FC-MoE: Flip Consistent Mixture of Experts for Unified Face Attack Detection

Anonymous ICME submission

***Abstract*—Unified face attack detection remains challenging for existing face attack detectors due to small intra-class similarities and large inter-class similarities. To address these issues, we propose partitioning the feature space according to facial regions by introducing the sparse mixture-of-experts (MoE) mechanism, where each expert handles one or more facial regions. Furthermore, considering the inherent flip semantic consistency of facial images, we propose a self-supervised loss function to guide the model in selecting the same expert for symmetrical facial regions. We construct a robust model, the Flip Consistent Mixture of Experts (FC-MoE) detector. Experimental results on the GrandFake and MS-UFAD datasets demonstrate that with an adaptive divide-and-conquer mechanism, the FC-MoE detector achieves good generalization ability in detecting various categories of attacks while maintaining excellent extensibility and application efficiency.**

***Index Terms*—Unified Face Attack Detection, Mixture-of-Experts, Flip Consistent Loss**

I. Introduction

Automated face recognition (AFR) systems have been deployed in various scenarios such as security surveillance, access control, and smart devices. Diverse and evolving attack techniques can be represented with an ever-growing attack tree, as illustrated in Fig. 1(a). Numerous detection methods have been proposed to protect AFR systems from attacks. Despite impressive detection rates, prevailing efforts have predominantly focused on detecting an individual category of attack techniques, such as adversarial attacks [1], digital manipulation attacks [2]–[6], or spoof attacks [7]–[11]. In practical applications, knowing the specific face attack category in advance may not be possible, so a unified detector capable of defending AFR systems against arbitrary attack categories is crucial.

Unified face attack detection (UFAD) is more challenging than detecting an individual category of attack techniques primarily due to the following two characteristics of its data, as illustrated in Fig. 1(b): (1) Small intra-class similarities. Attack samples exhibit dramatically different attack cues due to the varying underlying principles of each attack technique. (2) Large inter-class similarities. When training with all samples, some attack techniques have relatively subtle attack cues, causing their samples to be closer to bona fide samples in the feature space. Existing attack detectors are mainly dense models [2], [7], [13], [14]. Due to having only a single feature space, these detectors often exhibit performance degradation as the disparity in attack techniques increases. To address this issue, an intuitive idea is the divide-and-conquer strategy. Multi-branch detectors [12], [15] partition the feature space into multiple subspaces according to the similarity of attack techniques, with each subspace handled by a separate branch. Similar attack techniques are first clustered into one attack group and then are handled by a dedicated branch in a feature subspace. Since the attack samples within the same attack group can be more compact in the feature subspace, the multi-branch detectors can achieve better performance compared to dense detectors. However, the multi-branch mechanism results in low extensibility and application efficiency.

