This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DFD-HR: Generalizable Deepfake Detection via Hierarchical Routing Learning

Jiamu Sun[1], Zhiyuan Yan[2], Ke-Yue Zhang[1], Taiping Yao[1]*, Shouhong Ding[1]
[1]Tencent Youtu Lab, [2]Peking University
{genisun,zkyezhang,taipingyao,ericshding}@tencent.com, zhiyuanyan@stu.pku.edu.cn

Abstract

Developing generalizable deepfake detectors has become increasingly important with the rapid advancement of generative models. Adapting visual foundation models (VFMs), e.g. *CLIP, through parameter-efficient finetuning (PEFT), with only a small subset of parameters updated, has been proven highly effective for generalizable detection. However, the success of "fewer-parameters" training raises an important question: although only a few parameters are tuned, have existing PEFT-based detectors truly exploited the most informative ones while eliminating redundant parameters for better generalization? In this work, we move beyond standard PEFT by proposing a joint optimization strategy that operates at both the layer and token levels. Since latent features across layers capture different semantic abstractions and tokens within the same layer convey varied forgery cues, we propose integrating both layer-level and token-level routing to maximize representational synergy. Specifically, at the layer level, we introduce "Early Layer Pruning", an adaptive truncation mechanism that enables the model to adaptively learn distinct forward depths for different types of instances. At the token level, "Token Selection" is guided by the Spearman rank loss to filter tokens irrelevant to forgery learning, enabling the model to focus on the most discriminative cues. Furthermore, a unified MoE architecture is applied that encourages diversity and thus reduces the potential model's overfitting to specific forgery types. Extensive benchmarking results demonstrate the effectiveness of our designs and show the superior performance of our method over existing state-of-the-arts. Project: https://dfd-hr.github.io/.*

1. Introduction

DeepFake Detection (DFD) [2, 12, 39, 52, 58, 77] aims to protect facial privacy from digital-editing technology, *e.g.*, face-swapping [13, 35, 40], face-reenactment [31, 72, 73], and entire image synthesis [30]. However, as a real-time attack and defense task, deepfake detectors [63, 65, 71] suffer from overfitting to specific forgery artifacts, which severely hinders generalization for practical applications.

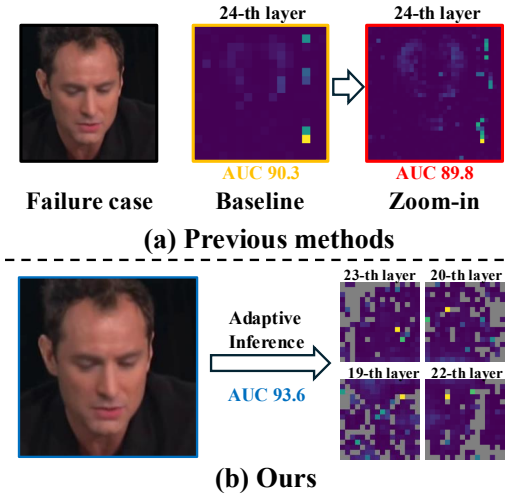
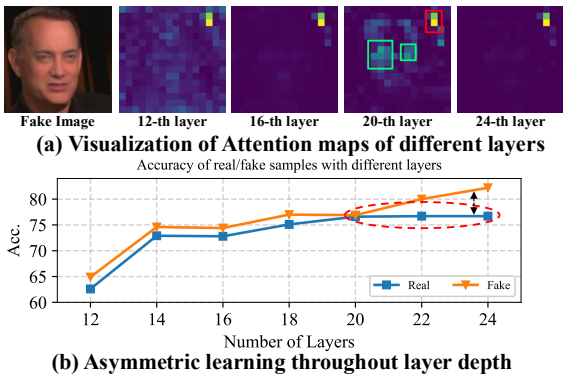


Figure 1. **Comparison of DFD-HR with previous methods.** Failure cases in forgery detection often exhibit no obvious artifacts at a glance. Attention is frequently drawn to incorrect regions even zoom-in. Our method overcome this by layer and token routing. Face images shown in figure are from Celeb-DF-v2 [40].

To address this issue, previous methods focus on the discovery of more general fake patterns, which can be divided into different directions: data augmentation [7, 37, 38, 65, 92], frequency clues [19, 45, 50, 58, 78], disentanglement learning [41, 43, 79, 86], reconstruction learning [75], and contrastive learning [11, 51]. Benefit from large-scale pretraining [21, 54, 57, 60, 89], researchers have attempted to leverage vision foundation models (VFMs) for DFD [11, 64, 83] through parameter-efficient finetuning (PEFT), achieving notable improvement in generalization.

However, despite their success, existing PEFT-based detectors still raise a fundamental question: *have these methods truly exploited the most informative parameters while eliminating redundant ones to achieve optimal generalization?* We argue that **simply updating a small number of parameters does not guarantee effective adaptation if the selected parameters or features are suboptimal**. In-

*Corresponding Author



(a) Visualization of Attention maps of different layers

(b) Asymmetric learning throughout layer depth

Figure 2. **Illustration of importance for hierarchical routing.** Sub-figure (a) shows that the attention maps differ among layers and convey varied forgery cues influenced by spurious tokens. Sub-figure (b) indicates that real samples converge rapidly, while forged ones remain challenging and are progressively uncovered at deeper layers, proving the necessity of Early Layer Pruning. Face images shown in figure are from Celeb-DF-v2 [40].

deed, directly transferring CLIP [60] to DFD often results in insensitivity to subtle forgery cues that are easily overlooked, as shown in Fig. 1(a). "Zoom-in" strategies [48] partially alleviate this issue by focusing on local regions, but they still fail to capture forgery-related semantics consistently. Since latent features across layers capture different semantic abstractions (Fig. 2(a)), which are asymmetric needed for real and fake samples (Fig. 2(b)). In addition, tokens within the same layer convey varied forgery cues influenced by spurious tokens (red region in Fig. 2(a)).

To this end, we move beyond standard PEFT and introduce a hierarchical routing strategy operating at both the layer and token levels. **At the layer level**, we propose Early Layer Pruning. This adaptive truncation mechanism allows the model to learn distinct forward depths for different instances, where real samples typically terminate earlier while fake samples propagate deeper and differ across forgery methods (Fig. 5(a)). **At the token level**, we introduce Token Selection guided by the Spearman rank loss to filter out tokens irrelevant to forgery learning, enabling the model to focus on the most discriminative cues (Fig. 6). Moreover, we employ a **unified Mixture-of-Experts (MoE) architecture** that encourages diversity and thus helps to discriminate real and fake samples (Fig. 5(b)). By integrating layer- and token-level routing within a unified sparse architecture, the model dynamically learns to concentrate learning resources on the most informative representations.

To validate DFD-HR, we apply DFD-HR to three representative VFMs, *i.e.* BEITv2 [57], SigLIP [88] and CLIP [59], and conduct extensive experiments on the popular benchmarks. Experimental results show that DFD-HR can greatly exceed the SoTA methods, *e.g.* +2.3% average Video-level AUC gains on cross-dataset setting and +3.8% average Video-level AUC gains on cross-method setting. More importantly, we present a comprehensive analysis of hierarchical routing for its effectiveness of decoupling real and fake samples at both the layer and token levels.

