Mixture-of-Attack-Experts with Class Regularization for Unified Physical-Digital Face Attack Detection

Shunxin Chen[1,2,3*], **Ajian Liu**[4,5*], **Junze Zheng**[5], **Jun Wan**[4,5,8†], **Kailai Peng**[6], **Sergio Escalera**[7], **Zhen Lei**[4,5,8,9]

[1]Nanjing University of Posts and Telecommunications, Nanjing, China
[2]Nanjing Artificial Intelligence Research of IA, Nanjing, China
[3]University of Chinese Academy of Sciences, Nanjing, China
[4]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[5]Macau University of Science and Technology (MUST), Macau, China
[6]Purple Mountain Laboratory, China
[7]Computer Vision Center (CVC), Barcelona, Catalonia, Spain
[8]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[9]CAIR, HKISI, Chinese Academy of Sciences, Hong Kong, China
chenshunxin23@mails.ucas.ac.cn, {ajian.liu,jun.wan}@ia.ac.cn

Abstract

Unified detection of digital and physical attacks in facial recognition systems has become a focal point of research in recent years. However, current multi-modal methods typically ignore the intra-class and inter-class variability across different types of attacks, leading to degraded performance.To address this limitation, we propose MoAE-CR, a framework that effectively leverages class-aware information for improved attack detection. Our improvements manifest at two levels, i.e., the *feature* and *loss* level. **At the feature level**, we propose Mixture-of-Attack-Experts (MoAEs) to capture more subtle differences among various types of fake faces. **At the loss level**, we introduce Class Regularization (CR) through the Disentanglement Module (DM) and the Cluster Distillation Module (CDM). The DM enhances class separability by increasing the distance between the centers of live and fake face classes. However, center-to-center constraints alone are insufficient to ensure distinctive representations for individual features. Thus, we propose the CDM to further cluster features around their class centers while maintaining separation from other classes. Moreover, specific attacks that significantly deviate from common attack patterns are often overlooked. To address this issue, our distance calculation prioritizes more distant features. Extensive experiments on two unified physical-digital attack datasets demonstrate the State-of-The-Art (SoTA) performance of the proposed method.

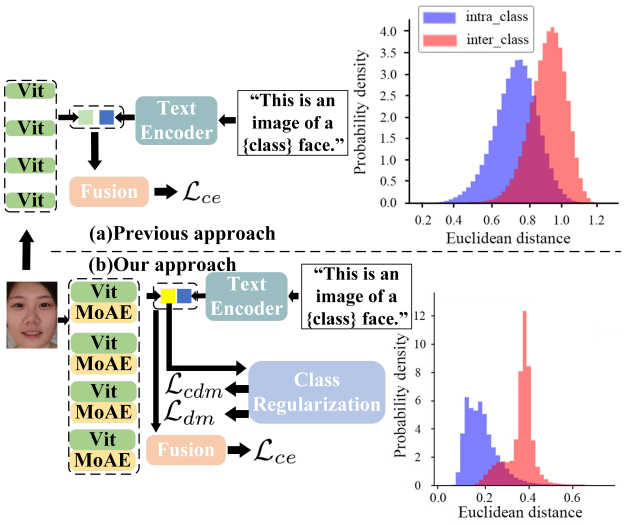

Figure 1: Comparison with existing methods. Greater overlap in histograms indicates poor class separation. (a) Previous methods focus on feature mining but overlook intra-class and inter-class variations. (b) Our method refines features and enforces constraints, achieving a more distinct and separable feature space.

Introduction

Facial recognition systems remain susceptible to a variety of attacks, broadly classified into physical and digital attacks. Each category comprises distinct types of attacks: physical attacks include print attacks (Zhang et al. 2020), replay attacks, and mask attacks (Liu et al. 2022a; Fang et al. 2023), whereas digital attacks encompass methods (Rössler et al. 2019) such as StyleGAN, FaceSwap, Deepfakes, and NeuralTextures. Research on physical attack detection (Liu, Jourabloo, and Liu 2018; Zhang et al. 2019; Yu et al. 2020; Cai et al. 2020; Liu et al. 2021) often involves the design of specialized networks to automatically extract spoofing cues and deceptive features from multiple modalities. In the realm of deepfake detection, numerous studies (Fei et al. 2022; Qian et al. 2020) leverage the spatial rich model,

*These authors contributed equally.
†Corresponding author

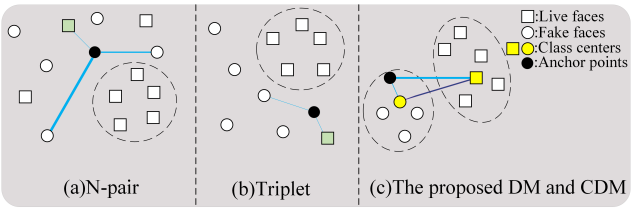

Figure 2: Comparison of popular class constraint methods and our approach. Small nodes represent the features of batch data. The connections defined by the loss are represented by edges, with thicker edges indicating larger gradients. (a) The N-pair loss reflects the hardness of the data but does not utilize all the data in the batch. (b) The triplet loss does not account for data hardness. The aggressive pushing mechanisms utilized by both (a) and (b) can lead to unintended class separation. Such forceful displacement may cause certain points, particularly green points, to diverge from their respective class clusters. (c) Our method considers all data in the batch, processes them with class centers, and simultaneously avoids class separation phenomena.

frequency details, and the relationships between facial action units to distinguish between genuine and fake faces. However, these methods cannot effectively address different types of attacks in different categories.

