 (V)** **(Live: 1800, Fake: 26906)**	**1800**	CASIA-SURF CeFA [Liu *et al.*, 2021a]	5400 (V)	Adv (6)	advdrop [Duan *et al.*, 2021]	1706 (V)
							alma [Rony *et al.*, 2021]	1800 (V)
							demiguise [Wang *et al.*, 2021c]	1800 (V)
							fgtm [Zou *et al.*, 2022]	1800 (V)
							ila_da [Yan *et al.*, 2022]	1800 (V)
							ssah [Luo *et al.*, 2022]	1800 (V)
						DeepFake (6)	FaceDancer [Rosberg *et al.*, 2023]	1800 (V)
							InsightFace [Heusch *et al.*, 2020]	1800 (V)
							SimSwap [Chen *et al.*, 2020]	1800 (V)
							SAFA [Wang *et al.*, 2021a]	1800 (V)
							DaGAN [Hong *et al.*, 2022]	1800 (V)
							OneShotTH [Wang *et al.*, 2021b]	1800 (V)
						summary	**12**	**21506 (V)**

Table 1: Comparison of our multimodal facial attack datasets with Grandfake and JFSFDB. Our UniAttackData dataset covers all attack types, containing advanced forgery methods from 2020 to 2023, using the same ID as the existing dataset CASIA-SURF CeFA. Images in UniAttackData of the amount of $2,526,432$ are 3 times more than GrandFake. [Keys: I=Image, V=Video]

region while preserving the background, our process initiates face detection on every frame of the video. Subsequently, we digitally manipulate solely the facial area and seamlessly integrate the background onto the altered face. In overview, as depicted in Fig. 2, our UniAttackData contains live faces from three ethnicities, two types of print attacks in distinct environments, one playback attack, and 6 digital editing attacks and 6 adversarial attacks.

3.2 Advantages of UniAttackData

In comparison to previously datasets, such as GrandFake [Deb *et al.*, 2023] and JFSFDB [Yu *et al.*, 2022], the UniAttackData has the following advantages: **Advantage 1.** Each ID contains a complete set of attack types. Because we conduct comprehensive physical and digital attacks on each live face, each ID contains a complete set of attack types. On the other hand, GrandFake and JFSFDB merge existing physical and digital attack datasets through integration, resulting in incomplete attack types for each ID, which easily leads to overfitting of the model to identity information. **Advantage 2.** Incorporation of the most advanced and comprehensive attack methods. Our UniAttackData contains 6 editing attacks and 6 adversarial attacks, utilizing algorithms developed from 2020 onwards for digital attack sample production. In contrast, JFSFDB only contains four types of digital editing attacks, and GrandFake uses algorithms predating 2020 for digital forgery. **Advantage 3.** The largest joint physical and digital attack dataset. As illustrated in Tab. 1, UniAttackData stands out as the dataset with the highest video count. Statistically, it comprises a total of $28,706$ videos, encompassing $1,800$ videos featuring live faces, $5,400$ videos showcasing physical attacks, and $21,506$ videos with digital attacks.

Protocol	Class	Types				# Total
		# Live	# Phys	# Adv	#Digital	
P1	train	3000	1800	1800	1800	8400
	eval	1500	900	1800	1800	6000
	test	4500	2700	7106	7200	21506
P2.1	train	3000	0	9000	9000	21000
	eval	1500	0	1706	1800	5006
	test	4500	5400	0	0	9900
P2.2	train	3000	2700	0	0	5700
	eval	1500	2700	0	0	4200
	test	4500	0	10706	10800	26006

Table 2: Amount of train/eval/test images of different types under three different protocols: P1, P2.1, and P2.2.

3.3 Protocols and Statistics

We define two protocols for UniAttackData. (1) Protocol 1 aims to evaluate under the unified attack detection task. As shown in Tab. 2, the training, validation, and test sets contain live faces and all attacks. (2) Protocol 2 evaluates the generalization to "unseen" attack types. The large differences and unpredictability between physical-digital attacks pose a challenge to the portability of the algorithms. In this paper, we use the "leave-one-type-out testing" approach to divide Protocol 2 into two sub-protocols, where the test set for each self-subprotocol is an unseen attack type. As shown in Tab 2, the test set of protocol 2.1 contains only physical attacks that have not been seen in the training and development sets, and the test set of protocol 2.2 contains only digital attacks that have not been seen in the training and development set.