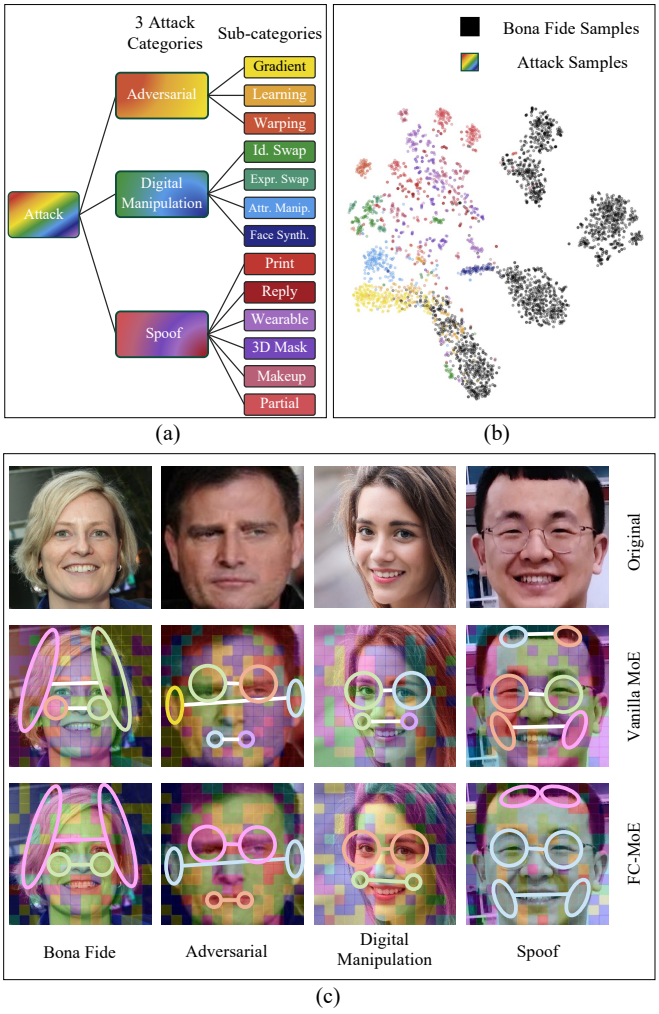

Fig. 1. (a) Ever-growing attack tree. (b) T-SNE visualization for samples from the GrandFake [12] dataset. The data exhibits characteristics of small intra-class similarities and large inter-class similarities. (c) Expert selection visualization before and after introducing flip consistent loss. Our proposed FS-MoE detector tends to select the same expert for symmetrical facial regions, whereas the Vanilla MoE detector does not exhibit this tendency.

In this paper, we propose partitioning the feature space according to facial regions instead of attack techniques. We introduce the sparse mixture-of-experts mechanism into the vision transformer, allowing each visual token to select the appropriate expert to process the features within its corresponding patch. Since the relationship between experts and facial regions is not explicitly specified but is instead learned by the model, our approach realizes an adaptive divide-and-conquer strategy. It effectively addresses the challenges present in the UFAD task: (1) Small intra-class similarities: Since each feature space focuses on a local facial region, it can better capture the common features of attack cues within that region. (2) Large inter-class similarities: An attack sample might resemble bona fide samples in some regions within the feature space, but it remains sufficiently distinguishable in other regions.

Furthermore, we aim to achieve more robust feature representations. Due to the inherent flip semantic consistency of facial images, symmetrical facial regions (e.g., the left eye and the right eye) should be processed by the same expert rather than different ones. This will allow each expert to learn more discriminative features within its responsible subspace, thereby improving the model's performance. To this end, we construct a self-supervised flip consistent loss function that encourages the selected experts for each patch to remain consistent before and after horizontal flipping. Specifically, we minimize the Jensen-Shannon divergence between the weight distributions assigned to the experts by the routing network before and after flipping. The resulting model is referred to as the Flip Consistent Mixture of Experts (FC-MoE) detector. Fig. 1(c) illustrates the expert selection before and after introducing the flip consistent loss. Our proposed FS-MoE detector tends to select the same expert for symmetrical facial regions, as highlighted in the third row of images, whereas the Vanilla MoE detector does not exhibit this tendency, as highlighted in the second row of images.

Experimental results demonstrate that the proposed FC-MoE detector is a robust learner for unified face attack detection. The FC-MoE detector only adds the computational burden of a lightweight routing network, achieving significantly improved performance while overcoming the limitations in extensibility and application efficiency associated with multi-branch detectors. On the reproduced GrandFake dataset, the FC-MoE detector achieves reductions in ACER by 1.31%, 1.29%, and 1.05% under the tiny, small, and base parameter settings, respectively, when compared to the vanilla MoE detector, which serves as a strong baseline.

II. Related Work

In recent years, there have been substantial advancements in face attack techniques. In the literature, attack techniques are broadly categorized into three types: (1) adversarial attacks [16]–[21], which involve adding imperceptible noise to face images or warping them to deceive AFR systems; (2) digital manipulation attacks [22]–[25], which use generative models to alter the identity, expression, or attributes of face images, or to synthesize entirely new face images; and (3) spoof attacks [26]–[30], which include physical artifacts such as printed images, replayed videos, and 3D masks.