Overall, the contributions of this paper are three-fold:

- We explicitly investigate the roles of parameters at both layer- and token- level for generalizable forgery learning.
- We present the first attempt of hierarchical routing network named DFD-HR with Early Layer Pruning, Token Selection and Expert Routing.
- DFD-HR achieves significant performance gains on cross-dataset and cross-method settings over SoTAs.

2. Related Work

Generalizable Deepfake Detectors. Deepfake detection (DFD) [2, 12, 39, 52, 58, 77] aims to detect synthetic media that can convincingly mimic the appearance and conduct of real people. Early DFD generalizable detectors [5, 37, 45, 50, 52, 53, 65, 81] mainly focus on discovering more general fake patterns, which can be roughly divided into four categories, *e.g.*, frequency-based methods [19, 45, 50, 58, 78], blend-based methods [65, 85], reconstruction-based methods [75, 90] and disentanglement-based methods [9, 43, 49, 79, 86]. However, these patterns suffer from limited generalization due to the rapid development of AI generative technologies [18, 22, 62, 84]. More recently, the remarkable transferability and generalization capabilities of large-scale pre-trained foundation models have attracted more attention. FFTG [70] proposes an accurate visual-linguistic dataset for the adaptation of CLIP [59] and MLLMs [48]. VLF-FFD [56] uses a novel external detector fusion solution for MLLM-enhanced face forgery detection. DeepShield [4] balances local sensitivity and global generalization to improve robustness on unseen forgeries.

Parameter Efficient Modeling. Parameter Efficient Fine-tuning (PEFT) of Vision Transformer (ViT) has made great progress in transfer learning [20, 47, 68], and can be broadly categorized into three approaches, *i.e.* prompt-tuning [6, 32, 42], Adapter-tuning [11, 34, 64, 85], and LoRA-based methods [83, 91]. In particular, Forensics Adapter [11] designs an adapter for CLIP to learn the unique blending boundaries of forged faces. Effort [83] employs SVD to LoRA [24] to construct two orthogonal subspaces of foundation models for efficient fake learning. More recently, sparsely-gated Mixture of Experts networks (MoEs) have demonstrated excellent scalability in Natural Language Processing [26] and Computer Vision [61]. MoE-FFD [34] updates lightweight mixture of LoRA and Adapter layers to scale the model's capacity and smartly select optimal forgery experts. More recently, token-level methods [55, 76], such as UDD [17], overcome position bias and content bias via token-level shuffling and mixing.

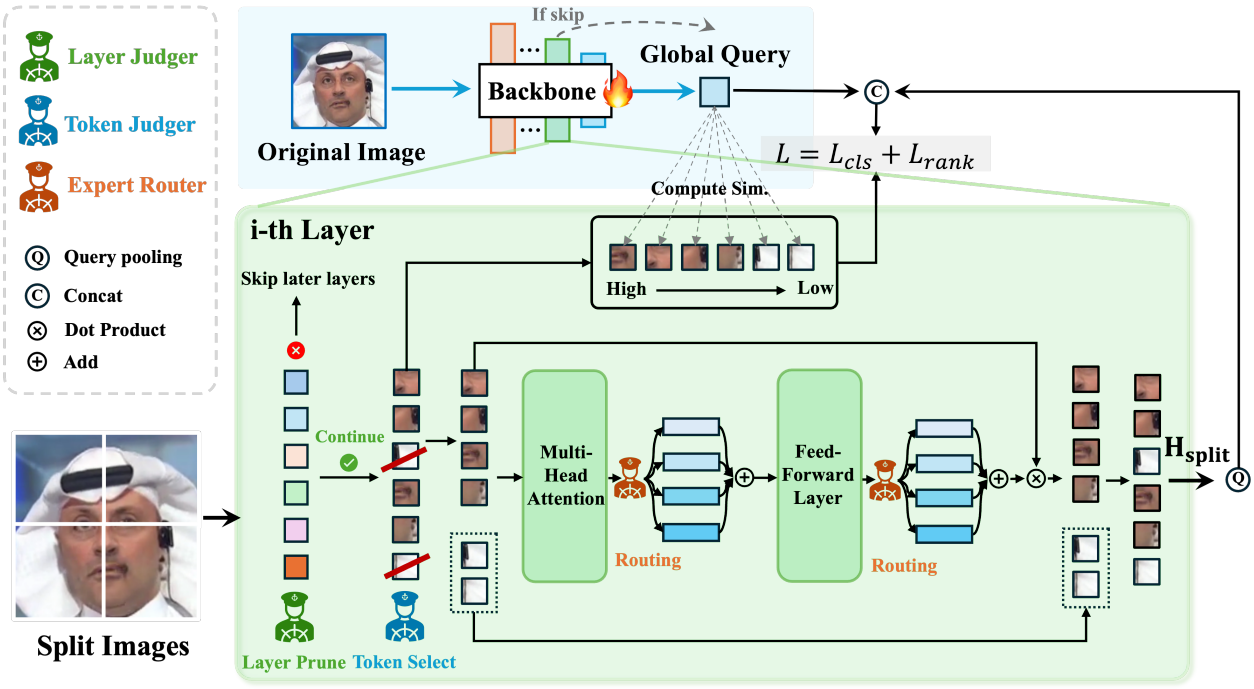


Figure 3. **The overall framework of DFD-HR.** DFD-HR explicitly exploits the most informative parameters via Hierarchical Routing. At the layer level, Early Layer Pruning enables the model to adaptively learn distinct forward depths for different instances. At the token level, Token Selection is guided by the Spearman rank loss to filter tokens irrelevant to forgery learning. Moreover, Expert Routing encourages diverse expert activation, which helps to discriminate real and fake samples. Face images shown in figure are from Celeb-DF-v2 [40].

3. Method

The overall pipeline of the proposed approach is illustrated in Fig. 3, designed to explicitly investigate the roles of parameters via hierarchical routing. Specifically, at the layer level, "Early Layer Pruning" is an adaptive truncation mechanism that enables distinct forward depths among instances. At the token level, "Token Selection" is used to filter forgery-irrelevant tokens and "Expert Routing" further gives differentiated weights between real and fake samples, thereby improving generalizable detection performance.

3.1. Hierarchical Routing (HR)

We first propose Hierarchical Routing to tackle the insensitivity to subtle forgery artifacts of foundation models. The most straightforward approach is to apply the "Zoom-in" strategies in Multi-Modal Large Language Model [48]. However, directly transferring this paradigm to DFD partially alleviate this issue by focusing on local regions, but still fail to capture forgery-related semantics consistently. Therefore, we divide the original image into splits and leverage global query information to help the model attend to facial forgeries. Given a set of labeled data $D = \{(\mathcal{I}_i), \mathcal{Y}_i\}_{i=1}^{N}$, the target of HR can be formulated as

$$\begin{aligned} \mathcal{I}^s &= \text{Split}(\text{Upsample}(\mathcal{I})), \\ &\min \mathcal{L}(\theta; \{(\mathcal{I}, \mathcal{I}^s), \mathcal{Y}\}). \end{aligned} \tag{1}$$

Here, $\mathcal{L}$ is the objective function. θ denotes HR proposed to exploit the most informative parameters. $\mathcal{I}$, $\mathcal{I}^s$ and $\mathcal{Y}$ denote images, split images and labels, respectively.