4 Proposed Method

4.1 Overview

Our UniAttackDetection framework is based on CLIP [Radford *et al.*, 2021]. As illustrated in Fig. 3, the backbone net-

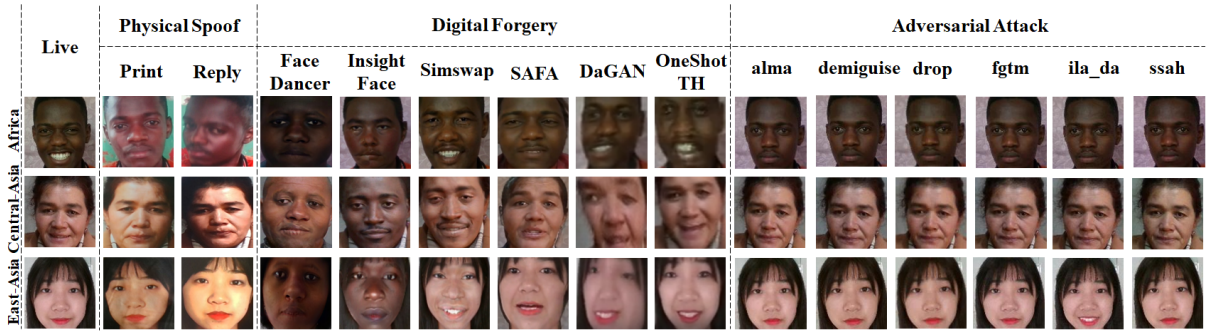

Figure 2: UniAttackData Dataset examples of all attack types corresponding to the same face ID. From top to bottom, they are Africans, Central Asians, and East Asians, respectively. The attack type of each sample is marked at the top.

work incorporates three key components: (1) The Teacher-Student Prompt (TSP) module, designed to extract unified and specific knowledge features by encoding both teacher prompts and student prompts, respectively; (2) The Unified Knowledge Mining (UKM) module, leveraging the Unified Knowledge Mining (UKM) loss to steer the student features, and facilitating the exploration of a tightly complete feature space. (3) The Sample-Level Prompt Interaction (SLPI) module, which maps learnable student prompts to visual modality prompts through the interaction projector. Subsequently, the language modality prompts obtained are linked to the original visual modality embeddings, and channel interactions take place during multi-modal prompt learning to extract sample-level semantics.

4.2 Teacher-Student Prompt

Teacher Prompt. Traditional FAS methods [Wang *et al.*, 2022] focus on extracting features commonly present in each category of images, such as facial contours and shapes. These features are then utilized to learn generic classification knowledge. We introduce this concept into the natural language field by utilizing template prompts to guide the extraction of unified knowledge in physical-digital attacks, which is based on the following fact: The textual descriptions for each category can be effectively represented using a consistent hard template, implying a shared semantic feature space among them. Specifically, given a set of categories consisting of unified class names $unified_c$, $c \in \{liveface, spoofface\}$, the teacher prompt group is then obtained by interconnecting the class name and the manually designed template, i.e., t_c = *a photo of a* $\{unified_c\}$. This group of prompts is passed through CLIP's text encoder $G_T(\cdot)$ to compute the unified knowledge features of the text $f_{tc} = G_T(t_c)$, $f_{tc} \in \mathbb{R}^{c_u \times d}$, where d is the dimension of CLIP. In this paper, we design multi-group artificial templates to fix discrete teacher anchors in the complete feature space. The multi-group teacher prompts are accumulated over the group dimension g and finally encoded as $f_{tc}^g \in \mathbb{R}^{c_u \times g \times d}$.

Student Prompt. In contrast to conventional single-attack detection tasks, the Unified Attack Detection (UAD) task presents substantial gaps within the same classes. To mitigate this problem, we construct a group of student prompts. Student prompts are learned by minimizing the classification error on a training set consisting of specific classes, and thus student prompts possess strong learning abilities in specific classes. Specifically, we prefix each specific class $specific_c$, $c \in \{realface, digitalattack, physicalattack\}$ with a series of learnable vectors $p_n \in \mathbb{R}^d, n \in 1, 2, 3, ..., N$ to get the student prompt. Then the student feature $f_{sc} = G_T(s_c)$ for learning specific knowledge is obtained, where $f_{sc} \in \mathbb{R}^{c_s \times d}$. To use the learned continuous tight features for binary classification, a lightweight head $\mathcal{H} : \mathbb{R}^{c_s \times d} \rightarrow \mathbb{R}^{c_u \times d}$ is adopted to map specific knowledge to a generic classification space.