Many methods have been developed to detect an individual category of attack techniques, while the research community has paid insufficient attention to unified face attack detection. UniFAD [12] constructs a unified multi-branch detector with the following pipeline. First, a clustering algorithm is executed to group attack techniques with high feature similarity. Next, attack samples from each attack group are mixed with bona fide samples to train a branch, which handles that specific attack group. Finally, all branches perform decision fusion. Additionally, JFSFD [15] has also implemented a multi-branch detector in its experiments.

The mixture of experts (MoE) mechanism is first proposed in the early 1990s [31]. Recently, the success of MoE in large language models [32] has drawn the attention of the computer vision community. The sparse MoE is introduced into computer vision tasks [33], demonstrating its effectiveness. In large language or vision models, MoE is primarily used to enhance model capacity without significantly increasing computational overhead. In our approach, however, the key role of MoE lies in providing a novel feature space partitioning method, which is suitable for addressing the issues of small intra-class similarities and large inter-class similarities in the UFAD task.

III. Method

In this section, we first introduce the overall architecture of the MoE detector. Then, we provide the formulation of the proposed self-supervised flip consistent loss function. Finally, we offer a comprehensive comparison between multi-branch and MoE detectors.

A. Overall Architecture of the MoE Detector

Fig. 2 illustrates model structures of dense, multi-branch, and MoE detectors. Similar to multi-branch detectors, MoE detectors employ early shared blocks in the low-level layers, while feature space partitioning is conducted in the high-level layers. This design addresses the issues of large intra-class similarities and small inter-class similarities that predominantly occur during semantic feature extraction in the high-level layers. In contrast, the low-level layers are responsible for modeling local geometric information, and not sharing parameters in these layers could lead to the risk of overfitting [34]. For both the multi-branch detector and the MoE detector, the optimal number of early shared blocks is determined through ablation studies in our paper and is set to 8 and 4, respectively. In the high-level layers of the MoE detector, sparse MoE transformer blocks and vanilla transformer blocks are alternately configured. The sparse MoE blocks are responsible for extracting features from one or several specific facial regions, whereas the vanilla transformer blocks handle feature fusion.

A sparse MoE transformer block is derived by replacing the MLP layer in a vanilla transformer block [14] with a sparse MoE layer. In detail, we denote independent experts

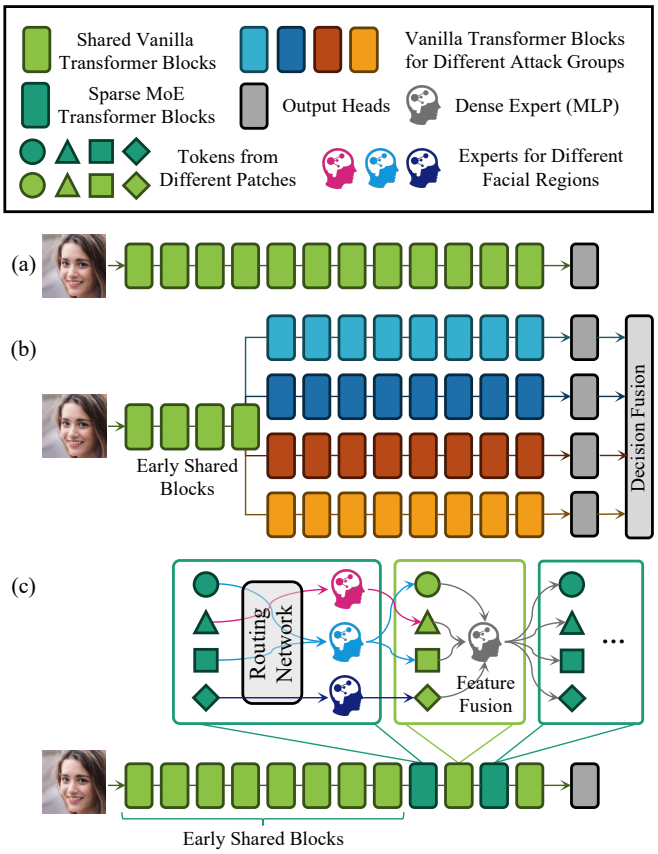

Fig. 2. The model structures of (a) dense, (b) multi-branch, and (c) MoE detectors.

