Fine-Grained DINO Tuning with Dual Supervision for Face Forgery Detection

Tianxiang Zhang, Peipeng Yu, Zhihua Xia*, Longchen Dai, Xiaoyu Zhou, Hui Gao

College of Cyber Security, Jinan University
{11027ztx, gaohui}@stu.jnu.edu.cn, {ypp865, xia_zhihua}@163.com, longchendai@stu2023.jnu.edu.cn, xiaoyuzhou68@stu2024.jnu.edu.cn

Abstract

The proliferation of sophisticated deepfakes poses significant threats to information integrity. While DINOv2 shows promise for detection, existing fine-tuning approaches treat it as generic binary classification, overlooking distinct artifacts inherent to different deepfake methods. To address this, we propose a DeepFake Fine-Grained Adapter (DFF-Adapter) for DINOv2. Our method incorporates lightweight multi-head LoRA modules into every transformer block, enabling efficient backbone adaptation. DFF-Adapter simultaneously addresses authenticity detection and fine-grained manipulation type classification, where classifying forgery methods enhances artifact sensitivity. We introduce a shared branch propagating fine-grained manipulation cues to the authenticity head. This enables multi-task cooperative optimization, explicitly enhancing authenticity discrimination with manipulation-specific knowledge. Utilizing only 3.5M trainable parameters, our parameter-efficient approach achieves detection accuracy comparable to or even surpassing that of current complex state-of-the-art methods.

Introduction

In recent years, the explosive growth of deep-fake technology has ushered in a revolutionary breakthrough in multimedia content generation, dramatically extending the practical boundaries of visual synthesis (Tolosana et al. 2020).The misuse of deepfake technology raises profound security concerns, eroding trust through identity theft, privacy violations, and misinformation. With the rapid advancement of generative models, synthetic content increasingly bypasses traditional forensic methods, making the development of broadly generalisable detection systems an urgent priority in AI security (Masood et al. 2023; Pei et al. 2024; Dai, Fei, and Huang 2024; Dai et al.).

To address the generalization challenges in deepfake detection, existing approaches can be broadly categorized into four types: detection based on physiological/physical artifacts (Haliassos et al. 2021), noise residual analysis (Jeong et al. 2022), feature consistency analysis (Hu et al. 2022b), and pre-trained large models based detection (Zhuang et al. 2022a). Recent research have increasingly adopted pre-trained large model to extract more expressive and semantically rich feature representations. Fine-tuning methods (Hu et al. 2022a; Page-Caccia et al. 2023) based on the ViT architecture have achieved state-of-the-art performance on multiple public deepfake datasets, validating their broad applicability and remarkable potential in forgery detection tasks, while offering a promising pathway to enhance model generalization (Lin et al. 2025a). Meanwhile, the emergence of recent benchmarks featuring diverse generative models and demographic variations (Lin et al. 2025b) and ILLUSION (Thakral et al. 2025) further emphasizes the need for detectors with stronger generalization and fairness. However, existing detection methods based on pre-trained large models either insert an adapter in the final Transformer block or apply LoRA to fine-tune parameters. These approaches lack task-specific design for deepfake detection and struggle to surpass the upper bound of generalization

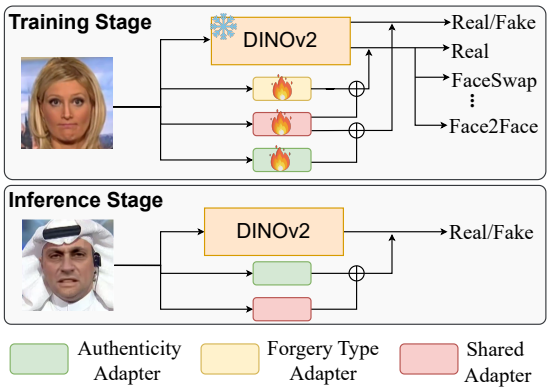


Figure 1: **Training and Inference Stages.** During training, the frozen DINOv2 backbone with the DFF-Adapter is augmented by three adapter heads: authenticity, forgery-type, and shared. The authenticity and forgery-type branches are jointly optimized, while the shared branch captures fine-grained forgery cues and transfers them to the authenticity stream. During inference, only the fused authenticity and shared branches are used for face forgery detection.