4.3 Unified Knowledge Mining

To enhance the model's exploration of the complete feature space, we design the Unified Knowledge Mining (UKM) loss as follows:

$$\mathcal{L}_{UKM} = \frac{1}{C} \sum_{g=0}^{G-1} \sum_{c=0}^{C-1} \left(1 - \frac{f'_{sc} f_{tc}^g}{\|f'_{sc}\| \|f_{tc}^g\|} \right) \quad (1)$$

This loss ensures the student features are sufficiently close to each teacher anchor in the unified knowledge feature space, yielding the following advantages: (1) The student prompt is no longer over-fitted with a specific knowledge distribution through the guidance of teacher anchors, and thus gains strong generalization capabilities. (2) Taking the learnable prompt as a carrier, the unlearnable unified feature space is fully mined to learn the tight and complete feature space.

Not only that, VLMs pre-trained on large-scale image-text pairs encapsulate abundant statistical common knowledge, which leads to the inclusion of potential semantic associations among text features extracted based on these models. Therefore, to capture the correlation between unified and specific knowledge features, we employ a fusion block to extract fusion features for multiple groups of teacher prompts and student prompts. Specifically, we concatenate multiple groups of teacher features f_{tc}^g and student features into complete features $f_{com} \in \mathbb{R}^{c_u \times (g+1) \times d}$ in the dimension of g. The fusion block consists of a self-attentive encoder $Att(\cdot)$ and a Multi-Layer Perceptron (MLP) $MLP_{fusion}(\cdot)$, where the self-attention encoder is given by the following equation:

$$Att(X) = \frac{softmax(W_Q X)(W_K X)^T}{\sqrt{d_k}} (W_V X) \quad (2)$$

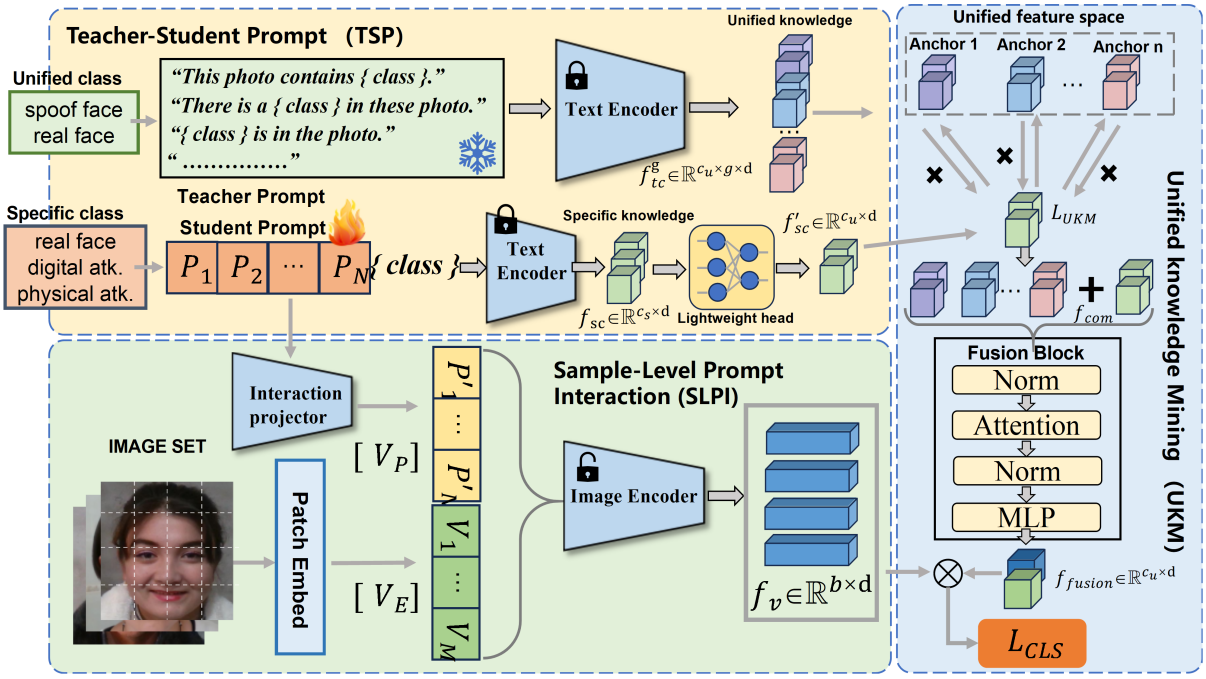

Figure 3: Our proposed UniAttackDetection architecture. The TSP module extracts unified and specific knowledge by constructing multiple groups of teacher prompts and learnable student prompts. The UKM module oversees the learning process by employing the unified knowledge mining loss, thereby enabling the model to acquire comprehensive insights across the entire feature space. The SLPI module maps the student prompts to the visual embedding space, allowing multi-modal prompt learning by making the student prompt learn sample-level semantics while allowing visual feature extraction to be guided by text.

Where W_Q, W_K, and W_V are the weight matrix of the query, key, and value, respectively, and d_k is the dimension of the model. Finally, the modulated fusion knowledge feature f_{fusion} is obtained by:

$$f_{fusion} = MLP_{fusion}(Att(f_{com})) \tag{3}$$

4.4 Sample Level Prompt Interaction

In recent FAS work [Zhou *et al.*, 2023], learned sample-level semantics of visual pictures contribute to enhancing model robustness. Therefore, in VLMs, we deem it essential to establish a multi-modal interaction between picture instance perception features and student prompts. Student prompts can dynamically understand the sample-level semantics in visual pictures and thus learn comprehensive adaptive features. In turn, the visual modality will be guided by the shared prompts to deeply mine the category-related picture features. Inspired by MaPLe [Khattak *et al.*, 2023], we design an interaction projector $p = Linear(\cdot)$. This interaction projector is implemented as a linear layer that maps the dimensions d_p of student prompt into the d_v of visual embeddings. CLIP's text encoder takes the learnable token $[p_1][p_2][p_3]\dots[p_n]$ projected into word embeddings $[P] \in \mathbb{R}^{n\times d_p}$. The interaction projector then acts as a bridge to map the student prompt dimension word embeddings $[P]$ into visual dimension embeddings $[V_P]$, where the projector is expressed by a linear layer. The overall formulation is as follows:

$$[V_P] = Linear([P]), [V] \in \mathbb{R}^{n\times d_v} \tag{4}$$

At the same time, the input image is encoded by the CLIP's patch embed layer as patch embeddings $[V_E] \in \in \mathbb{R}^{m\times d_v}$, where m is the patch dimension of the CLIP. Then, $[V_E]$ and $[V_P]$ are concatenated as the final visual embedding $V = [V_E, V_P]$, where $[\cdot,\cdot]$ refers to the concatenation operation. Finally, the visual embedding is passed through the CLIP image encoder G_V to get the visual features $f_v = G_v(V)$. During the training process, we compute the cosine similarity between the image features and the labeled features to get the prediction probability of the image for each category and finally use the cross-entropy loss L_{CLS} for binary class supervision. Together with the Unified Knowledge Mining (UKM) loss, the final objective is defined as:

$$L_{Total} = L_{CLS} + \lambda \cdot L_{UKM} \tag{5}$$

where λ is a hyper-parameter to trade-off between two losses.

5 Experiments

5.1 Experimental Setting

Datasets. To evaluate the performance of the proposed method and existing approaches, we employ four datasets for face forgery detection, i.e., Our proposed UniAttackData, FaceForensics++ (FF++) [Rossler *et al.*, 2019], OULU-NPU [Boulkenafet *et al.*, 2017] and JFSFDB [Yu *et al.*, 2022]. Our approach first demonstrates the superiority of our method on UniAttackData. Subsequently, we conducted sufficient experiments on other UAD datasets, such as the OULU-FF data protocol composed of OULU-NPU and FF++, and the JFSFDB dataset, to further demonstrate the generalization capability of our approach. Finally, our ablation study demonstrates the effectiveness of each component of our approach.