Previous approaches (Fang et al. 2024) seek to achieve classification by exploring a comprehensive feature space but do not account for the large intra-class differences and small inter-class differences present in physical and digital attack data. The inherent characteristics of the data make it extremely challenging to identify such a space, and neglecting these characteristics leads to suboptimal performance, as shown in Fig. 1. Common techniques for implementing class regularization include triplet loss (Schroff, Kalenichenko, and Philbin 2015; Wang et al. 2014) and N-pair loss (Sohn 2016). As illustrated in Fig. 2, N-pair loss connects an anchor point to a single positive data point and multiple negative data points, pulling the positive point closer to the anchor while pushing the negative points away, with consideration of their hardness. However, N-pair loss does not fully utilize the entire data batch, as it samples an equal number of points from each negative class, which may result in the exclusion of informative samples during training. On the other hand, triplet loss fails to adequately account for the difficulty of the data, resulting in limited sensitivity when processing distant features. Additionally, the direct pull-and-push mechanisms in both triplet loss and N-pair loss pose challenges when dealing with classes that have enveloping relationships, making it difficult for dispersed classes to cluster effectively, thereby limiting the model's ability to learn a unified feature representation.

In this paper, we propose the Mixture-of-Attack-Experts with Class Regularization (MoAE-CR) framework, which incorporates SoftMoE (Puigcerver et al. 2023) into the image encoder of CLIP (Radford et al. 2021) to enable more refined processing at the feature level. We further refined and proposed MoAE, enabling it to process features from multiple perspectives with greater granularity. Additionally, we introduce two novel constraint modules: the Disentanglement Module (DM) and the Cluster Distill Module (CDM). These modules account for all data within a batch during computation and employ a relational matrix to prevent class separation caused by simple pushing mechanisms. DM enhances the separation between these classes to address the challenge of small inter-class differences, particularly in distinguishing between real and fake faces. Meanwhile, CDM promotes the clustering of features around their respective class centers while maintaining separation from other class centers. Furthermore, distance is utilized as a constraint reference to mitigate the model's overlooking rare attacks. In summary, the main contributions of this paper are as follows:

- We propose a novel MoAE-CR framework, which incorporates MoAE and two regularization modules, DM and CDM. It demonstrates undeniable advantages over SoTA methods on two unified physical-digital attack datasets.
- At the feature level, we integrate SoftMoE into the image encoder of CLIP. To enable finer feature processing and capture the nuances of various attack types, MoAE enhances SoftMoE through the application of multi-head attention mechanisms.
- At the loss level, we utilize two constraint modules, DM and CDM. These modules ensure that live and fake faces exhibit greater intra-class aggregation and inter-class separation. In processing live and fake faces, all data within a batch is considered, with careful attention to the impact of distances. By assigning larger gradients to more distant features, we more effectively address attacks with skewed feature distributions.

Related Works

Face Anti-Spoofing

Face anti-spoofing is a technique designed to identify whether a face captured by sensors is genuine or a presentation attack, i.e., prints (Zhang et al. 2020), video replays, or 3D mask attacks (Liu et al. 2022a; Fang et al. 2023). With the advancement of deep learning, researchers (Liu, Jourabloo, and Liu 2018; Yu et al. 2020; Cai et al. 2020) have developed specialized networks that automatically extract spoofing cues. However, these algorithms suffer from performance degradation when facing unknown domains. To address this issue, recent methods have employed DA-based techniques (Liu et al. 2022b; Yue et al. 2023; Liu et al. 2024c) and DG-based approaches (Zheng et al. 2024b,c; Liu et al. 2023b; Cai et al. 2024; Liu et al. 2024b; Liu 2024) aim to learn domain-invariant features across multiple source domains. Also, incremental learning (IL) methods (Guo et al. 2022; Wang et al. 2024) are considered to tackle the catastrophic forgetting problem in the context of domain discontinuity in FAS. With the increasing advancement of physical presentation mediums, an increasing number of algorithms are mining complementary information from visible light, depth map, and near-infrared modes to identify spoofing clues, including multi-modal fusion (Zhang et al. 2020;

George et al. 2019), cross modal transformation (George and Marcel 2021; Liu et al. 2021), flexible modal (Liu and Liang 2023; Liu et al. 2023a; Yu et al. 2023b,a; Zhang et al. 2024; Liu et al. 2024a), and missing modality (Lin et al. 2024; Zheng et al. 2024a; Li et al. 2024).

Face Forgery Detection

Digital attack detection (Zhao et al. 2021; Song et al. 2024) aims to distinguish authentic images from digitally manipulated facial artifacts, or diffusion generated video. Numerous endeavors have been undertaken to enhance the efficacy of Digital attack detection techniques (Nguyen, Yamagishi, and Echizen 2019; Tolosana et al. 2020). 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 focus on identifying more reliable forgery patterns, including noise statistics, local textures, and frequency information. Zhao et al. (Zhao et al. 2021) introduced a texture enhancement block in shallow layers to extract and enhance texture features by applying average pooling to filter out texture details from feature maps and subtracting the result from the original image. For both cnn-synthesized and image editing forgery domains, HiFi-IFDL (Guo et al. 2023) and HiFi-Net++ (Guo et al. 2024) formulate the image forgery detection and localization (IFDL) as a hierarchical fine-grained classification problem, and classify the individual forgery method of given images via predicting the entire hierarchical path. DD-VQA (Zhang et al. 2025) extends deepfake detection from a conventional binary classification to a VQA task.

Physical-Digital Attack Detection

Recent studies have made significant strides in the detection of face fraud and forgery. A pioneering benchmark for detecting face fraud and forgery was established, integrating visual and physiological rPPG signals to address issues of generalization (Yu et al. 2024). Additionally, a comprehensive analysis of 25 documented attack types introduced a method that utilizes a multi-task learning framework alongside k-means enhancement techniques to differentiate between genuine identities and various attacks (Deb, Liu, and Jain 2023). La-SoftMoE (Zou et al. 2024) leverages re-weighted SoftMoE with linear attention, achieving satisfactory performance on tasks with sparse feature distributions. But its generalizaiton would not meet practical requirements. Further contributions include the development of a new benchmark using existing physical and digital attack datasets, employing reconstruction learning for detection (Cao et al. 2024). However, these studies have not investigated unified attack detection based on ID consistency.