3.1.1. Early Layer Pruning

To distinguish among instances, we first explore the relationships between layers. Considering that the latent features across layers capture different semantic abstractions (Fig. 2(a)) which are asymmetric needed for real and fake samples (Fig. 2(b)), it suggests that layer routing can be optimized for better interactions. Based on these observations, we further propose an adaptive truncation for layers.

Given input hidden states $\mathbf{H}_i \in \mathbb{R}^{B\times L\times D}$ of i-th layer, we first pass [CLS] token $h_i^{cls} \in \mathbb{R}^{B\times D}$ to layer judge:

$$\text{logits} = \mathbf{W_2}(\text{ReLU}(\mathbf{W_1}(h_i^{cls}))), \tag{2}$$

where $\mathrm{W}_1 \in \mathbb{R}^{D\times d}$ and $\mathrm{W}_2 \in \mathbb{R}^{d\times 1}$ are learnable weights. Then, we use Gumbel-Sigmoid [27] to perform hard routing during training. First, the Gumbel noise distribution is converted from uniform distribution sampling:

$$\begin{aligned} U &\sim \text{Uniform}(0, 1), \\ g &= -\log(-\log(U + \epsilon) + \epsilon), \end{aligned} \tag{3}$$

where U and g are the uniform and Gumbel noise distribution variables, respectively. ϵ is a numerical stability con-

stant. The corresponding probability can be formulated as:

$$p_{\text{soft}} = \sigma\left(\frac{\text{logits} + g}{\tau}\right), \\ p_{\text{hard}} = \mathbb{I}\left[p_{\text{soft}} > 0.5\right], \tag{4}$$

where $\sigma(\cdot)$ is Sigmoid function, τ is a temperature parameter. The soft routing probability $p_{\text{soft}} \in [0, 1]$ and the corresponding hard one $p_{\text{hard}} \in \{0, 1\}$.

During training, the output probs p is then defined as:

$$p = \text{sg}\left(p_{\text{hard}}\right) - \text{sg}\left(p_{\text{soft}}\right) + p_{\text{soft}}, \tag{5}$$

where $\text{sg}(\cdot)$ is stop-gradient function and the gradient backpropagation is from p_{soft} during training. And p_{hard} is used directly during inference. The output hidden states $\mathbf{H}_{i+1} \in \mathbb{R}^{B\times L\times D}$ of i-th layer can be formulated as:

$$\mathbf{H}_{i+1} = p \cdot \text{Layer}_i\left(\mathbf{H}_i\right) + (1 - p) \cdot \mathbf{H}_i. \tag{6}$$

To encourage the difference between instances, once $p = 0$ in i-th layer, that instance will not participate in all subsequent layers and is directly used for classification.

3.1.2. Token Selection

Previous work [17] shows vision foundation models exhibit semantic bias in DFD. We attribute it to the fact that tokens within the same layer convey varied forgery cues influenced by spurious tokens (Fig. 2(a)). Based on the observation, "Token Selection" is proposed to filter tokens irrelevant to forgery learning via global query prior.

To distinguish the importance of each token for forgery detection, we first introduce token judge to score all tokens:

$$\text{scores}_i = \mathbf{W}_{\text{t}}([h_i^{cls}, h_i^{patch}]), \tag{7}$$

where $\text{scores}_i \in \mathbb{R}^{B\times L}$ is the forgery scores of all tokens in i-th layer, and $\mathbf{W}_{\text{t}} \in \mathbb{R}^{D\times 1}$ is the learnable weight. For each layer, only the TopK score tokens will be selected:

$$k = \max(1, \lfloor r \times L \rfloor), \\ \mathbf{H}_i^{select} = \mathbf{H}_i[\text{TopK}\left(\text{scores}_i, k\right)], \tag{8}$$

where r and k are the proportion and number of selected tokens, L is the number of all tokens in i-th layer, and $\mathbf{H}_i^{select} \in \mathbb{R}^{B\times k\times D}$ is the selected tokens scored by token judge. Then, filtered tokens $\mathbf{H}_i^{filter} \in \mathbb{R}^{B\times(L-k)\times D}$ are directly propagated to the output of i-th layer.

Based on the prior knowledge that forgery cues appear only in face regions, We impose a rank constraint on the token judge to better guide token selection. Given the final output [CLS] token h_{global}^{cls} of the image $\mathcal{I}$ which represents global face information, we calculate cosine similarity between h_{global}^{cls} and input hidden states H_i of split images $\mathcal{I}^s$:

$$\text{cosine}_{\text{i}} = \frac{\mathbf{H}_i}{\|\mathbf{H}_i\|_2} \cdot \left(\frac{h_{global}^{cls}}{\|h_{global}^{cls}\|_2}\right)^{\text{T}}, \tag{9}$$

where $\|\cdot\|_2$ denotes the L2 norm, $\text{cosine}_{\text{i}} \in \mathbb{R}^{B\times L}$ denotes the output similarity obtained by the product of $\mathbf{H}_i \in \mathbb{R}^{B\times L\times D}$ and $h_{global}^{cls} \in \mathbb{R}^{B\times 1\times D}$. To encourage the monotonic relationship between scores_i and cosine_{i}, we define rank loss using the Spearman correlation coefficient [66]:

$$\rho_{\text{spearman}} = \frac{\text{Cov}(\text{rank}(\text{scores}_{\text{i}}), \text{rank}(\text{cosine}_{\text{i}}))}{\sigma_{\text{rank}(\text{scores}_{\text{i}})} \cdot \sigma_{\text{rank}(\text{cosine}_{\text{i}})}}, \\ \mathcal{L}_{rank} = 1 - \rho_{\text{spearman}}. \tag{10}$$

Here, $\text{rank}(\cdot)$ calculates token ranks according to corresponding metrics, $\text{Cov}(\cdot, \cdot)$ is the covariance of two ranked sequences, and $\sigma_{(\cdot)}$ is the standard deviation function. The output $\rho_{\text{spearman}} \in [-1, 1]$ and $\mathcal{L}_{rank}$ is optimized for consistent correlation. Equipped with the well-learned token judge, the selected tokens $\mathbf{H}_i^{select}$ are passed to Transformer block Layer_i to get the output $\mathbf{H}_{i+1}^{select}$. Then, the output are reweighted according to scores of selected tokens:

$$\text{scores}_i^{select} = \text{Softmax}(\text{TopK}\left(\text{scores}_i, k\right)), \\ \mathbf{H}_{i+1}^{select} = \mathbf{H}_{i+1}^{select} \odot (1 + \text{scores}_i^{select}), \tag{11}$$

where Layer_i consists of Multi-Head Attention (MHA) and MLP function in Eq. (12). $\text{scores}_i^{select} \in \mathbb{R}^{B\times k}$ denotes scores of selected tokens and $\mathbf{H}_{i+1}^{select} \in \mathbb{R}^{B\times k\times D}$ is the final output of selected tokens. Then, the final output $\mathbf{H}_{i+1}$ of i-th layer is the merge of $\mathbf{H}_{i+1}^{select}$ and filtered tokens $\mathbf{H}_i^{filter}$, maintaining their original relative order.