Evaluation metrics. For a comprehensive measure of the algorithm's UAD performance, we adopt the common metrics used in both physical forgery detection and digital forgery detection work. All experiments were conducted using average classification error rate ACER, overall detection accuracy ACC, the area under the curve (AUC), and equivalent error rate (EER) for performance evaluation. ACER and ACC on each test set are determined by the performance thresholds on the development set.

Implementation Details. To better demonstrate the effectiveness of our method, we consider a series of existing competitors in the field of face anti-spoofing and base network backbone. The considered comparison methods are the ResNet50, ViT-B/16, FFD [Dang *et al.*, 2020], CDCN [Yu *et al.*, 2020b] and the Auxiliary(Depth) [Liu *et al.*, 2018]. In addition, we incorporated our baseline approach into the comparative analysis. Notably, the student prompts for the baseline method were constructed through the unified class. In this way, we validate the superiority of our overall framework and the effectiveness of learning a continuous and compact specific class space.

5.2 Experiments on Proposed UniAttackData

Prot.	Method	ACER(%)↓	ACC(%)↑	AUC(%)↑	EER(%)↓
1	ResNet50	1.35	98.83	99.79	1.18
	VIT-B/16	5.92	92.29	97.00	9.14
	Auxiliary	1.13	98.68	99.82	1.23
	CDCN	1.40	98.57	99.52	1.42
	FFD	2.01	97.97	99.57	2.01
	Baseline(Our)	0.79	99.44	99.96	0.53
	UniAttackDetection(Our)	**0.52**	**99.45**	**99.96**	**0.53**
2	ResNet50	34.60±5.31	53.69±6.39	87.89±6.11	19.48±9.10
	VIT-B/16	33.69±9.33	52.43±25.88	83.77±2.35	25.94± 0.88
	Auxiliary	42.98±6.77	37.71±26.45	76.27±12.06	32.66±7.91
	CDCN	34.33±0.66	53.10±12.70	77.46±17.56	29.17±14.47
	FFD	44.20±1.32	40.43±14.88	80.97±2.86	26.18±2.77
	Baseline(Our)	28.90± 10.85	58.28± 26.37	89.48± 4.94	19.04± 5.81
	UniAttackDetection(Our)	**22.42± 10.57**	**67.35± 23.22**	**91.97± 4.55**	**15.72± 3.08**

Table 3: The results of intra-testing on two protocols of UniAttackData, where the performance of Protocol 2 quantified as the mean±std measure derived from Protocol 2.1 and Protocol 2.2.

Experiments on Protocol 1. In protocol 1 of UniAttackData, the data distribution is relatively similar across sets. The training, development, and test sets all contain both physical and digital attacks. This protocol is suitable for evaluating the performance of the algorithms in UAD tasks. We present the performance results for commonly used backbone networks, networks for physical attack detection, and networks for digital attack detection. As shown in Tab 3, our algorithm outperforms others in all four metrics—ACER, ACC, AUC, and EER. This proves the superiority of our method for unified attack detection.

Experiments on Protocol 2. In protocol 2 of UniAttackData, the test set comprises attack forms that are "unseen" in the training or validation sets, aiming to assess the algorithm's generalizability. As shown in Tab 3, an interesting observation emerges: the performance of single-attack detection networks FFD, CDCN, and Auxiliary largely degrades on Protocol 2, and almost all of their four metrics are lower than classic backbone ViT and ResNet, which exposes the flaw that the single-attack detector is over-fitting specific attacks. On the other hand, our method ranks first in performance on all metrics, which proves that our method not only learns unified knowledge of the UAD task but also captures a compact and continuous feature space for attack categories.

5.3 Experiments on Other UAD Datasets

Data.	Method	ACER(%)↓	ACC(%)↑	AUC(%)↑	EER(%)↓
JFS-FDB	ResNet50	7.70	90.43	98.04	6.71
	VIT-B/16	8.75	90.11	98.16	7.54
	Auxiliary	11.16	87.40	97.39	9.16
	CDCN	12.31	86.18	95.93	10.29
	FFD	9.86	89.41	95.48	9.98
	Baseline(Our)	2.84	96.93	99.54	2.90
	UniAttackDetection(Our)	**1.66**	**98.23**	**99.74**	**1.78**
OULU-FF	ResNet50	7.45	92.64	96.92	7.60
	VIT-B/16	9.95	90.44	97.30	9.83
	Auxiliary	16.43	83.68	92.98	16.80
	CDCN	17.42	82.56	92.37	17.35
	FFD	19.13	80.32	88.90	19.00
	Baseline(Our)	2.22	97.52	99.60	2.22
	UniAttackDetection(Our)	**1.63**	**98.00**	**99.81**	**1.80**

Table 4: The results of Unified attack data protocol JFSFDB and OULU-FF.