*Corresponding author: Zhihua Xia .

performance.

To address this issue, we propose the DeepFake Fine-Grained Adapter (DFF-Adapter), a fine-grained tuning method for deepfake detection based on DINOv2. Specifically, DFF-Adapter is integrated into every Transformer block of DINOv2 to enable the joint optimization of two tasks: authenticity detection and forgery type classification. The architecture comprises three branches: an authenticity detection head, a forgery type classification head, and a shared head adapter. The shared head participates in feature modeling for both tasks, aiming to effectively transfer the fine-grained forgery cues extracted by the forgery type branch to the authenticity branch. This allows the main task to go beyond relying solely on coarse global signals and instead benefit from the detailed artifact patterns learned by the auxiliary task, leading to more generalizable forgery detection (see in Figure. 1). Additionally, we divide the input features of each adapter head into multiple subspaces and introduce a multi-head composition mechanism that enables different subspaces to focus on distinct aspects of forgery-related features, thereby achieving multi-view feature fusion.

Our main contributions are summarized as follows:

- We devise a DeepFake Fine-Grained Adapter architecture that intertwines a shared branch with task-specific branches, enabling the authenticity detector to inherit fine-grained forgery cues from the forgery-type classification task.
- We propose a Forgery-Aware Multi-Head Router that partitions intermediate Transformer features into multiple subspaces and, for each subspace, dynamically routes to a learned top-3 set of LoRA experts. This per-subspace expert optimization fully mines localized forgery artefacts and enables fine-grained, multi-view feature fusion.
- Our method achieves superior generalization in cross-dataset evaluations, outperforming state-of-the-art approaches on multiple challenging benchmarks. Furthermore, in cross-manipulation evaluations conducted on the recently proposed DF40 dataset, it also achieves the best overall performance among all competing methods.

Related Work

Classical Forgery Detection Methods

To improve the generalization ability of face forgery detection models against unseen manipulation methods, researchers have proposed numerous detection methods.Early research concentrated on identifying physiological or physical artifacts introduced during the synthesis process, targeting visually observable inconsistencies directions (Haliassos et al. 2021; Li and Lyu 2018). However, with the rapid advancement of generative models, these artifacts have become increasingly subtle and less discernible. In response, researchers have proposed noise residual analysis methods that exploit subtle anomalies in the frequency domain (Jeong et al. 2022; Wang et al. 2023a; Tan et al. 2024); Methods based on feature consistency focus on detecting latent inconsistencies among internal representations within forged images, capturing contradictions between manipulated and authentic regions (Zhao et al. 2021; Zhai et al. 2023; Hu et al. 2024).In contrast to these traditional techniques, recent methods based on pre-trained large models leverage powerful visual representations learned from diverse, large-scale datasets. These models exhibit superior global perception, semantic abstraction, and robust generalization to unseen forgeries. pre-trained large models offer a unified and extensible framework for face forgery detection across diverse manipulation techniques and data domains.

Pre-trained Large Models for Detection

Methods based on pre-trained large models leverage knowledge distilled from vast image corpora and fine-tune deep networks for the forgery detection task. Prior studies have explored various architectural improvements to the Vision Transformer (Fu et al. 2025a), and some research focuses on uncovering the visual priors embedded in clip (Yu et al. 2025; Yan et al. 2025). By employing lightweight fine-tuning strategies, these methods have effectively enhanced the model's generalization ability in cross-dataset scenarios.

DINO is a visual foundation model based on Vision Transformer (Caron et al. 2021). It employs self-supervised knowledge distillation to learn structure-aware and highly generalizable visual features from large-scale unlabeled image datasets, thereby encoding rich prior knowledge. In deepfake detection tasks, simply appending a linear classifier to DINOv2 can significantly outperform supervised models such as DeiT-III and CLIP during testing (Nguyen, Yamagishi, and Echizen 2024). Compared to CLIP, which operates with dual image-text branches and focuses on semantic alignment, DINOv2 adopts a purely visual self-supervised learning paradigm, preserving fine-grained local textures and geometric structures more effectively. It has shown superior performance in dense prediction tasks such as semantic segmentation and patch matching compared to weakly-supervised models like OpenCLIP (Oquab et al. 2023), making it more sensitive to subtle forgery traces in images.