as $E_i, i = 1, \ldots, N$. N is set to 6 in our scenario for optimal performance. Each expert is an MLP with specific learnable weights and can be denoted as:

$$E_i(\mathbf{x}) = \mathrm{MLP}_i(\mathbf{x}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^D$ is an input token to the layer from a certain patch, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the parameters of the two-layer MLP and $\sigma(\cdot)$ is the non-linear activation function GeLU. To determine the input-conditioned weight of experts for each token $\mathbf{x}$, we employ a routing network, defined as:

$$G(\mathbf{x}) = \mathrm{TOP}_k(\mathrm{Softmax}(\mathbf{W}\mathbf{x} + \epsilon)), \tag{2}$$

where $\mathbf{W}$ is the parameter of the routing network. TOP_k operation is a one-hot embedding that sets all elements of the output vector to zero except for the k largest values. We choose $k = 1$ in our scenario for optimal performance. The term ϵ is sampled from $\epsilon \sim \mathcal{N}(0, 1/N^2)$. Ultimately, the MoE layer can be represented as:

$$\mathrm{MoE}(\mathbf{x}) = \sum_{i=1}^{N} G(\mathbf{x})_i \cdot E_i(\mathbf{x}). \tag{3}$$

The MoE mechanism enables the effective utilization of different experts for different patches, thereby enhancing the model's capacity to capture attack cues in specific facial regions.

B. Flip Consistent Loss

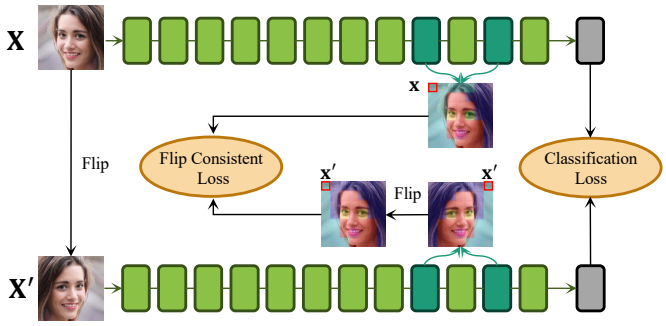

Fig. 3. The training framework of the proposed FC-MoE detector.

The overall training framework of our proposed FC-MoE detector is shown in Fig. 3. Given an input face image $\mathbf{X}$, it can be divided into L non-overlapping patches. Each patch is then encoded into a visual token $\mathbf{x} \in \mathbb{R}^D$ and passed through transformer blocks. As described in Section III-A, the routing network within the sparse MoE transformer blocks provides weight distributions assigned to experts for each $\mathbf{x}$, which can be expressed as:

$$\mathbf{P}(\mathbf{x}) = \mathrm{Softmax}(\mathbf{W}\mathbf{x} + \epsilon). \tag{4}$$

We horizontally flip the image $\mathbf{X}$, resulting in image $\mathbf{X}'$. The corresponding visual token in $\mathbf{X}'$ for each token $\mathbf{x}$ in the original image $\mathbf{X}$ can be denoted as $\mathbf{x}'$. The routing network similarly provides weight distributions assigned to experts for $\mathbf{x}'$, as:

$$\mathbf{Q}(\mathbf{x}') = \mathrm{Softmax}(\mathbf{W}\mathbf{x}' + \epsilon). \tag{5}$$

For any visual token $\mathbf{x}$, we define the flip consistent loss function using the Jensen-Shannon divergence as:

$$L_{fc}(\mathbf{x}) = \mathrm{D}_{JS}(\mathbf{P}(\mathbf{x}) \parallel \mathbf{Q}(\mathbf{x}')), \tag{6}$$

constraining the selected experts to remain consistent before and after horizontal flipping for each patch.