3.1.3. Token-level Expert Routing

To mitigate conflicts of forgery learning of different types in the i-th layer, we insert Adapter after Multi-Head Attention (MHA) and MLP via expert routing. Given selected token $\mathbf{H}_i^{select}$ ($\mathbf{X}$ for simplify), Layer_i can be formulated as:

$$\mathbf{X}_{\text{attn}} = \mathbf{X} + \text{MOA}(\text{Attention}(\text{LN}_1(\mathbf{X}))), \\ \mathbf{X}_{\text{out}} = \mathbf{X}_{\text{attn}} + \text{MOA}(\text{MLP}(\text{LN}_2(\mathbf{X}_{\text{attn}}))). \tag{12}$$

Here, $\text{MOA}(\cdot)$ is our Mixture of Adapter design, which is composed of gate network with noise, Adapter Tuning and routing mixture. First, gate network can be formulated as:

$$\mathbf{L} = \mathbf{X}\mathbf{W}_g, \\ \mathbf{G} = \text{Softmax}(\mathbf{L} + \boldsymbol{\epsilon} \odot \text{Softplus}(\mathbf{L})), \tag{13}$$

where N is the expert number and $\mathbf{W}_g \in \mathbb{R}^{D\times N}$ denotes the learnable gating parameter. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is added only during training and $\mathbf{G} \in \mathbb{R}^{B\times k\times N}$ denotes gating weights for each selected tokens. The final output of MOA is:

$$\text{Adapter}(\mathbf{X}) = \mathbf{W_2}(\text{ReLU}(\mathbf{W_1}(\mathbf{X})), \\ \text{MOA}(\mathbf{X}) = \mathbf{X} + \sum_{i=1}^{N} \mathbf{G}_i \odot \text{Adapter}_i(\mathbf{X}). \tag{14}$$

3.2. Multi-Scale Fusion (MSF)

Following D^3 [87], images at multiple scales exhibit distinct representations that complement one another. This motivates us to fuse between split images and original images, jointly learning of shared forgery artifacts.

$$\mathbf{h}_{split} = \mathrm{MHA}(\mathbf{Q}_{\text{token}}, \mathbf{H}_{\text{split}}, \mathbf{H}_{\text{split}}), \tag{15}$$

where $\mathbf{Q}_{\text{token}} \in \mathbb{R}^{B\times 1\times D}$ is the learnable query token, $\mathbf{H}_{\text{split}} \in \mathbb{R}^{B\times n\times D}$ denotes the output hidden states of split images, and n is the number of splits per image. Finally, $\mathbf{h}_{split}$ and $\mathbf{h}^{cls}_{global}$ are concatenated for classification.

Overall, the complete procedure of DFD-HR is described in Algorithm 1. The training objective combines the classification loss (*i.e.*, cross-entropy loss for binary classification) and rank loss (*i.e.*, the Spearman correlation coefficient for better Token Selection):

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \frac{1}{m}\sum_{i}^{m} \mathcal{L}^i_{\text{rank}}. \tag{16}$$

where λ_1 is hyperparameter that balances the importance of the regularization term, and m represents the number of Hierarchical Routing inserted into the branch.

Algorithm 1 *DFD-HR* Approach Algorithm

Input: Training data $\mathcal{D} = \{\mathcal{I}, \mathcal{Y}\}$; Pretrained CLIP model θ; Total block number M; HR inserted block number m; Token Selection ratio r; Hyperparameter λ_1
Output: Layer judge $\{\theta^i_l\}^{M-1}_{i=M-m}$; Token judge $\{\theta^i_t\}^{M-1}_{i=M-m}$; Expert network $\{\theta^i_e\}^{M}_{i=1}$

```
► Training Loop
for each epoch do
    for each batch in 𝒟 do
        ▷ Original Image Forward Pass
        ◇ Same as Split Images Forward Pass shown bellow
        ◇ Extract h^cls_global of last hidden states
        ▷ Split Images Forward Pass
        ◇ Split images by Eq. (1)
        for each Layer_i do
            if i >= M − m then
                ◇ Forward Layer judge by Eq. (2)
                ◇ Early Layer Pruning by Eq. (6)
                if p = 0 then
                    Break
                end if
                ◇ Forward Token judge by Eq. (7)
                ◇ Compute ℒ_rank using h^cls_global by Eq. (10)
                ◇ Token Selection by Eq. (8)
            end if
            ◇ Forward Expert network by Eq. (12)
        end for
        ▷ Multi-Scale Fusion
        ◇ Fuse the final output by Eq. (15)
        ▷ Total Loss
        ◇ Compute classification loss ℒ_cls
        ◇ ℒ = ℒ_cls + λ1 ℒ_rank
        ▷ Backward Pass and Optimization
        ◇ Update θ_l, θ_t, θ_e using gradient descent
    end for
end for
return Updated weight of θ_l, θ_t, θ_e
```

4. Experiment

4.1. Datasets and Metric

Following previous works [11, 82, 83], we adopt two widely used protocols for evaluation: **Protocol-1:** cross-dataset evaluation and **Protocol-2:** cross-manipulation evaluation within the FF++ data domain. For **Protocol-1**, we conduct evaluations by training the models on FaceForensics++ (FF++) [63] and testing them on other seven deepfake detection datasets: Celeb-DF-v2 (CDF-v2) [40], DeepfakeDetection (DFD) [15], Deepfake Detection Challenge (DFDC) [14], the preview version of DFDC (DFDCP) [16], DeeperForensics (DFo) [28], WildDeepfake (WDF) [94], and FFIW [93]. Note that FF++ has three different compression versions and we adopt the c23 version for training all methods following most existing works [81]. For **Protocol-2**, we evaluate the models on the latest deepfake dataset DF40 [82], which contains the forgery data generated within the FF++ domain, ensuring the fake methods different while the data domains remain unchanged. For fair comparison with other works [11, 44, 65, 82, 83], both **Video-level and Frame-level Area Under the Curve (AUC)** are reported, which measure the quality of deepfake detectors. We compute the average model's output probabilities of each video to obtain the video-level AUC.

4.2. Experimental Settings

Following the pre-processing and training pipeline of DeepfakeBench [80], we utilize CLIP ViT-L/14 [59] as our default vision backbone. We use full finetuning with learning rate of $1e-6$ as our baseline. Additionally, following [65], we sample 8 frames from each video for training with 16 batch size and 32 frames for inference with 32 batch size. DFD-HR is optimized by Adam [33] using the fixed learning rate of $1e-4$. We also use data augmentations following Effort [83]. For better performance, We insert HR into the last 4 layer of the backbone and equip other layers with expert designs. The ratio of token selection is set to 75% and λ_1 is 0.1. The expert number for each MOA is 4 and the hidden dimension of Adapter is set to 256.