Current detection methods based on DINOv2 commonly insert adapters only into the final Transformer block or fine-tune merely the last few layers (Kundu, Balachandran, and Roy-Chowdhury 2025; Pellicer, Li, and Angelov 2024). However, such a limited fine-tuning scope constrains the backward propagation of task-specific signals to earlier layers, thereby hindering the modeling of low-level forgery cues and lacking specificity for the deepfake detection task.To address these limitations, we propose the DeepFake Fine-Grained Adapter (DFF-Adapter), which injects both task-specific and shared low-rank adapters throughout the entire DINOv2 backbone. This design is tailored specifically for face forgery detection, aiming to capture the distinctive artifact patterns characteristic of various manipulation methods. By modeling fine-grained forgery cues through task-aware adaptation, DFF-Adapter enhances the model's sensitivity to manipulation-specific traces and effectively improves the generalization capability of deepfake detection approaches.

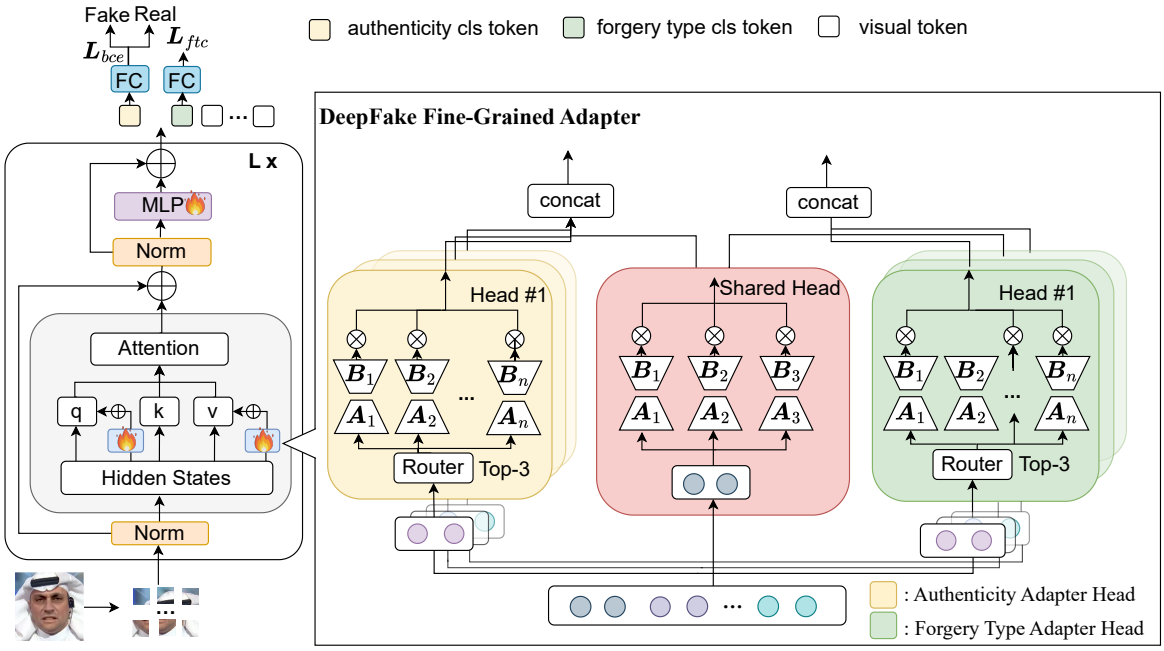


Figure 2: The framework of our method augments a frozen DINOv2 backbone with DFF-Adapters placed in each Transformer block. Each adapter contains three low-rank heads—authenticity, forgery-type, and shared—whose multi-head routers select the top-3 LoRA experts per feature subspace. The shared head transfers fine-grained cues to the authenticity stream. During training, the authenticity and forgery-type CLS tokens are supervised by a binary cross-entropy loss L_{bce} and a multi-class loss L_{ftc}, respectively.