For the input image $\mathbf{X}$ and its flipped counterpart $\mathbf{X}'$, we average the flip consistent loss $L_{fc}(\mathbf{x})$ over all visual tokens processed by both sparse MoE transformer blocks (9-th and 11-th layers), yielding the final flip consistent loss $L_{fc}(\mathbf{X})$. Additionally, we average the cross-entropy loss for images $\mathbf{X}$ and $\mathbf{X}'$, yielding the final classification loss $L_c(\mathbf{X})$. The overall loss function is defined as:

$$L = L_c(\mathbf{X}) + \lambda L_{fc}(\mathbf{X}). \tag{7}$$

In our experiments, λ is set to 0.01.

C. Comparison between Multi-Branch and MoE Detectors

To address the issue of small intra-class similarities and large inter-class similarities in UFAD data, both the multi-branch detector and the MoE detector adopt a divide-and-conquer strategy, partitioning the feature space based on attack groups and facial regions, respectively. This divide-and-conquer process can be expressed using conditional statements, as shown in Table I.

TABLE I
COMPARISON BETWEEN MULTI-BRANCH AND MOE DETECTORS.

Detector	Multi-Branch	Vanilla MoE / FC-MoE
Partitioning According to What?	Attack Groups	Facial Regions
Divide-And-Conque	Pre-determined	Adaptive
Fusion Operation	Decision Fusion	Feature Fusion
Extensibility	Low	High
Application Efficiency	Low	High
Conditional Statements	If there is an attack clue in any attack group, facial region, then the sample is considered an attack sample.	

For the multi-branch detector, partitioning the feature space allows the attack samples within the same attack group to be more compact in the feature subspace, thereby improving performance. However, the multi-branch detector has notable limitations: (1) Low extensibility. The grouping is based on the current attack tree, and for newly emerging attack techniques, the clustering algorithm must be re-executed. (2) Low application efficiency. As the attack tree continues to grow, the number of branches will also increase.

The MoE detector, due to its more unified structure, avoids these limitations. Moreover, the MoE detector achieves performance comparable to the multi-branch detector while maintaining computational overhead close to that of the dense detector, demonstrating the potential of the MoE architecture. Our proposed flip consistent loss further enhances the performance of the MoE detector. The resulting FS-MoE detector exhibits significant advantages in terms of extensibility, application efficiency, and overall performance.

IV. EXPERIMENTS

A. Implementation Details

To demonstrate the effectiveness of our proposed FC-MoE detector, we conduct experiments on the reproduced GrandFake [12] dataset (as shown in Fig. 2(a)) and the MS-UFAD [35] dataset. Both datasets encompass adversarial attacks, digital manipulation attacks, and spoof attacks, including 25 and 52 attack techniques, respectively. Our FC-MoE detector was trained using the Adam optimizer for 3 epochs with a batch size of 32. The initial learning rate is set to 1×10^{-4} and halves after each epoch. Similar to V-MoE [33], we also use slight variants of two of the auxiliary losses proposed in [36] to encourage load balancing. All training is conducted on 1 NVIDIA RTX 4090 GPU.

B. Performance Comparison of Dense, Multi-Branch, Vanilla MoE, and FC-MoE Detectors

Table II compares the dense detector, the multi-branch detector, the vanilla MoE detector, and our proposed FC-MoE detector, in terms of performance, parameter count, and computational overhead. For a fair comparison, the dense and multi-branch detectors leverage ViT [14] and UniFAD-ViT respectively, with the same width and depth of transformer blocks as the vanilla MoE and FC-MoE detectors. Here,

TABLE II
COMPARISON OF PERFORMANCE, PARAMETER COUNT, AND COMPUTATIONAL OVERHEAD OF DENSE, MULTI-BRANCH, AND MOE DETECTORS AT DIFFERENT SCALES.