To verify the superiority of our proposed dataset and algorithm in more detail. We have done extensive experiments on other UAD datasets. As shown in Tab 4, we obtained the best performance for all metrics on the previously proposed JFSFDB dataset. Furthermore, we combine the training, validation, and testing sets provided by FF++ [Rossler *et al.*, 2019] and OULU-NPU [Boulkenafet *et al.*, 2017] to form a unified attack data protocol named OULU-FF. Our algorithm also outperformed competing methods on the OULU-FF protocol. This validates the ability of our algorithm to detect unified attacks. Notably, we observed a significant decline in the performance of all evaluated methods on both joint attack datasets in comparison to our proposed dataset. This finding underscores the challenge algorithms face in discerning spoofing traces without ID consistency.

5.4 Ablation Experiment

Method	ACER(%)↓	ACC(%)↑	AUC(%)↑	EER(%)↓
UniAttackDetection w/o s_c	1.76	96.95	99.90	2.03
UniAttackDetection w/o t_c	0.83	99.40	99.95	0.60
UniAttackDetection w/o L_{UKM}	0.71	99.25	99.95	0.71
UniAttackDetection w/o V_P	2.56	98.70	99.90	0.93
UniAttackDetection(Our)	**0.52**	**99.45**	**99.95**	**0.53**

Table 5: The ablation study of different Components. The evaluation protocol is P1.

Effectiveness of Different Components. To verify the superiority of our UniAttackDetection as well as the contributions of each component, multiple incomplete models are built up by controlling different variables. All results are measured in the same manner, as shown in Tab 5. First, to verify the validity of student prompts and teacher prompts, we conducted experiments without student prompts or teacher prompts, respectively. The experiments showed that both unified and specific knowledge play a positive role in the UAD

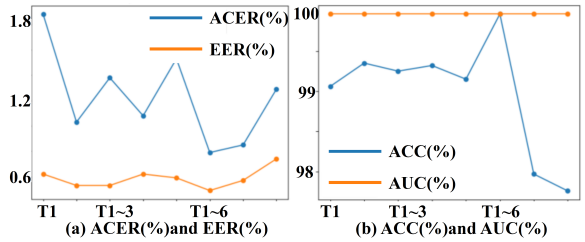

Figure 4: Ablation experiments of the selection of teacher prompts. The horizontal coordinates indicate which teacher prompts from Tab 6 were selected. For example, T1∼6 indicates the selection of the first six prompts from Tab 6.

task. In addition, to validate the effectiveness of the Unified Knowledge Mining (UKM) module, we conducted experiments with UniAttackDetection $w/o\ L_{UKM}$. Specifically, we directly let the unified knowledge features and the specific knowledge features be connected as the final text features. The experiments demonstrate that the UKM module helps the model learn unified knowledge and enhances the generalization to different attacks. Then, to demonstrate the importance of the Sample-Level Prompt Interaction (SPLI) module, we conducted experiments without adding visual prompts. The quantitative results show that multi-modal prompt learning facilitates improved learning of instance-level features.

Effectiveness of Teacher Prompts. To validate the effect of teacher prompts on experimental performance, we incrementally introduced new templates into the teacher prompts group. Tab 6 shows the specific language descriptions covering the real and spoof categories. Fig 4 depicts the variation in experimental performance as we systematically increased the number of teacher prompt groups from one to eight. Examination of the figure reveals that, with the exception of the AUC (Area Under the Curve) metric, there is a discernible trend where the performance across the other three metrics initially improves with the addition of more prompts, only to subsequently decline. All four experimental metrics obtain the best results when six templates are used as teacher prompts. This proves that learning the complete feature space through multiple sets of teacher anchors is beneficial for the UAD task, but too many teacher anchors may instead hinder the adaptive learning ability of student prompts.