DeepFake Fine-Grained Adapter

Overview

Vision Transformers such as DINOv2 have recently been adopted for forgery detection and already outperform conventional CNNs. Most existing methods treat deepfake detection as a binary classification task and fine-tune only the final layers of the backbone accordingly. However, such methods overlook the fact that different forgery methods often generate distinct artefactual patterns. These method-specific cues are informative yet underutilized, limiting the model's ability to generalize across manipulation types.

To address this limitation, we propose the DeepFake Fine-Grained Adapter (DFF-Adapter), a lightweight, low-rank tuning scheme tailored for deepfake detection. DFF-Adapter jointly optimizes authenticity discrimination and forgery-type classification by injecting task-specific and shared adapters into every Transformer block of a frozen DINOv2 backbone. To further enhance feature diversity and generalization, we design two core modules: a Forgery-Aware Multi-Head Router, which captures diverse forgery artefacts by adaptively routing feature subspaces to specialized experts, and a Shared-Enhanced Task Fusion module, which integrates multi-level task-specific and shared representations to transfer fine-grained forgery cues from the auxiliary branch, thereby enhancing the authenticity detector's sensitivity to manipulation artefacts.

Forgery-Aware Multi-Head Router

To capture the diverse forgery artefacts that manifest in different channel sub-spaces, we propose a lightweight DeepForgery Multi-Head Router(DF-MHR) into our DeepFake Fine-Grained Adapter. DF-MHR partitions the input feature map along the channel axis into h disjoint heads, enabling independent analysis of each sub-space. All head adapters share a pool of N low-rank LoRA adapters; for every head, a router scores the adapters and activates the top-3 that best fit the statistics of that sub-space. This per-sub-space expert allocation equips each head with a specialised adapter set tuned to the artefacts present in its own feature slice. Finally, the head adapter outputs are concatenated into a unified representation that strengthens forgery modelling while introducing only a modest number of parameters.

Given the hidden states of a Transformer block $\mathbf{X} \in \mathbb{R}^{L\times d}$, we split the channel axis into h head adapters, $\mathbf{X}^{(k)} \in \mathbb{R}^{L\times d_h}$ with $d_h = d/h$. The router keeps two routing-logit tables: a task-specific tensor $\mathbf{Z}_{\text{task}} \in \mathbb{R}^{T\times h_t\times N}$ serving the h_t task head adapters, and a global vector $\mathbf{Z}_{\text{shared}} \in \mathbb{R}^{N}$ used by the shared head adapter.

Task-Specific Head Adapter Each task-specific head adapter draws from a shared bank of N low-rank LoRA experts $\{(\mathbf{A}_j, \mathbf{B}_j)\}_{j=1}^{N}$. For a slice width d_h, the two projection matrices in expert j have shapes $d_h \times r_h$ and $r_h \times d_h$, where the per-head rank is $r_h = r/h$. For each task-specific head adapter $k = 1, \ldots, h_t$ under task t, we compute the output

feature f_t^k as follows:

$$f_t^k = \beta \sum_{j \in S_{t,k}} \tilde{g}_{t,k}^{(j)} \mathbf{B}_j \left(\mathbf{A}_j \mathbf{X}^{(k)} \right) \tag{1}$$

$$\tilde{g}_{t,k}^{(j)} = \frac{g_{t,k}^{(j)}}{\sum_{\ell \in S_{t,k}} g_{t,k}^{(\ell)}}, \tag{2}$$

$$S_{t,k} = \text{Top3}\,(N, g_{t,k}) \tag{3}$$

$$g_{t,k} = \sigma\left(\mathbf{Z}_{\text{task}}[t,k] \right) \tag{4}$$

where $\sigma(\cdot)$ denotes the softmax function. $\text{Top}\,3(N,\cdot)$ denotes the set comprising the indices of the three highest routing probabilities among those computed for head adapter k of task t across all N LoRA experts. $g_{t,k}$ is the gating scores with head adapter k of task t. $\mathbf{Z}_{\text{task}}[t,k]$ is the score for the N LoRA experts associated, and β is a scaling coefficient controlling the magnitude of aggregated LoRA expert outputs.