Method

Our approach is based on a vision-language model, as illustrated in Fig. 3. It involves the adjustment of features through an enhanced MoAE, while introducing a CR module at the loss layer to facilitate intra-class aggregation and inter-class separation. Subsequently, we will provide a detailed introduction to the proposed MoAE-CR.

Preliminary

In this paper, we utilize large-scale Vision-Language Models (Jiang et al. 2024; Huang et al. 2024) (VLMs) like CLIP (Contrastive Language-Image Pre-Training) (Radford et al. 2021) based on contrastive learning to dynamically adjust classifier weights using textual features. CLIP integrates text and image encoders: a Transformer-based model converts text into fixed-size vectors, while convolutional neural networks (ResNet or Vision Transformer) convert images into vectors of the same dimensionality. Through contrastive learning, the model aligns image and text representations within a shared embedding space. The training objective is to minimize the cosine distance for positive pairs and maximize it for negative pairs using a symmetric cross-entropy loss function.

In our model, we keep the cross entropy of the similarity between images and texts in the CLIP model. The formula is shown as Eq. 1, where S is the similarity matrix, and the element $S_{i,j}$ of the similarity matrix S denotes the similarity between the i^{th} image embedding and the j^{th} text embedding.

$$\mathcal{L}_{ce} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(S_{i,i})}{\sum_{j=1}^{N}\exp(S_{i,j})} + \log\frac{\exp(S_{i,i})}{\sum_{j=1}^{N}\exp(S_{j,i})}\right). \quad (1)$$

In our task, the text in CLIP is limited to two categories: live and fake, preventing the exploration of specific attack types. To address data sparsity, we employ Soft Mixture of Experts (Soft MoEs). Unlike sparse and discrete routing mechanisms that assign tokens to experts definitively, Soft MoEs use a flexible approach by mixing tokens to handle features. This involves calculating multiple weighted averages of all tokens, with weights determined by both tokens and experts. These weighted averages are then processed by their respective experts. Soft MoEs show exceptional performance in handling sparse tasks in visual recognition.

Mixture-of-Attack-Experts

As illustrated in Fig. 3, the MoAE module is integrated into the image encoder, where it operates in parallel with the MLP within the transformer block, and the results are subsequently combined. Given the subtle differences among various types of deception attacks in our task, we propose the MoAE module. In our enhanced MoAE, the primary improvements involve the introduction of multi-head attention (Katharopoulos et al. 2020) to augment the model's representational capacity and learning ability, as well as the parallel processing of representations from different heads within the expert networks. Multiple experts can capture various attack features, and the use of soft routing allows for the weighted aggregation of features processed by all experts. The addition of multiple heads enables different branches to learn distinct attack traces, thereby enhancing the effectiveness of the anti-spoofing task.

Specifically, given an input $x \in \mathbb{R}^{n\times p\times d}$ where n denotes the batch size set to 32, p represents the sequence length,

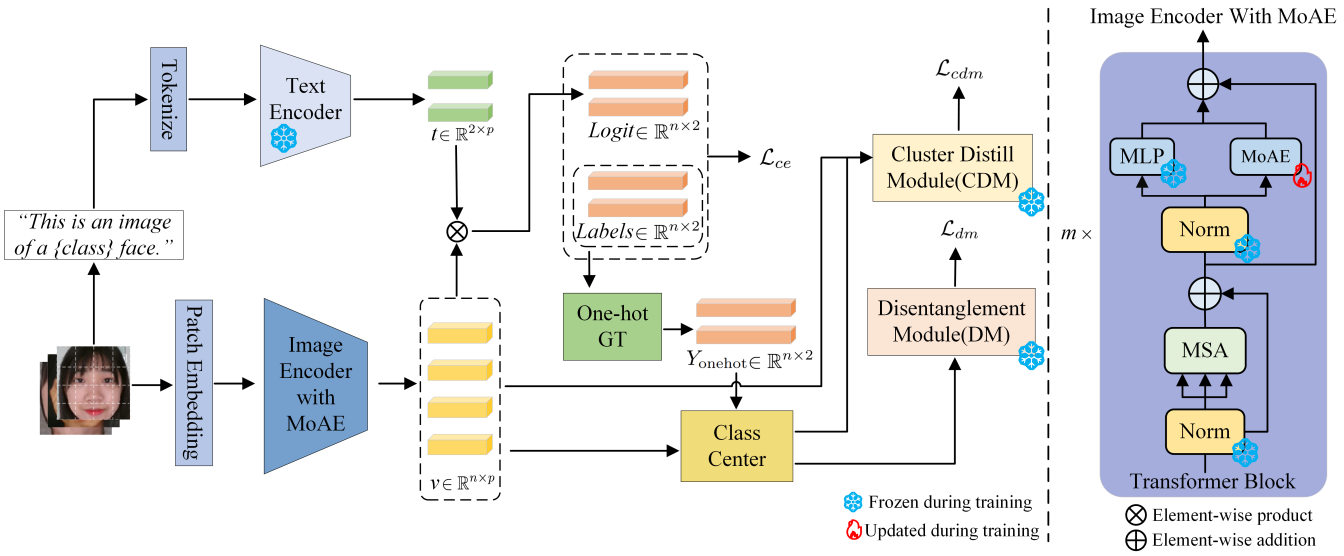

Figure 3: Our proposed MoAE-CR framework. This article primarily utilizes the Uniattack Datasetoder, and is designed to adapt to joint physical and digital attack tasks through contributions at two levels: (1) The image encoder incorporates MoAEs, which are composed of m Transformer Blocks. Our MoAEs facilitate more nuanced learning from multiple perspectives, resulting in superior feature representation. (2) Two constraint modules: the Disentanglement Module (DM) and the Cluster Distill Module (CDM). These modules maximize intra-class cohesion and inter-class separation between live and fake faces.