Prompt No.	Teacher Prompts (Templates)
T1	This photo contains {real face}/ {spoof face}.
T2	There is a {real face}/ {spoof face} in this photo.
T3	{real face}/ {spoof face} is in this photo.
T4	A photo of a {real face}/ {spoof face}.
T5	This is an example of a {real face}/ {spoof face}.
T6	This is how a {real face}/ {spoof face} looks like.
T7	This is an image of {real face}/ {spoof face}.
T8	The picture is a {real face}/ {spoof face}.

Table 6: The multiple groups of manual templates used are fixed during training.

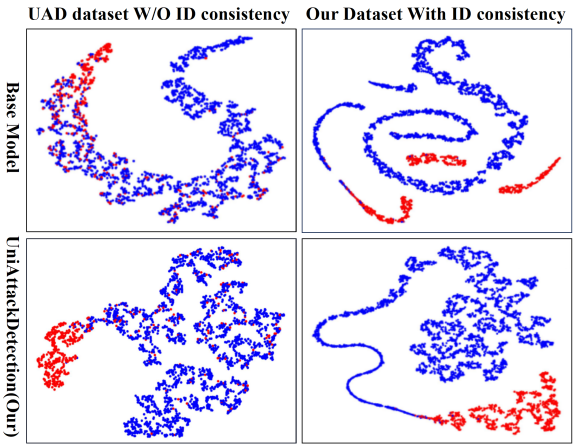

Figure 5: Feature distribution comparison on UAD data protocol (OULU-FF) and UniAttackData using t-SNE. Different colors denote features from different classes.

5.5 Visualization Analysis

Here, we visualize the distribution of features learned by the baseline model ResNet50 and our method on UniAttackData's P1 and the previously mentioned UAD data protocol, respectively. As depicted in Fig 5, our study reveals two significant findings: (1) ID Consistency Impact. Both the base model ResNet50 and our proposed method exhibit limitations in classifying datasets lacking ID consistency assurance. The feature space displays a mixing of lives and attacks across various regions. In contrast, on the proposed dataset, both methods distinctly segregate lives from attacks. This underscores the critical role of ID consistency, enabling the model to concentrate on acquiring deception features rather than being influenced by ID-related noise features. (2) Enhanced Category Feature Space. In comparison to the baseline model, our approach demonstrates the capacity to learn a closely connected category feature space. This highlights our proficiency in acquiring category-specific knowledge. Moreover, our method successfully establishes a distinct category boundary on both datasets, affirming its effectiveness in extracting unified knowledge.

6 Conclusion

In this paper, we proposed a dataset that combines physical and digital attacks, called UniAttackData. This is the first unified attack dataset with guaranteed ID consistency. In addition, we proposed a unified detection framework based on CLIP, namely UniAttackDetection. The method introduces language information into the UAD task and substantially improves the performance of unified attack detection through multi-modal prompt learning. Finally, we conduct comprehensive experiments on UniAttackData and three other datasets to verify the importance of the datasets for the attack detection task and the effectiveness of the proposed method.

Acknowledgments

This work was supported by the National Key Research and Development Plan under Grant 2021YFF0602105, Beijing Natural Science Foundation JQ23016, the Science and Technology Development Fund of Macau Project 0123/2022/A3, and 0070/2020/AMJ, CCF-Zhipu AI Large Model OF 202219 and the Chinese National Natural Science Foundation Project U23B2054, 62276254, and the InnoHK program.

Contribution Statement

Hao Fang and Ajian Liu are equally contributed in this paper. Jun Wan is the Corresponding Author of the paper.

References

[Boulkenafet *et al.*, 2017] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FG*, pages 612–618, 2017.

[Chen *et al.*, 2020] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, pages 2003–2011, 2020.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2018.

[Dabouei *et al.*, 2019] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Fast geometrically-perturbed adversarial faces. In *WACV*, pages 1979–1988. IEEE, 2019.

[Dang *et al.*, 2020] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020.

[Deb *et al.*, 2019] Debayan Deb, Jianbang Zhang, and Anil K. Jain. Advfaces: Adversarial face synthesis, 2019.

[Deb *et al.*, 2023] Debayan Deb, Xiaoming Liu, and Anil K Jain. Unified detection of digital and physical face attacks. In *FG*, pages 1–8. IEEE, 2023.