4.3. Experimental Results

4.3.1. Generalization Evaluation

Comparison with competitive detectors. We first compare our DFD-HR with state-of-the-arts on video-level AUC in Tab. 1. DFD-HR obtains obvious gains on all settings, *i.e.*, +2.3% on cross-dataset and +3.8% on cross-

Table 1. **Benchmarking Results of Cross-dataset Evaluations (Protocol-1) and Cross-method Evaluations (Protocol-2)**. All detectors are trained on FF++_c23 [63] and evaluated on other fake data **(Video-level AUC)**. All the results are cited from [10, 80, 81, 83].

Methods	Cross-dataset Evaluation								Cross-method Evaluation								
	CDF-v2	DFD	DFDC	DFDCP	DFo	WDF	FFIW	Avg.	UniFace	BleFace	MobSwap	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg.
F3Net [58]	0.789	0.844	0.718	0.749	0.730	0.728	0.649	0.743	0.809	0.808	0.867	0.494	0.717	0.845	0.757	0.674	0.746
SPSL [45]	0.799	0.871	0.724	0.770	0.723	0.702	0.794	0.769	0.747	0.748	0.885	0.514	0.666	0.812	0.643	0.665	0.710
SRM [50]	0.840	0.885	0.695	0.728	0.722	0.702	0.794	0.767	0.749	0.704	0.779	0.704	0.659	0.772	0.793	0.694	0.732
CORE [53]	0.809	0.882	0.721	0.720	0.765	0.724	0.710	0.762	0.871	0.843	0.959	0.679	0.774	0.958	0.855	0.724	0.833
RECCE [5]	0.823	0.891	0.696	0.734	0.784	0.756	0.711	0.779	0.898	0.832	0.925	0.683	0.848	0.949	0.848	0.768	0.844
SLADD [7]	0.837	0.904	0.772	0.756	0.800	0.690	0.683	0.777	0.878	0.882	0.954	0.765	0.825	0.943	0.879	0.794	0.865
SBI [65]	0.886	0.827	0.717	0.848	0.899	0.703	0.866	0.821	0.724	0.891	0.952	0.750	0.594	0.803	0.712	0.701	0.766
UCF [79]	0.837	0.867	0.742	0.770	0.808	0.774	0.697	0.785	0.831	0.827	0.950	0.731	0.862	0.937	0.809	0.647	0.824
IID [25]	0.838	0.939	0.700	0.689	0.808	0.666	0.762	0.789	0.839	0.789	0.888	0.766	0.844	0.927	0.789	0.644	0.811
LSDA [81]	0.875	0.881	0.701	0.812	0.768	0.797	0.724	0.794	0.872	0.875	0.930	0.694	0.721	0.939	0.855	0.793	0.835
ProDet [10]	0.926	0.901	0.707	0.828	0.879	0.781	0.751	0.828	0.908	0.929	0.975	0.771	0.747	0.928	0.837	0.844	0.867
CDFA [44]	0.938	0.954	0.830	0.881	0.973	0.796	0.777	0.878	0.762	0.756	0.823	0.631	0.803	0.942	0.772	0.757	0.781
Effort [83]	0.956	0.965	0.843	**0.909**	0.977	0.848	0.921	0.917	0.962	0.873	0.953	0.983	0.926	0.957	0.936	0.926	0.940
DFD-HR (Ours)	**0.960**	**0.980**	**0.865**	0.901	**0.997**	**0.907**	**0.968**	**0.940**	**0.988**	**0.956**	**0.986**	**0.993**	**0.970**	**0.984**	**0.981**	**0.970**	**0.978**

Table 2. **Cross-dataset evaluation results (Frame-level AUC)**. All methods are trained on FF++ and evaluated on other datasets. All the results are cited from [11]. The **best** results are indicated in bold and the second-best results are underlined.

Method	Venue	CDF-v2	DFDC	DFD	WDF	Avg.
Xception [63]	ICCV'19	0.737	0.708	0.816	0.662	0.731
EfficientB4 [71]	ICML'19	0.749	0.696	0.815	0.610	0.718
F3Net [58]	AAAI'20	0.735	0.702	0.798	0.677	0.728
X-ray [37]	CVPR'20	0.679	0.633	0.766	-	-
FFD [12]	CVPR'20	0.744	0.703	0.802	-	-
SPSL [45]	CVPR'21	0.765	0.704	0.812	-	-
SRM [50]	CVPR'21	0.755	0.700	0.812	-	-
Recce [5]	CVPR'22	0.732	0.713	0.812	-	-
SBI [65]	CVPR'22	0.813	-	0.774	0.672	-
UCF [79]	ICCV'23	0.753	0.719	0.807	-	-
ED [1]	AAAI'24	0.864	0.721	-	-	-
CFM [46]	TIFS'24	0.828	-	0.915	0.784	-
FoCus [74]	TIFS'24	0.720	0.669	-	0.722	-
LSDA [81]	CVPR'24	0.830	0.736	0.880	-	-
DiffusionFake [69]	NeurIPS'24	0.805	-	0.904	0.801	-
MoE-FFD [34]	ArXiV'25	0.867	-	0.904	0.806	-
Dual-Adapter [64]	IJCV'25	0.717	0.727	-	-	-
UDD [17]	AAAI'25	0.869	0.758	0.910	-	-
Effort [83]	ICML'25	0.901	0.798	0.923	-	-
FFTG [70]	CVPR'25	0.832	-	0.948	0.851	-
ForAda [11]	CVPR'25	0.900	**0.843**	0.933	0.803	0.870
DFD-HR (Ours)	-	**0.910**	**0.843**	**0.953**	**0.856**	**0.891**

Table 3. **Ablation studies on Hierarchical Routing (HR) and Multi-Scale Fusion (MSF) using Video-level AUC (%).**

Ours		CDF-v2	FaceShifter	Avg.
HR	MSF			
×	×	91.8$_{+0.0}$	88.8$_{+0.0}$	90.3$_{+0.0}$
✓	×	95.0$_{+3.2}$	90.7$_{+1.9}$	92.9$_{+2.6}$
×	✓	92.9$_{+1.1}$	90.1$_{+1.3}$	91.5$_{+1.2}$
✓	✓	**96.0**$_{+4.2}$	**91.2**$_{+2.4}$	**93.6**$_{+3.3}$

Table 4. **Ablation studies on Early Layer Pruning, Token Selection and Expert Routing using Video-level AUC (%).**

Hierarchical Routing			CDF-v2	FaceShifter	Avg.
ELP	TS	ER			
×	×	×	91.8$_{+0.0}$	88.8$_{+0.0}$	90.3$_{+0.0}$
✓	×	×	93.4$_{+1.6}$	89.2$_{+0.4}$	91.3$_{+1.0}$
✓	✓	×	94.3$_{+2.5}$	90.4$_{+1.6}$	92.4$_{+2.1}$
✓	✓	✓	**95.0**$_{+3.2}$	**90.7**$_{+1.9}$	**92.9**$_{+2.6}$

method evaluation compared to Effort [83]. We further compare DFD-HR with other methods on frame-level AUC in Tab. 2. DFD-HR also performs best on all four datasets and gains +2.1% average AUC over ForAda [11]. Besides, DFD-HR outperforms existing Adapter-based methods (MoE-FFD [34], Dual-Adapter [64], Effort [83], ForAda [11]) and token-level methods (UDD [17]). More importantly, DFD-HR is the first hierarchical routing model, *i.e.*, the number of layers is reduced by up to 17%, where the number of tokens is reduced by 25%.