and d is the feature dimension, we first perform a linear transformation to obtain the query vector q, the key vector k, and the value vector v. Each MoE layer uses a set of m expert functions applied on individual tokens, namely $\left\{f_i : \mathbb{R}^d \to \mathbb{R}^d\right\}_{1:m}$. These vectors are divided into multiple heads. Subsequently, the attention scores for each head are computed and normalized, followed by applying the softmax function to the attention scores to obtain the attention weights, the attention output P is obtained by applying weighted summation of the value vectors v using the attention weights:

$$P = \text{softmax}\left(\frac{q \cdot k^T}{\sqrt{d_h}}\right) \cdot v, \tag{2}$$

where $d_h = \frac{d}{h}$, h is the number of heads.

In the improvement of the expert network, we process the output of each attention head separately through the expert network, allowing each head to focus on different feature representations. Specifically, given the output of each head, we need to reshape it to match the input format of the expert network. Subsequently, these reshaped outputs are processed through the expert network:

$$\mathrm{Y}_i = f_{\lfloor i/p \rfloor}\left(P_i\right). \tag{3}$$

Finally, the output Y of the expert network is reshaped back to its original form.

Class Regularization

The Class Regularization module comprises two constraint components: the Disentanglement Module (DM) and the Cluster Distillation Module (CDM). The DM increases the distance between the centers of real and fake face classes, thereby facilitating the separation of these classes and mitigating any potential dependencies between them. The CDM further refines feature clustering around their respective class centers while maximizing separation from centers of other classes, thereby optimizing class-specific characteristics. Additionally, to address the unique features found in combined physical and digital attacks, we prioritize features that are more distant during distance calculations.

Disentanglement Module. As depicted in Fig. 4, the DM module is computed using a relationship matrix based on class centers.Assume the input features $X \in \mathbb{R}^{n\times p}$ and labels $Y \in \mathbb{R}^n$. Then we generate one-hot labels $Y_{\text{onehot}} \in \mathbb{R}^{n\times 2}$. By utilizing simple summation and statistical methods, the masks class mask $\in \mathbb{R}^{2\times n}$ for the two classes of live faces and fake faces can be calculated. The formula for the class center is as follows:

$$R_{\text{center}} = \frac{R_{\text{s}}}{N}, \tag{4}$$

where N is the number of non-zero elements for each class, R_{s} is the sum of features for each class. N retains this meaning throughout the text.

The DM is utilized to compute the relative differences in features between classes, thereby promoting the distinct separation of features across different classes. To achieve this, we obtain the inter-class relation matrix R_{c} for both live and fake faces:

$$R_{\text{c}} = \frac{R_{\text{center}} \cdot R_{\text{center}}^T}{\sqrt{p}}, R_{\text{c}} \in \mathbb{R}^{2\times 2}. \tag{5}$$

Then, we remove diagonal elements $\text{diag} = I_{2\times 2}$ and

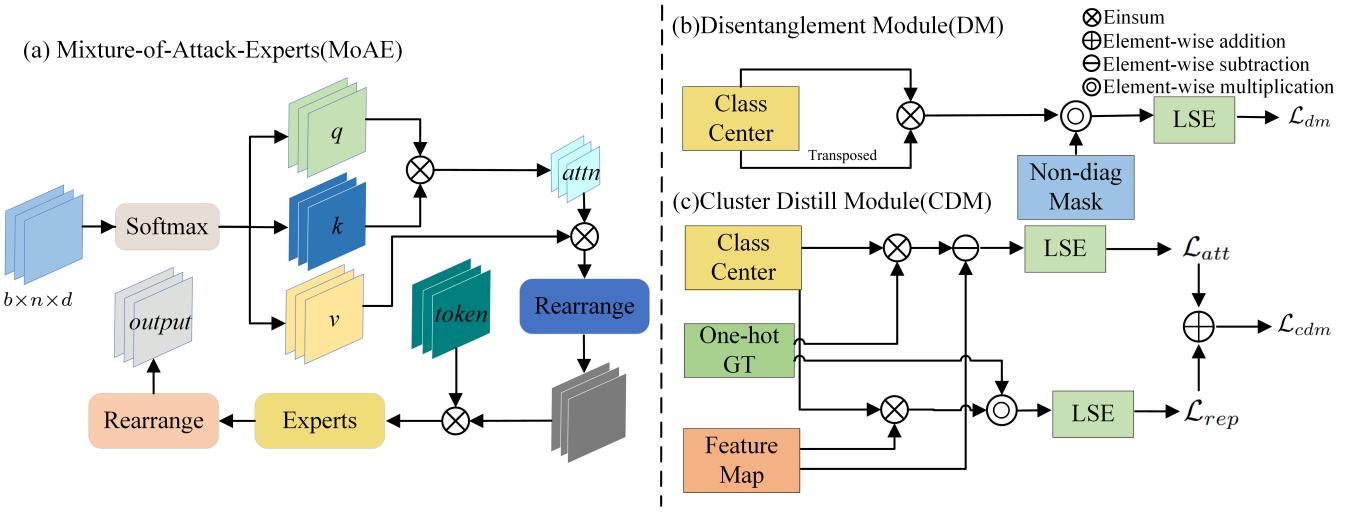

Figure 4: The detailed implementation structures of the Mixture-of-Attack-Experts (MoAEs), Disentanglement Module (DM), and Cluster Distill Module (CDM) are as follows. The MoAEs build upon the Soft MoEs by incorporating a multi-head attention mechanism to enhance feature processing. The DM utilizes a relationship matrix based on class centers to increase the distance between the centers of different classes. The CDM leverages this relationship matrix to bring each feature closer to its corresponding class center while distancing it from other class centers. Both DM and CDM employ the Log-Sum-Exp (LSE) function to prioritize more distant features.

compute the relation R_{or} between other classes:

$$R_{\text{or}} = R_{\text{c}} \cdot (1 - \text{diag}). \tag{6}$$

Building on this basis, we apply the threshold to calculate the relation difference Q:

$$Q = \max(R_{\text{or}} - \text{t}, 0), \tag{7}$$

where the t is a hyperparameter, which we set to 0.5.