Scale	Detector	ACER	Params (M)	FLOPs (GMac)
Tiny	Dense	10.37	5.5	1.08
	Multi-branch	7.36	16.2	3.19
	Vanilla MoE	7.32	8.5	1.08
	FC-MoE (Ours)	**6.01**	8.5	1.08
Small	Dense	9.21	21.7	4.26
	Multi-branch	6.98	64.2	12.66
	Vanilla MoE	6.58	33.5	4.26
	FC-MoE (Ours)	**5.29**	33.5	4.26
Base	Dense	6.89	85.8	16.88
	Multi-branch	5.93	255.9	50.42
	Vanilla MoE	4.95	133.0	16.88
	FC-MoE (Ours)	**3.90**	133.0	16.88

UniFAD-ViT refers to the transformer version of UniFAD [12] that we have reproduced. At various parameter scales, the vanilla MoE detector and the multi-branch detector exhibit comparable performance, significantly surpassing the dense detector ViT. The proposed FC-MoE detector demonstrates substantial improvements over the vanilla MoE detector. It is important to note that the computational overhead of the FC-MoE detector during inference is significantly lower than that of UniFAD-ViT and nearly identical to that of ViT, yet it achieves markedly superior performance compared to UniFAD-ViT. Additionally, FC-MoE-Tiny even outperforms UniFAD-ViT-Small, and FC-MoE-Small exceeds the performance of UniFAD-ViT-Base, showcasing the strong potential of the MoE architecture for UFAD tasks. The partitioning of the feature space according to facial regions proves to be an effective and efficient method for addressing UFAD challenges, demonstrating advantages in both performance and efficiency.

C. Performance Evaluation on GrandFake and MS-UFAD

Table III presents the results comparing the proposed FS-MoE detector with existing state-of-the-art dense detectors and

TABLE III
COMPARISON WITH SOTA DETECTORS ON THE REPRODUCED GRANDFAKE DATASET AND THE MS-UFAD DATATSET.

Detector	Method	GrandFake			MS-UFAD
		ACER	ERR	AUC	ACER
Dense	FFD	8.17	5.85	96.42	3.01
	CDCN	9.57	7.06	91.66	3.41
	CDCN++	9.19	6.53	93.39	3.37
	Resnet18	8.84	7.29	95.89	2.93
	ViT-Tiny	10.37	9.58	95.41	3.84
Multi -Branch	UniFAD -Semantic -ViT-Tiny	9.39	6.76	94.68	3.06
	UniFAD -ViT-Tiny	7.36	5.63	97.62	-
MoE	Vanilla -MoE-Tiny	7.32	4.10	98.02	2.62
	FC-MoE -Tiny(Ours)	**6.01**	**3.21**	**98.63**	**2.21**

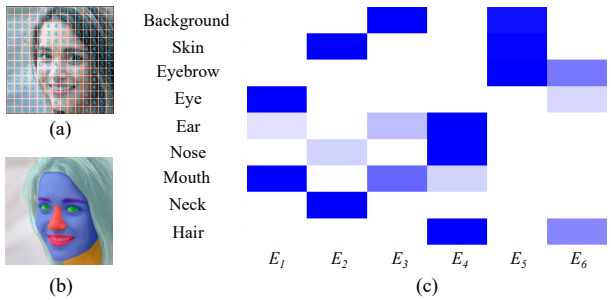

Fig. 4. Model analysis. (a) Expert selection. (b) Facial regions. (c) Correlation between experts and facial regions.

multi-branch detectors. Among them, UniFAD-Semantic-ViT-Tiny is the version that divides attack groups according to attack categories. It can be seen that our method achieves superior performance compared to the existing methods on both GrandFake and MS-UFAD.

D. Model Analysis

We train the FS-MoE detector on the GrandFake to explore the correlation between expert and facial regions. In FS-MoE, each MoE Block contains 6 experts. For each sample, we can record the expert selection for each patch (Fig. 4(a)), while simultaneously performing face parsing to obtain the segmentation results of facial regions (Fig. 4(b)). By aggregating the expert selections corresponding to each facial region across all samples, we can generate a 2D histogram correlation, as illustrated in Fig. 4(c). It can be observed that, even without any visual supervision signals for facial regions, the ability to correlate facial regions automatically emerges during training. For example, the features of the eye region are primarily extracted by Expert 0. This indicates that MoE detector is an adaptive divide-and-conquer mechanism, and partitions the feature space according to facial regions, with each expert being responsible for one or more regions.