4.3.2. Ablation Study

The impact of Hierarchical Routing and Multi-Scale Fusion. In Table 3, we first ablate Hierarchical Routing (HR) on cross-method evaluation. We can see that the routing metrics of both layer and token level can bring obvious performance gains on DFD datasets, *i.e.*, + 3.2% on CDF-v2 and + 1.9% on FaceShifter. Such gains also validate the necessity of hierarchical routing via explicit modeling. Then, Multi-Scale Fusion can also improve 0.7% average video-level AUC, further learning shared forgery artifacts.

The impact of components in Hierarchical Routing. In Tab. 4, we ablate the effectiveness of components in HR. At the layer level, Early Layer Pruning brings a significant improvement compared to our baseline, *e.g.*, +1.6% video-level AUC on CDF-v2. Early Layer Tuning learns distinct forward depths for different instances for better interactions. Moreover, at the token level, Token Selection further gives +1.1% improvement on average video-level AUC, suggesting that spurious tokens are filtered, enabling the model to focus on the most discriminative cues. Lastly, Expert Routing decouples forgery learning between real and fake samples, *e.g.*, +0.7% video AUC on CDF-v2. These results well validate the effectiveness of components in HR.

Comparison of Token Selection with other designs. Token Selection is compared with other popular designs in Tab. 5. PCA is a feature selection strategy used in LongCLIP [89]. FastV [8] is a popular token selection metric in

Table 5. **Comparison of methods for Token Selection on video-level AUC (%).** Our settings are in blue.

Settings	CDF-v2	FaceShifter	Avg.
Early Layer Pruning	93.4$_{+0.0}$	89.2$_{+0.0}$	91.3$_{+0.0}$
PCA [89]	93.3$_{-0.1}$	90.0$_{+0.8}$	91.7$_{+0.4}$
FastV [8]	93.5$_{+0.1}$	90.1$_{+0.9}$	91.8$_{+0.5}$
Sparse Attention [67]	92.8$_{-0.6}$	89.2$_{+0.0}$	91.0$_{-0.3}$
Token Selection	**94.3**$_{+0.9}$	**90.4**$_{+1.2}$	**92.4**$_{+1.1}$

Table 6. **Comparison of methods for multi-scale complementarity on video-level AUC (%).** Our settings are in blue.

Settings	CDF-v2	FaceShifter	Avg.
Baseline	91.8$_{+0.0}$	88.8$_{+0.0}$	90.3$_{+0.0}$
Split Ensemble	90.8$_{-1.0}$	89.9$_{+1.1}$	90.4$_{+0.1}$
Multi CLS Pooling	91.4$_{-0.4}$	86.5$_{-2.3}$	89.0$_{-1.3}$
D^3 [87]	89.4$_{-2.4}$	88.7$_{-0.1}$	89.1$_{-1.2}$
RINE [36]	83.9$_{-7.9}$	84.4$_{-4.4}$	84.2$_{-6.1}$
MRA [48]	91.4$_{-0.4}$	89.8$_{+1.0}$	90.6$_{+0.3}$
Multi-Scale Fusion	**92.9**$_{+1.1}$	**90.1**$_{+1.3}$	**91.5**$_{+1.2}$

MLLM based on attention weights. Sparse Attention [67] is a sparse and discrete manner searching for non-semantic features used in image manipulation localization. Both approaches can select forgery artifacts but our approach performs better, *e.g.*, +1.0% AUC on CDF-v2 compared to PCA and +1.4% average AUC compared to SparseViT.

Comparison of Multi-Scale Fusion with other designs. Multi-Scale Fusion is also compared with other viable choices. An ensemble of original images and split images is a direct solution during training and inference. Other classic methods such as RINE [36] and Multi CLS Pooling use [CLS] tokens of all layers for multi-scale information. D^3 [87] uses patch shuffling to focus on local forgery. MRA [48] introduces DINOv2 [27] for high-resolution fusion. Our Multi-Scale Fusion is both simple and effective, *e.g.*, +1.5% AUC on CDF-v2 compared to MRA.

Comparison of different PEFT methods. We further compare our proposed DFD-HR with other popular parameter-efficient finetuning (PEFT) methods (Supp.). The first observation is that Adapter [23] and LoRA [24] are better among PEFT variants. Then, LoRA variants like rsLoRA [29] and OLoRA [3] are better than traditional LoRA. Moreover, our proposed DFD-HR outperforms all other PEFT methods on the DFD benchmark.

Compatibility with other vision foundation models. In Tab. 7, following Effort [83], we transfer our multi-level routing method to other visual backbones, *i.e.*, BEIT-v2 [57] and SigLIP [88]. These results show that our method is compatible with various backbones for better generalization, *i.e.*, +4.7% average AUC with BEIT-v2 and +3.3% average AUC with SigLIP compared to Effort.

Table 7. **Compatibility with other vision foundation models** trained on FF++ (c23) and tested on CDF-v2 and SimSwap.

VFMs	#ImgSize	CDF-v2	SimSwap	Avg.
BEIT-v2 [57]	224	0.855	0.821	0.838
Effort [83]	224	0.894	0.850	0.872
Ours	224	**0.912**	**0.925**	**0.919**
SigLIP [88]	256	0.877	0.713	0.795
Effort [83]	256	0.895	0.778	0.867
Ours	256	**0.907**	**0.892**	**0.900**
CLIP [59]	224	0.857	0.860	0.859
Effort [83]	224	0.956	0.926	0.941
Ours	224	**0.960**	**0.970**	**0.965**

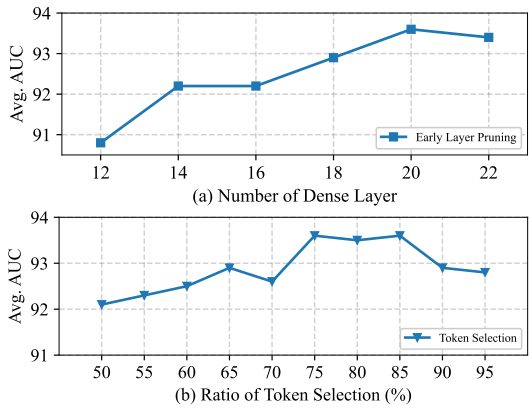


Figure 4. **Hyperparameter of Early Layer Pruning and Token Selection.** The default setting of the number of dense layer $M-m$ and selection ratio are set to 20 and 75% respectively.