Thus, the DM Center-to-center Loss $\mathcal{L}_{dm}$ is defined by:

$$\mathcal{L}_{dm} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\sum_{i\neq j}\exp\left(Q_{i,j} - \max Q_i\right)\right)^2, \tag{8}$$

where $maxQ_i$ is the maximum value of $Q_{i,j}$ across its last feature dimensions.It can be seen that we use Log-Sum-Exp here. Due to the properties of Log-Sum-Exp, the actual loss will pull and push all features in the batch, but with varying intensities depending on their relative hardness. We fully considered the impact of distance and subtracted the maximum value to reduce data fluctuations caused by Log-Sum-Exp, ensuring numerical stability.

Cluster Distillation Module. Although the DM module separates the centers of the two classes, there might still be instances where live and fake faces are not completely classified based on specific features. To address this, we further designed the CDM (Cluster Discrepancy Minimization). The goal of CDM is to achieve more compact intra-class clustering and greater inter-class separation of features. Specifically, we identify the centers of the live and fake face classes and ensure that each feature in a batch is close to its respective class center while being distant from the center of the other class. This is achieved by designing an attraction loss $\mathcal{L}_{att}$ and a repulsion loss $\mathcal{L}_{rep}$. Based on the class centers calculated in Eq. 4, the differences between the features and the mean features are computed. These differences reflect the distance between the features and the class centers:

$$R' = |\text{X} - R'_{\text{center}}|, R' \in \mathbb{R}^{n\times p}, \tag{9}$$

where R'_{center} is the corresponding class center.

$$\mathcal{L}_{att} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\sum_{i}\exp\left(R'_{i,j} - \max R'_i\right)\right)^2, \tag{10}$$

where similar to Eq. 8, $\max R'_i$ is the maximum value of $R'_{i,j}$ across its last feature dimensions.We similarly employ Log-Sum-Exp to emphasize the importance of distance and subtract the maximum value to stabilize the numerical calculations.

The relationship between the features and the class centers can be determined as follows:

$$R'' = \text{softmax}(X \cdot R^T_{\text{center}}), R'' \in \mathbb{R}^{n\times 2}, \tag{11}$$

$$\mathcal{L}_{rep} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\sum_{j\neq i}\exp\left(R''_{i,j} - \max R''_i\right)\right)^2. \tag{12}$$

We define $\mathcal{L}_{cdm}$ as:

$$\mathcal{L}_{cdm} = \mathcal{L}_{att} + \mathcal{L}_{rep}. \tag{13}$$

Model Training and Inference. We calculate the cross-entropy loss using the visual and textual features processed by the CLIP model. In the training phase, we update the parameters of the image encoder with MoAEs and text encoder. The full training objective of MoAE-CR is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{dm} + \mathcal{L}_{cdm}. \tag{14}$$

In the inference stage, the image encoder with MoAEs will adaptively engage different experts based on each example instance. Simultaneously, our DM and CDM, through constraints applied during training, ensure that the image processor achieves intra-class aggregation and inter-class separation, thereby better distinguishing between live and fake faces.

Experiments

To evaluate the performance of our proposed method in comparison to existing methods, we employed two publicly available datasets, UniAttackData (Fang et al. 2024) and JFSFDB (Yu et al. 2024), for face forgery detection. UniAttackData serves as the primary evaluation dataset due to its advantage of ID consistency. Our method exhibited superior performance and generalization capabilities across both datasets. Additionally, to substantiate the effectiveness of each proposed module, we conducted comprehensive ablation studies.

Experimental Settings

Datasets. UniAttackData extends the CASIA-SURF CeFA dataset by including digital forgery techniques, featuring 1,800 subjects from three ethnic groups and two physical attacks (Print and Replay), with each subject facing 12 digital attacks from six editing and six adversarial methods. It contains 28,706 videos, offering a broader attack variety per identity compared to GrandFake and JFSFDB. Two protocols are defined: Protocol 1 evaluates unified attack detection with all attack types in training, validation, and test sets, while Protocol 2 tests generalization to unseen attacks using a 'leave-one-type-out' approach, divided into P2.1 (unseen physical attacks) and P2.2 (unseen digital attacks). The dataset provides a robust framework for developing and evaluating advanced attack detection methods.

In addition to the UniAttackData, the JFSFDB (Yu et al. 2024) dataset, introduced by Yu et al., integrates nine subsets. The dataset provides two main protocols: separate training, where models address Presentation Attack (PA) and Deepfake Attack (DA) tasks independently, and joint training, which allows simultaneous handling of both tasks. In our study, we employ both protocols to evaluate the effectiveness of our method.

Implementation Details. We configured ViT-B/16 as the image encoder, with the number of experts and heads in the MoAE set to 4 and 2, respectively. The Adam optimizer was employed, with a learning rate of 1e-6 and a weight decay of 5e-4. The model was trained for 300 iterations.