We conducted a further comparison of expert selection before and after introducing the flip consistent loss, as illustrated in Fig. 1(c). The flip consistent loss enables the detector to better model the flip semantic consistency of face images. As highlighted in the second row of images, the Vanilla MoE detector typically assigns different experts to process the left and right sides of regions such as the hair, eyes, ears, and cheeks. In contrast, the FC-MoE detector tends to select the same expert for symmetrical facial regions, as highlighted in the third row of images.

E. Ablation Study

Hyperparameter in sparse MoE transformer blocks. As shown in Table IV, we vary the number of experts and the value of k in the TOP_k operation within the sparse MoE transformer blocks and then evaluate the performance for FS-MoE-Tiny. The results indicate that the optimal performance is achieved when the number of experts is set to 6 and k is set to 1.

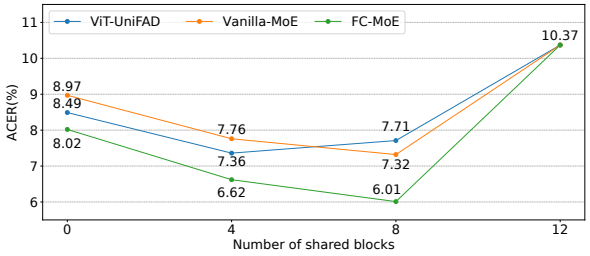

Fig. 5. The relationship between model performance and the number of early shared blocks.

Number of early shared blocks. Fig. 5 shows the performance trend of the multi-branch and MoE detectors as the number of early shared blocks changes. The experiments are conducted on a tiny scale. All three types of detectors exhibit a trend where the ACER first decreases and then increases. This indicates that sharing parameters in low-level layers can enhance model performance, whereas continuing to share parameters in high-level layers leads to performance degradation due to the small intra-class similarities and large inter-class similarities of semantic features. The optimal number of early shared blocks for the multi-branch detector, the vanilla MoE detector, and the FC-MoE detector are 4, 8, and 8, respectively.

Face attack classification. We aim to determine whether the FC-MoE detector is also effective for face attack classification tasks. The results in Table V demonstrate that, when classifying based on the attack categories, sub-categories, and subsets within the attack tree of the GrandFake dataset, our proposed FC-MoE detector consistently outperforms the ViT baseline and the vanilla MoE detector.

V. Conclusion

The data for unified face attack detection (UFAD) is characterized by small intra-class similarities and large inter-class similarities. Dense detectors, which operate within a single feature space, are not well-suited for the UFAD task. While multi-branch detectors improve performance by partitioning the feature space based on attack groups, they suffer from

TABLE IV
Performance of the FS-MoE detector with different hyperparameter settings in sparse MoE transformer blocks.

Number of Experts	k in TOP_k Operation	ACER
6	**1**	**6.01**
6	2	7.52
3	1	6.21
12	1	6.34

TABLE V
Comparison on the face attack classification task.

Method	Catagory	Sub-catagory	Subset
ViT-Tiny	95.55	95.15	93.17
Vanilla-MoE-Tiny	96.18	95.72	93.56
FC-MoE-Tiny (Ours)	**97.53**	**96.31**	**94.32**

low extensibility and application efficiency. We are the first to introduce the sparse mixture-of-experts (MoE) mechanism into the UFAD task. MoE detectors leverage the ability of experts to automatically correlate facial regions during training, thereby partitioning the feature space according to facial regions. To further improve performance, we propose a self-supervised loss function to guide the detector in modeling the flip semantic consistency of facial images. Experimental results demonstrate that the resulting FC-MoE detectors are effective learners for UFAD, achieving performance significantly superior to multi-branch detectors and vanilla MoE detectors while maintaining efficiency similar to dense detectors.