[Ding *et al.*, 2020] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *EURASIP JIS*, pages 1–12, 2020.

[Duan *et al.*, 2021] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *ICCV*, pages 7506–7515, 2021.

[Fang *et al.*, 2023] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *TIFS*, 2023.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[Heusch *et al.*, 2020] G. Heusch, A. George, D. Geissbuhler, Z. Mostaani, and S. Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *TBIOM*, pages 1–1, 2020.

[Hong *et al.*, 2022] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023.

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

[Kowalski, 2018] Marek Kowalski. Faceswap github. *Faceswap github",[online] Available: https://github.com/MarekKowalski/FaceSwap*, 2018.

[Li *et al.*, 2018] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *TIFS*, 2018.

[Liu and Liang, 2022] Ajian Liu and Yanyan Liang. Mavit: Modality-agnostic vision transformers for face anti-spoofing. In *IJCAI*, pages 1180–1186, 2022.

[Liu *et al.*, 2016] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*, pages 85–100. Springer, 2016.

[Liu *et al.*, 2018] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018.

[Liu *et al.*, 2019a] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, pages 3673–3682, 2019.

[Liu *et al.*, 2019b] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, 2019.

[Liu *et al.*, 2021a] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *WACV*, pages 1179–1187, 2021.

[Liu *et al.*, 2021b] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *TIFS*, pages 2759–2772, 2021.

[Liu *et al.*, 2022a] Ajian Liu, Jun Wan, Ning Jiang, Hongbin Wang, and Yanyan Liang. Disentangling facial pose and appearance information for face anti-spoofing. In *ICPR*, pages 4537–4543. IEEE, 2022.

[Liu *et al.*, 2022b] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *TIFS*, pages 2497–2507, 2022.

[Liu *et al.*, 2023a] Ajian Liu, Zichang Tan, Yanyan Liang, and Jun Wan. Attack-agnostic deep face anti-spoofing. In *CVPR Workshops*, pages 6335–6344, 2023.

[Liu *et al.*, 2023b] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, S. Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *TIFS*, pages 4775–4786, 2023.

[Luo *et al.*, 2022] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *CVPR*, pages 15315–15324, 2022.

[Madry *et al.*, 2019] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[Moosavi-Dezfooli *et al.*, 2016] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016.

[Qiu *et al.*, 2020] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rony *et al.*, 2021] Jérôme Rony, Eric Granger, Marco Pedersoli, and Ismail Ben Ayed. Augmented lagrangian adversarial attacks. In *ICCV*, pages 7738–7747, 2021.

[Rosberg *et al.*, 2023] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: pose-and occlusion-aware high fidelity face swapping. In *WACV*, pages 3454–3463, 2023.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[Thies *et al.*, 2016] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016.

[Thies *et al.*, 2019] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm TOG*, pages 1–12, 2019.

[Wan *et al.*, 2023] Jun Wan, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. *Advances in Face Presentation Attack Detection*. Springer, 2023.

[Wang *et al.*, 2021a] Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *3DV*, pages 679–688, 2021.

[Wang *et al.*, 2021b] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049, 2021.

[Wang *et al.*, 2021c] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. *arXiv:2107.01396*, 2021.

[Wang *et al.*, 2022] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *CVPR*, pages 4123–4133, 2022.

[Wen *et al.*, 2015] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *TIFS*, pages 746–761, 2015.

[Yan *et al.*, 2022] Chiu Wai Yan, Tsz-Him Cheung, and Dit-Yan Yeung. Ila-da: Improving transferability of intermediate level attack with data augmentation. In *ICLR*, 2022.

[Yu *et al.*, 2020a] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z. Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. In *TPAMI*, 2020.

[Yu *et al.*, 2020b] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, pages 5295–5305, 2020.

[Yu *et al.*, 2022] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *arXiv:2208.05401*, 2022.

[Zhao *et al.*, 2021] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, pages 2185–2194, 2021.

[Zhou *et al.*, 2023] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *CVPR*, pages 20453–20463, 2023.

[Zi *et al.*, 2020] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020.

[Zou *et al.*, 2022] Junhua Zou, Yexin Duan, Boyu Li, Wu Zhang, Yu Pan, and Zhisong Pan. Making adversarial examples more transferable and indistinguishable. In *AAAI*, volume 36, pages 3662–3670, 2022.