Performance

To evaluate the proposed algorithm in Unified Attack Detection (UAD), we use standard metrics from physical and digital forgery detection: average classification error rate (ACER), overall detection accuracy (ACC), area under the curve (AUC), and equal error rate (EER). ACER and ACC are calculated based on thresholds from the development set.

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
	UniAttackDetection	0.52	99.45	99.96	0.53
	MoAE-CR(Our)	**0.37**	**99.47**	**99.97**	**0.49**
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
	UniAttackDetection	22.42±10.57	67.35±23.22	91.97±4.55	15.72±3.08
	MoAE-CR(Our)	**15.13±12.10**	**85.41±6.85**	**92.09±7.11**	**13.81±8.71**

Table 1: 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.

Prot.	Method	ACER(%)↓	ACC(%)↑	AUC(%)↑	EER(%)↓
1	SupContrastive	0.64	99.32	99.95	0.68
	N-pair	1.78	98.62	99.75	1.38
	Triplet	0.67	98.95	99.76	1.04
	Hard Triplet	1.36	98.90	99.90	1.09
	MoAE-CR(Our)	**0.37**	**99.47**	**99.97**	**0.49**
2	SupContrastive	16.44±14.14	68.63±15.08	84.77±14.91	24.31±22.13
	N-pair	18.79±13.68	79.80±12.34	85.05±14.14	18.66±13.88
	Triplet	20.00±14.80	69.44±15.93	82.52±15.64	25.86±20.30
	Hard Triplet	19.56±15.43	75.07±11.64	88.39±4.13	16.91±7.09
	MoAE-CR(Our)	**15.13±12.10**	**85.41±6.85**	**92.09±7.11**	**13.81±8.71**

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

Additionally, the robustness of our method is demonstrated through comparisons with a range of established competitors in face anti-spoofing and backbone networks, including ResNet50, ViT-B/16, FFD (Dang et al. 2020), CDCN (Yu et al. 2020), Auxiliary (Depth) (Liu, Jourabloo, and Liu 2018), and UniAttackDetection (Fang et al. 2024).

Table 1 shows that our method, MoAE-CR, achieves state-of-the-art performance across all metrics (ACER, ACC, AUC, EER) on UniAttackData. Its strong results in Protocol 2 demonstrate excellent generalization to "unseen" attacks.

In Protocol 1, our method achieves an ACER of 0.37%, surpassing the previous best of 0.52%. The ACC reaches 99.47%, exceeding the prior best of 99.45%. Both AUC and EER also show notable improvements.

In Protocol 2, our method demonstrates even more significant advances, with an average ACER of 15.13% and an average ACC of 85.41%, far outperforming previous methods. AUC and EER similarly show substantial improvements.

To validate the effectiveness of our DM and CDM modules, we replaced them with conventional techniques such as triplet loss and supervised contrastive loss. As shown in Table 2, our DM and CDM modules outperform these mainstream methods in both protocols, highlighting their superiority in handling combined digital and physical attack tasks.

To further assess the efficacy of the MoAE-CR method, we conducted additional tests on the JFSFDB dataset. As shown in Table 3, our proposed method, MoAE-CR, achieves SoTA performance with an ACER of 4.40%, an

Method	ACER(%)↓	ACC(%)↑	AUC(%)↑	EER(%)↓
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
MoAE-CR(Our)	**4.40**	**95.33**	**98.97**	**4.66**

Table 3: This table presents the results on the JFSFDB dataset under the p2 intra protocol, where our proposed MoAE-CR has the SoTA performance.

CLIP	SoftMoE	MoAE	DM	CDM	ACER(%)↓	ACC(%)↑	AUC(%)↑
✓	-	-	-	-	0.79	98.91	99.76
✓	✓	-	-	-	0.66	99.01	99.82
✓	-	✓	-	-	0.49	99.24	99.79
✓	-	✓	✓	-	0.95	98.73	99.79
✓	-	✓	-	✓	0.54	99.28	99.86
✓	-	✓	✓	✓	**0.37**	**99.47**	**99.97**

Table 4: Ablation of each component was conducted on the UniAttackData under Protocol I.

ACC of 95.33%, an AUC of 98.97%, and an EER of 4.66%. These results further substantiate the superior performance of our method compared to previous works.

Ablation Study

Effectiveness of Each Component. To assess the contribution of each component in our framework, we conducted ablation studies beginning with the baseline framework, CLIP. Specifically, Soft MoEs were employed to process different features through distinct experts and merge them via a soft routing mechanism, as shown in Table 4, leading to performance changes of -0.13% (ACER), +0.1% (ACC), and +0.06% (AUC).

Upon introducing fine-grained improvements to Soft MoEs, the optimized results for the three metrics improved to 0.49% (ACER), 99.24% (ACC), and 99.79% (AUC). This suggests that incorporating a multi-head attention mechanism within Soft MoEs further enhances the model's ability to represent features. While the individual introduction of DM and CDM into the Fine-Grained MoEs framework resulted in slight decreases across each metric, their combined application produced optimal outcomes of 0.37% (ACER), 99.47% (ACC), and 99.97% (AUC). These results indicate that DM and CDM exhibit a positive synergistic effect, and their integration delivers significant improvements.

Effects of the Number of Experts and Heads. We evaluated the impact of the number of heads and experts in MoAE. As shown in Table 5, the model generally performs better with four experts. However, increasing the number of attention heads leads to a decline in performance, suggesting that excessive focus can negatively impact the model's effectiveness. Similarly, having too many experts results in performance degradation, indicating that an excessive number of experts may cause overfitting or optimization difficulties. Therefore, we recommend using two attention heads and four experts as a better trade-off between performance and efficiency.

ACER(%)↓	Num of heads		
Experts	×2	×4	×8
×2	0.90	0.41	0.71
×4	**0.37**	0.41	0.52
×8	0.67	1.09	1.60

Table 5: This table shows the effects of the number of experts and attention heads.