Detailed ablations of proposed designs. We ablate detailed designs in our DFD-HR. **For Multi-Scale Fusion**, compared to mean or self-attention pooling, a learnable query aggregates local information of split images better (Tab. 8a). **For Early Layer Pruning**, the layer judges are Gumbel-Sigmoid hard routers on [CLS] token during training, different across layers for better performance (Tab. 8b). **For Token Selection**, the learnable token judge constraint by spearman rank loss is better than using cosine similarity weight to ensemble the score (Tab. 8c). Besides, [CLS] and patch tokens can be treated equally for selection (Supp.). The reweighting metric performed in front of residual of MLP is worse (Supp.). Token judge shared across layers performs worse (Supp.). **For Token-level Expert Routing**, expert design with gate noise during training is better than Top2 hard routing or balance loss constraint (Tab. 8d). **For Hyperparameters**, the more experts will not be better (Tab. 8e), λ_1 set to 0.1 is better (Tab. 8f), the number of sparse layers m and token selection ratio set to 4 and 75% respectively are better (Fig. 4).

4.3.3. Further Analysis

Analysis of Early Layer Pruning. We first analyze the distribution of Early Layer Pruning (Fig. 5 (a)). The first observation is that all distributions asymptotically assume a

case	CDF.	Fsh.	Avg.
Mean-pooling	95.7	91.1	93.4
SA-pooling	92.3	90.4	91.4
Query-select	**96.0**	**91.2**	**93.6**

(a) **Fusion metrics**. A learnable query can aggregate local information better.

case	CDF.	Fsh.	Avg.
Early Pruning	96.0	**91.2**	**93.6**
w.o. Gumbel	80.8	68.7	74.8
w.i. Shared	95.3	88.2	91.8
w.i. GAP	**96.1**	88.6	92.4

(b) **Layer Pruning metrics**. Layer judges differ across layers with Gumbel on [CLS].

case	CDF.	Fsh.	Avg.
Random	94.8	90.8	92.8
Learnable	95.4	91.0	93.2
Learn.+Weight	95.6	90.9	93.3
Learn.+Rank	**96.0**	**91.2**	**93.6**

(c) **Token Selection metrics**. Spearman Rank loss on the learnable token judges is better.

case	CDF.	Fsh.	Avg.
MoE	95.7	91.0	93.4
MoE+Top2	95.9	89.8	92.9
MoE+Balance	94.2	90.5	92.4
MoE+Noise	**96.0**	**91.2**	**93.6**

(d) **Expert Routing designs**. Experts with noise gate perform better.

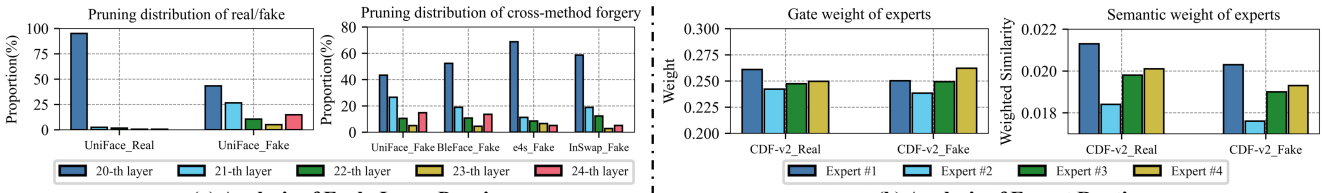
num	CDF.	Fsh.	Avg.
2	94.9	90.5	92.7
4	**96.0**	**91.2**	**93.6**
6	95.3	90.9	93.1
8	95.0	91.0	93.0

(e) **Expert Number**. The larger expert number will not be better.

param	CDF.	Fsh.	Avg.
0.1	**96.0**	**91.2**	**93.6**
0.5	95.4	90.7	93.1
1.0	94.6	90.3	92.5

(f) **Loss balancing**. Rank loss balanced by 0.1 is better.

Table 8. **DFD-HR detailed ablation experiments** with CLIP ViT-L/14 trained on FF++_c23 [63]. We report video-level AUC (%) on CDF-v2 and Faceshifter. If not specified, default settings are marked in blue .



(a) Analysis of Early Layer Pruning **(b) Analysis of Expert Routing**

Figure 5. **Analysis of Hierarchical Routing.** Sub-figure (a) shows that real samples predominantly cut off early while fake ones tend to flow deeper and differ across methods. Sub-figure (b) indicates that Expert #4 and Expert #1 act as the forgery expert and semantic expert, respectively. For all datasets, these two experts dominate the learning between real and fake samples.

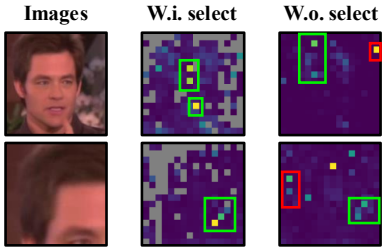


Figure 6. **Visualization of Token Selection of multi-scale images.** Face images shown in figure are from Celeb-DF-v2 [40].

monotonically decreasing shape. In particular, for pruning distribution comparison between real and fake samples, real distribution exhibits a pronounced peak, while the fake one is more uniformly spread. This indicates that real samples predominantly cut off early while fake ones tend to flow deeper and differ across methods, thus providing strong evidence for the necessity of Early Layer Tuning (Fig. 2).

Analysis of Token-level Expert Routing. We further analyze the weight distribution of gate network for all selected tokens (Fig. 5 (b)). Our preliminary observation is that the distributions are largely stable, enabling each expert to specialize without becoming a bottleneck. In addition, an interesting phenomenon is that for all datasets, Expert #1 and Expert #4 dominate the learning between real and fake samples. We further multiple the gate weights of selected tokens with corresponding cosine similarity multiplied by global semantic query h^{cls}_{global}. For all datasets, Expert #1 dominates the learning of semantic information, and we observe no statistically significant difference between the semantic distributions of real and fake samples. Consequently, Expert #4 and Expert #1 function as forgery and semantic expert, respectively, thereby decoupling forgery learning from semantic representation (Fig. 2(a)).

Visualization of Token Selection. To gain more insights into Token Selection, we visualize attention maps of multi-scale images (Fig. 6). We observe that spurious tokens (red region) are filtered by Token Selection with Spearman rank loss for better forgery learning (green region).

5. Conclusion

In this paper, we start from an import question: although only a few parameters are tuned, have existing PEFT-based detectors truly exploited the most informative ones while eliminating redundant parameters for better generalization? Since latent features across layers capture different semantic abstractions and tokens within the same layer convey varied forgery cues, we move beyond standard PEFT by integrating both layer- and token-level routing to maximize representational synergy. Specifically, at the layer level, we introduce "Early Layer Pruning", an adaptive truncation mechanism that enables the model to adaptively learn distinct forward depths for different instances. At the token level, 'Token Selection" is guided by the Spearman rank loss to filter tokens irrelevant to forgery learning. Furthermore, a unified MoE architecture is applied that reduces the potential model's overfitting to specific forgery types. Extensive results demonstrate the effectiveness of our designs and superior performance over existing SoTAs.