Shared Head Adapter Given the global logit vector $\mathbf{Z}_{\text{shared}} \in \mathbb{R}^N$, the shared head adapter output is computed in a single expression:

$$f_{\text{share}} = \beta \sum_{j=1}^{N} \sigma\big(\mathbf{Z}_{\text{shared}}\big)_j \, \mathbf{B}_j \left(\mathbf{A}_j \mathbf{X}^{(0)} \right), \tag{5}$$

where $\mathbf{X}^{(0)}$ is the feature slice for the shared head adapter and β is a scaling coefficient controlling the magnitude of aggregated LoRA expert outputs.

Shared-Enhanced Task Fusion

To improve generalization under diverse manipulation techniques, this module enhances task-aware representation learning by residually fusing shared and task-specific updates across all Transformer blocks, thereby enabling the authenticity detector to inherit auxiliary task knowledge and capture manipulation-specific cues.

The DFF-Adapter inserts low-rank updates into every query, value, and dense projection of each Transformer block. At block l it yields one shared update $f_{\text{share}}^{(l)}$ and n task-specific updates $\{f_i^{(l),k}\}_{k=1}^n$, where $i \in \{0,1\}$ indexes authenticity and forgery-type tasks, respectively.

Across the L blocks, these updates are added residually to the token stream, so the CLS token f_{bin} can be generally expressed as:

$$f_{\text{bin}} = h_{\text{cls}} + \text{concat}(f_{\text{share}}, f_0^1, \ldots, f_0^n), \tag{6}$$

where h_{cls} is the CLS token from the frozen backbone, and $\text{concat}(\cdot)$ stacks the shared output f_{share} with the n authenticity-specific outputs $\{f_0^k\}$ along the channel dimension.Providing f_{bin} as input to the binary classifier, the binary-cross-entropy loss $\mathcal{L}_{\text{bce}}$ is formulated as follows:

$$\mathcal{L}_{\text{bce}} = \text{BCE}\big(y,\, h_{\text{bce}}(f_{\text{bin}})\big), \tag{7}$$

where $\text{BCE}(\cdot,\cdot)$ denotes the binary cross-entropy loss function that measures the discrepancy between the predicted probability and the ground truth label, $h_{\text{bce}}(\cdot)$ denotes the binary classification layer that converts its input into the predicted probability of the fake class,y is the ground truth label.

Similarly, the CLS token for the forgery-type branch is

$$f_{\text{ftc}} = h_{\text{cls}} + \text{concat}(f_{\text{share}}, f_1^1, \ldots, f_1^n), \tag{8}$$

where $\text{concat}(\cdot)$ stacks the shared output f_{share} with the n forgery-type–specific outputs $\{f_1^k\}$ along the channel dimension.Providing f_{ftc} as input to the forgery type classifier, the forgery type classification loss $\mathcal{L}_{\text{ftc}}$ is formulated as follows:

$$\mathcal{L}_{ftc} = -\sum_{c=1}^{C} y_c \, \log h_{\text{ftc}}(f_{\text{ftc}}), \tag{9}$$

where y_c represents the integer-encoded ground-truth label indicating the correct forgery type among C categories,and the function $h_{\text{ftc}}(\cdot)$ denotes the predicted probability for class c.

Training Details

To fully leverage the complementary strengths of authenticity discrimination and forgery-type recognition, we adopt a dual-branch training strategy where each task is optimized in a separate forward pass within the same mini-batch. This task-wise decoupling avoids gradient interference, allowing each branch to focus on learning task-specific patterns. Meanwhile, the shared adapter facilitates cross-task knowledge transfer, enabling the authenticity branch to benefit from the fine-grained forgery cues captured by the auxiliary classification task.