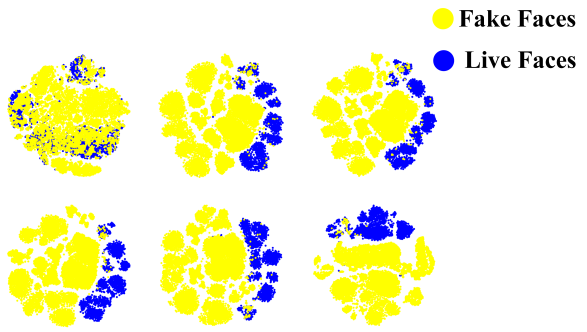

Figure 5: The figure presents the feature distribution visualization analysis of UniAttackData using the following methods: vanilla CLIP (top left), CLIP with SoftMoE (top center), CLIP with MoAE (top right), MoAE with DM (bottom left), MoAE with CDM (bottom center), and MoAE-CR (bottom right).

Visualization and Analysis

Using t-SNE and Matplotlib for feature visualization, as shown in 5, the feature distribution significantly improves upon introducing SoftMoE. The proposed MoAE-CR further optimizes the feature distribution compared to SoftMoE. The introduction of DM or CDM constraints enhances within-class clustering and improves class separation. The most notable improvements are observed when both DM and CDM constraints are applied simultaneously. This combined approach leads to clearer inter-class separation and more compact within-class distributions, facilitating the identification of the decision boundary. Further experimental analyses can be found in the supplementary materials.

Conclusion

In this work, we introduce the MoAE-CR framework to effectively tackle the challenges arising from combined digital and physical attacks, addressing these at both the feature and loss levels. At the feature level, our framework integrates Soft MoEs, and we further propose MoAE to enhance feature processing. At the loss level, we incorporate two constraint modules, DM and CDM, to facilitate more accurate classification by promoting a more balanced distribution between live and fake faces. Extensive experiments and visual analyses substantiate the superiority of the proposed MoAE-CR framework.

Acknowledgments

This work was supported by Beijing Natural Science Foundation JQ23016, the Chinese National Natural Science Foundation Projects 62476273, 62406320, U23B2054, 62276254, the Science and Technology Development Fund of Macau Project 0044/2024/AGJ, 0123/2022/A3, 0070/2020/AMJ, 0096/2023/RIA2, Ant Group and InnoHK program.

References

Cai, R.; Li, H.; Wang, S.; Chen, C.; and Kot, A. C. 2020. DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16: 937–951.

Cai, R.; Soh, C.; Yu, Z.; Li, H.; Yang, W.; and Kot, A. C. 2024. Towards Data-Centric Face Anti-spoofing: Improving Cross-Domain Generalization via Physics-Based Data Synthesis. *International Journal of Computer Vision*, 1–22.

Cao, J.; Zhang, K.-Y.; Yao, T.; Ding, S.; Yang, X.; and Ma, C. 2024. Towards Unified Defense for Face Forgery and Spoofing Attacks via Dual Space Reconstruction Learning. *International Journal of Computer Vision*, 1–26.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 5781–5790.

Deb, D.; Liu, X.; and Jain, A. K. 2023. Unified detection of digital and physical face attacks. In *FG*, 1–8. IEEE.

Fang, H.; Liu, A.; Wan, J.; Escalera, S.; Zhao, C.; Zhang, X.; Li, S. Z.; and Lei, Z. 2023. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*.

Fang, H.; Liu, A.; Yuan, H.; Zheng, J.; Zeng, D.; Liu, Y.; Deng, J.; Escalera, S.; Liu, X.; Wan, J.; et al. 2024. Unified physical-digital face attack detection. *arXiv preprint arXiv:2401.17699*.

Fei, J.; Dai, Y.; Yu, P.; Shen, T.; Xia, Z.; and Weng, J. 2022. Learning second order local anomaly for general face forgery detection. In *CVPR*, 20270–20280.

George, A.; and Marcel, S. 2021. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, 7882–7891.

George, A.; Mostaani, Z.; Geissenbuhler, D.; Nikisins, O.; Anjos, A.; and Marcel, S. 2019. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE transactions on information forensics and security*, 15: 42–55.

Guo, X.; Liu, X.; Masi, I.; and Liu, X. 2024. Language-guided Hierarchical Fine-grained Image Forgery Detection and Localization. *International Journal of Computer Vision*.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3155–3165.

Guo, X.; Liu, Y.; Jain, A.; and Liu, X. 2022. Multi-domain Learning for Updating Face Anti-spoofing Models. In *In Proceeding of European Conference on Computer Vision*.

Huang, W.; Zheng, X.; Ma, X.; Qin, H.; Lv, C.; Chen, H.; Luo, J.; Qi, X.; Liu, X.; and Magno, M. 2024. An Empirical Study of LLaMA3 Quantization: From LLMs to MLLMs. arXiv:2404.14047.

Jiang, Y.; Yan, X.; Ji, G.-P.; Fu, K.; Sun, M.; Xiong, H.; Fan, D.-P.; and Khan, F. S. 2024. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1): 17.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, 5156–5165. PMLR.

Li, Z.; Yu, Z.; Lin, X.; Selvaraj, N. M.; Guo, X.; Shen, B.; Kong, A. W.; and Kot, A. C. 2024. Flexible-Modal Deception Detection with Audio-Visual Adapter. In *IJCB*, 1–10.

Lin, X.; Wang, S.; Cai, R.; Liu, Y.; Fu, Y.; Tang, W.; Yu, Z.; and Kot, A. 2024. Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 211–221.