REFERENCES

[1] Debayan Deb, Xiaoming Liu, and Anil K. Jain, "Faceguard: A self-supervised defense against adversarial face images," in *FG*. 2023, pp. 1–8, IEEE.

[2] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain, "On the detection of digital face manipulation," in *CVPR*. 2020, pp. 5780–5789, Computer Vision Foundation / IEEE.

[3] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, "Multi-attentional deepfake detection," in *CVPR*. 2021, pp. 2185–2194, Computer Vision Foundation / IEEE.

[4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *CVPR*. 2022, pp. 18689–18698, IEEE.

[5] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *CVPR*. 2023, pp. 3994–4004, IEEE.

[6] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu, "UCF: uncovering common features for generalizable deepfake detection," in *ICCV*. 2023, pp. 22355–22366, IEEE.

[7] 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*. 2020, pp. 5294–5304, Computer Vision Foundation / IEEE.

[8] 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*. 2022, pp. 4113–4123, IEEE.

[9] Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu, "Rethinking domain generalization for face anti-spoofing: Separability and alignment," in *CVPR*. 2023, pp. 24563–24574, IEEE.

[10] 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*. 2023, pp. 20453–20463, IEEE.

[11] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei, "CFPL-FAS: class free prompt learning for generalizable face anti-spoofing," *CoRR*, vol. abs/2403.14333, 2024.

[12] Debayan Deb, Xiaoming Liu, and Anil K. Jain, "Unified detection of digital and physical face attacks," in *FG*. 2023, pp. 1–8, IEEE.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*. 2016, pp. 770–778, IEEE Computer Society.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*. 2021, OpenReview.net.

[15] 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," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–15, 2024.

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *ICLR (Poster)*, 2015.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR (Poster)*. 2018, OpenReview.net.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *CVPR*. 2016, pp. 2574–2582, IEEE Computer Society.

[19] Debayan Deb, Jianbang Zhang, and Anil K. Jain, "Advfaces: Adversarial face synthesis," in *IJCB*. 2020, pp. 1–10, IEEE.

[20] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li, "Semanticadv: Generating adversarial examples via attribute-conditioned image editing," in *ECCV (14)*. 2020, vol. 12359 of *Lecture Notes in Computer Science*, pp. 19–37, Springer.

[21] Ali Dabouei, Sobhan Soleymani, Jeremy M. Dawson, and Nasser M. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *WACV*. 2019, pp. 1979–1988, IEEE.

[22] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*. 2019, pp. 1–11, IEEE.

[23] 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*. 2019, pp. 3673–3682, Computer Vision Foundation / IEEE.

[24] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*. 2018, pp. 8789–8797, Computer Vision Foundation / IEEE Computer Society.

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*. 2020, pp. 8107–8116, Computer Vision Foundation / IEEE.

[26] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1179–1187.

[27] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *ECCV (12)*. 2020, vol. 12357 of *Lecture Notes in Computer Science*, pp. 70–85, Springer.

[28] Shan Jia, Xin Li, Chuanbo Hu, Guodong Guo, and Zhengquan Xu, "3d face anti-spoofing with factorized bilinear coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4031–4045, 2021.

[29] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 42–55, 2020.

[30] Xiao Guo, Yaojie Liu, Anil K. Jain, and Xiaoming Liu, "Multi-domain learning for updating face anti-spoofing models," in *ECCV (13)*. 2022, vol. 13673 of *Lecture Notes in Computer Science*, pp. 230–249, Springer.

[31] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[32] William Fedus, Barret Zoph, and Noam Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, pp. 120:1–120:39, 2022.

[33] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby, "Scaling vision with sparse mixture of experts," in *NeurIPS*, 2021, pp. 8583–8595.

[34] Jonathan Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Mach. Learn.*, vol. 28, no. 1, pp. 7–39, 1997.

[35] Ltd. Mashang Consumer Finance Co., "Ms-ufad: A large-scale dataset for real-world unifed face attack detection with text descriptions," https://ms-ufad.github.io/, 2024.

[36] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *ICLR (Poster)*. 2017, OpenReview.net.