References

[1] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 719–728, 2024. 6

[2] Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24709–24719, 2023. 1, 2

[3] Kerim Büyükakyüz. Olora: Orthonormal low-rank adaptation of large language models. *arXiv preprint arXiv:2406.01775*, 2024. 7

[4] Yinqi Cai, Jichang Li, Zhaolun Li, Weikai Chen, Rushi Lan, Xi Xie, Xiaonan Luo, and Guanbin Li. Deepshield: Fortifying deepfake video detection with local and global forgery analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12524–12534, 2025. 2

[5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4113–4122, 2022. 2, 6

[6] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023. 2

[7] 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*, pages 18710–18719, 2022. 1, 6

[8] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 6, 7

[9] Ruoxin Chen, Zhe Wang, Ke-Yue Zhang, Shuang Wu, Jiamu Sun, Shouli Wang, Taiping Yao, and Shouhong Ding. Decoupled data augmentation for improving image classification. *arXiv preprint arXiv:2411.02592*, 2024. 2

[10] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *NeurIPS*, 2024. 6

[11] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19207–19217, 2025. 1, 2, 5, 6

[12] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, 2020. 1, 2, 6

[13] DeepFakes., 2021. www.github.com/deepfakes/faceswap Accessed 2021-04-24. 1

[14] Deepfake detection challenge., 2020. https://www.kaggle.com/c/deepfake-detection-challenge Accessed 2021-04-24. 5

[15] DFD., 2020. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html Accessed 2021-04-24. 5

[16] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 5

[17] Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *AAAI*, 2025. 2, 4, 6

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[19] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *AAAI*, pages 735–743, 2022. 1, 2

[20] Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22995–23005, 2025. 2

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[23] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 7

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 7

[25] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *CVPR*, pages 4490–4499, 2023. 6

[26] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

[27] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3, 7

[28] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020. 5

[29] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023. 7

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[32] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024. 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[34] Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*, 2024. 2, 6

[35] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 1

[36] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024. 7

[37] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 1, 2, 6

[38] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 1

[39] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *WIFS*, 2018. 1, 2

[40] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, 2020. 1, 2, 3, 5, 8

[41] Jiahao Liang, Huafeng Shi, and Weihong Deng. Exploring disentangled content information for face forgery detection. In *ECCV*, pages 128–145. Springer, 2022. 1

[42] Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5262–5270, 2025. 2

[43] Kaiqing Lin, Zhiyuan Yan, Ke-Yue Zhang, Li Hao, Yue Zhou, Yuzhen Lin, Weixiang Li, Taiping Yao, Shouhong Ding, and Bin Li. Guard me if you know me: Protecting specific face-identity from deepfakes. *arXiv preprint arXiv:2505.19582*, 2025. 1, 2

[44] Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. *arXiv preprint arXiv:2409.14444*, 2024. 5, 6

[45] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*, 2021. 1, 2, 6

[46] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE TIFS*, 19:1168–1182, 2023. 6

[47] Gen Luo, Yiyi Zhou, Jiamu Sun, Xiaoshuai Sun, and Rongrong Ji. A survivor in the era of large-scale pretraining: An empirical study of one-stage referring expression comprehension. *IEEE Transactions on Multimedia*, 26:3689–3700, 2023. 2

[48] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 2, 3, 7

[49] Gen Luo, Jiamu Sun, Lei Jin, Yiyi Zhou, Qiang Xu, Rongrong Fu, Xiaoshuai Sun, and Rongrong Ji. Domain incremental learning for object detection. *Pattern Recognition*, 170:111882, 2026. 2

[50] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021. 1, 2, 6

[51] Long Ma, Zhiyuan Yan, Jin Xu, Yize Chen, Qinglang Guo, Zhen Bi, Yong Liao, and Hui Lin. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *arXiv preprint arXiv:2504.04827*, 2025. 1

[52] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024. 1, 2

[53] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *CVPRW*, pages 12–21, 2022. 2, 6

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[55] Chunlei Peng, Xiaoyi Luo, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Semantic token transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2025. 2

[56] Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. Mllm-enhanced face forgery detection: A vision-language fusion solution. *arXiv preprint arXiv:2505.02013*, 2025. 2

[57] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1, 2, 7

[58] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020. 1, 2, 6

[59] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2, 5, 7

[60] 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 *ICML*, pages 8748–8763. PMLR, 2021. 1, 2

[61] 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. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 2

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[63] 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. 1, 5, 6, 8

[64] Rui Shao, Tianxing Wu, Liqiang Nie, and Ziwei Liu. Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision*, 133(6): 3613–3628, 2025. 1, 2, 6

[65] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022. 1, 2, 5, 6

[66] C Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15 (1):72–101, 1904. 4

[67] Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2025. 7

[68] Jiamu Sun, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Refteacher: A strong baseline for semi-supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19144–19154, 2023. 2

[69] Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024. 6

[70] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19576–19586, 2025. 2, 6

[71] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 1, 6

[72] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 1

[73] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *TOG*, 38(4):1–12, 2019. 1

[74] Jiahe Tian, Peng Chen, Cai Yu, Xiaomeng Fu, Xi Wang, Jiao Dai, and Jizhong Han. Learning to discover forgery cues for face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:3814–3828, 2024. 6

[75] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *CVPR*, 2021. 1, 2

[76] Ping Wang, Kunlin Liu, Wenbo Zhou, Hang Zhou, Honggu Liu, Weiming Zhang, and Nenghai Yu. Adt: Anti-deepfake transformer. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2899–2903. IEEE, 2022. 2

[77] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57(3):1–35, 2024. 1, 2

[78] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7278–7287, 2023. 1, 2

[79] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. *arXiv preprint arXiv:2304.13949*, 2023. 1, 2, 6

[80] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *NeurIPS*, pages 4534–4565. Curran Associates, Inc., 2023. 5, 6

[81] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, pages 8984–8994, 2024. 2, 5, 6

[82] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024. 5

[83] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. In *ICML*, 2025. 1, 2, 5, 6, 7

[84] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025. 2

[85] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter

tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12615–12625, 2025. 2

[86] Tianyun Yang, Juan Cao, Qiang Sheng, Lei Li, Jiaqi Ji, Xirong Li, and Sheng Tang. Learning to disentangle gan fingerprint for fake image attribution. *arXiv preprint arXiv:2106.08749*, 2021. 1, 2

[87] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23850–23859, 2025. 5, 7

[88] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 2, 7

[89] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 1, 6, 7

[90] Ke-Yue Zhang, Ruoxin Chen, Jiamu Sun, Jiangming Wang, Hao Yang, Taiping Yao, and Shouhong Ding. A generalizable face security detection method via unified texture and semantic feature framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3190–3198, 2025. 2

[91] Tianshuo Zhang, Li Gao, Siran Peng, Xiangyu Zhu, and Zhen Lei. Devfd: Developmental face forgery detection by learning shared and orthogonal lora subspaces. *arXiv preprint arXiv:2509.19230*, 2025. 2

[92] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV*, 2021. 1

[93] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5778–5788, 2021. 5

[94] 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. 5