Liu, A. 2024. CA-MoEiT: Generalizable Face Anti-spoofing via Dual Cross-Attention and Semi-fixed Mixture-of-Expert. *International Journal of Computer Vision*, 1–14.

Liu, A.; and Liang, Y. 2023. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. *arXiv preprint arXiv:2304.07549*.

Liu, A.; Ma, H.; Zheng, J.; Yuan, H.; Yu, X.; Liang, Y.; Escalera, S.; Wan, J.; and Lei, Z. 2024a. FM-CLIP: Flexible Modal CLIP for Face Anti-Spoofing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8228–8237.

Liu, A.; Tan, Z.; Wan, J.; Liang, Y.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16: 2759–2772.

Liu, A.; Tan, Z.; Yu, Z.; Zhao, C.; Wan, J.; Lei, Y. L. Z.; Zhang, D.; Li, S. Z.; and Guo, G. 2023a. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*.

Liu, A.; Xue, S.; Gan, J.; Wan, J.; Liang, Y.; Deng, J.; Escalera, S.; and Lei, Z. 2024b. CFPL-FAS: Class Free Prompt Learning for Generalizable Face Anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, A.; Zhao, C.; Yu, Z.; Wan, J.; Su, A.; Liu, X.; Tan, Z.; Escalera, S.; Xing, J.; Liang, Y.; et al. 2022a. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17: 2497–2507.

Liu, Y.; Chen, Y.; Dai, W.; Gou, M.; Huang, C.-T.; and Xiong, H. 2022b. Source-Free Domain Adaptation with Contrastive Domain Alignment and Self-supervised Exploration for Face Anti-Spoofing. In *ECCV*.

Liu, Y.; Chen, Y.; Dai, W.; Gou, M.; Huang, C.-T.; and Xiong, H. 2024c. Source-Free Domain Adaptation With Domain Generalized Pretraining for Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Chen, Y.; Gou, M.; Huang, C.-T.; Wang, Y.; Dai, W.; and Xiong, H. 2023b. Towards Unsupervised Domain Generalization for Face Anti-Spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, Y.; Jourabloo, A.; and Liu, X. 2018. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 389–398.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2307–2311. IEEE.

Puigcerver, J.; Riquelme, C.; Mustafa, B.; and Houlsby, N. 2023. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, 86–103. Springer.

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

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision (ICCV)*.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 1–11.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Song, X.; Guo, X.; Zhang, J.; Li, Q.; Bai, L.; Liu, X.; Zhai, G.; and Liu, X. 2024. On Learning Multi-Modal Forgery Representation for Diffusion Generated Video Detection. In *NeurIPS*.

Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64: 131–148.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1386–1393.

Wang, K.; Zhang, G.; Yue, H.; Liu, A.; Zhang, G.; Feng, H.; Han, J.; Ding, E.; and Wang, J. 2024. Multi-domain incremental learning for face presentation attack detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5499–5507.

Yu, Z.; Cai, R.; Cui, Y.; Liu, A.; and Chen, C. 2023a. Visual prompt flexible-modal face anti-spoofing. *arXiv preprint arXiv:2307.13958*.

Yu, Z.; Cai, R.; Li, Z.; Yang, W.; Shi, J.; and Kot, A. C. 2024. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *IEEE Transactions on Dependable and Secure Computing*.

Yu, Z.; Liu, A.; Zhao, C.; Cheng, K. H.; Cheng, X.; and Zhao, G. 2023b. Flexible-modal face anti-spoofing: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6346–6351.

Yu, Z.; Wan, J.; Qin, Y.; Li, X.; Li, S. Z.; and Zhao, G. 2020. NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9): 3005–3023.

Yue, H.; Wang, K.; Zhang, G.; Feng, H.; Han, J.; Ding, E.; and Wang, J. 2023. Cyclically disentangled feature translation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3358–3366.

Zhang, S.; Liu, A.; Wan, J.; Liang, Y.; Guo, G.; Escalera, S.; Escalante, H. J.; and Li, S. Z. 2020. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2): 182–193.

Zhang, S.; Wang, X.; Liu, A.; Zhao, C.; Wan, J.; Escalera, S.; Shi, H.; Wang, Z.; and Li, S. Z. 2019. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 919–928.

Zhang, Y.; Colman, B.; Guo, X.; Shahriyari, A.; and Bharaj, G. 2025. Common Sense Reasoning for Deepfake Detection. In *European Conference on Computer Vision*, 399–415. Springer.

Zhang, Y.; Zhu, X.; Liu, A.; Lin, X.; Wan, J.; Zhang, J.; and Lei, Z. 2024. CPL-CLIP: Compound Prompt Learning for Flexible-Modal Face Anti-Spoofing. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-attentional deepfake detection. In *CVPR*, 2185–2194.

Zheng, G.; Liu, Y.; Dai, W.; Li, C.; Zou, J.; and Xiong, H. 2024a. Towards Unified Representation of Invariant-Specific Features in Missing Modality Face Anti-Spoofing. In *ECCV*.

Zheng, T.; Li, B.; Wu, S.; Wan, B.; Mu, G.; Liu, S.; Ding, S.; and Wang, J. 2024b. MFAE: Masked Frequency Autoencoders for Domain Generalization Face Anti-Spoofing. *IEEE Trans. Inf. Forensics Secur.*, 19: 4058–4069.

Zheng, T.; Yu, Q.; Chen, Z.; and Wang, J. 2024c. FAMIM: A Novel Frequency-Domain Augmentation Masked Image Model Framework for Domain Generalizable Face Anti-Spoofing. In *ICASSP*, 4470–4474.

Zou, H.; Du, C.; Zhang, H.; Zhang, Y.; Liu, A.; Wan, J.; and Lei, Z. 2024. La-SoftMoE CLIP for Unified Physical-Digital Face Attack Detection. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 1–11. IEEE.