During each mini-batch, the frozen backbone processes the same input images twice. In the first forward pass, the task flag is set to zero to indicate the authenticity prediction task, and the network produces the authenticity logits. In the second forward pass, the task flag is switched to one to indicate the forgery-type classification task, and the network outputs the forgery-type logits. The total loss is defined as:

$$\mathcal{L} = \lambda_0 \, \mathcal{L}_{\text{bce}} + \lambda_1 \, \mathcal{L}_{\text{ftc}}, \tag{10}$$

where λ_0 and λ_1 are weighting hyperparameters that control the relative contributions of the authenticity and forgery-type branches. This total loss is back-propagated to update only the DeepFake Fine-Grained Adapters and task-specific classifiers, while keeping all backbone parameters frozen.

Experiments

Experimental Settings

Datasets The FaceForensics++ (FF++) (Rossler et al. 2019) dataset includes 1,000 real videos and 4,000 forgery videos across four deepfake categories, which is one of the most widely-used datasets for deepfake detection. CDF-v1, CDF-v2 (Li et al. 2020b), DFDCP (Dolhansky et al. 2020a), DFDC (Dolhansky et al. 2020b), and DF40 (Yan et al. 2024b) are commonly used datasets for evaluating generalization performance in deepfake detection. The datasets employed in this work are collected from DeepFakeBench (Yan et al. 2023b) and DF40, which serve as standardized benchmarks for deepfake detection. To be consistent with the previous deepfake detection approaches, we trained only on c23 compression version of FF++ dataset.

Method	Venue	Intra-dataset	Cross-dataset				
		FF++	CDF-v2	DFDC	CDF-v1	DFDCP	Avg.
Xception (Rossler et al. 2019)	ICCV'19	97.23	81.65	–	80.98	69.90	–
FaceXRay (Li et al. 2020a)	CVPR'20	–	79.50	–	80.58	80.92	–
F3Net (Qian et al. 2020)	ECCV'20	98.20	78.88	71.77	81.11	73.50	76.32
SPSL (Liu et al. 2021)	CVPR'21	96.91	79.86	66.16	85.02	75.86	76.73
RECCE (Cao et al. 2022)	CVPR'22	99.32	82.31	69.58	81.49	71.49	76.21
SBI (Shiohara and Yamasaki 2022)	CVPR'22	99.15	93.82	74.47	93.44	90.95	88.17
UIA-ViT (Zhuang et al. 2022b)	ECCV'22	99.33	82.41	–	86.59	75.80	–
TALL (Xu et al. 2023)	ICCV'23	**99.87**	90.79	76.78	–	–	–
SeeABLE (Larue et al. 2023)	ICCV'23	–	87.3	75.9	–	86.3	–
AltFreezing (Wang et al. 2023b)	CVPR'23	93.81	89.50	64.75	88.48	64.05	76.70
UCF (Yan et al. 2023a)	CVPR'23	98.69	83.73	75.11	86.08	80.50	81.36
LSDA (Yan et al. 2024a)	CVPR'24	–	91.10	77.00	–	–	–
CFM (Luo et al. 2023)	TIFS'24	–	89.65	80.22	–	–	–
InfoClue (Ba et al. 2024)	AAAI'24	–	93.6	75.4	–	90.2	–
LVLM-DFD (Yu et al. 2025)	ICML'25	99.53	<u>94.71</u>	79.12	**97.62**	**91.81**	<u>90.82</u>
VB-StA (Yan et al. 2025)	CVPR'25	–	94.7	<u>84.3</u>	–	90.9	–
ProDet (Cheng et al. 2024)	NIPS'25	–	92.62	71.52	94.48	82.83	85.36
UDD (Fu et al. 2025b)	AAAI'25	–	93.13	81.21	–	88.11	–
Ours	–	<u>99.56</u>	**95.26**	**89.96**	<u>96.14</u>	<u>91.57</u>	**93.23**

Table 1: **Comparison of intra-dataset and cross-dataset performance between our method and existing deepfake detection methods.** The best AUC scores are highlighted in bold, and the second-best scores are underlined. All results are taken from the original publications or LVLM-DFD(Yu et al. 2025).

Evaluation metrics Following existing approaches, we report video-level Area Under the Receiver Operating Characteristic Curve (AUC) for a fair comparison with prior works. The video-level scores are computed by averaging the frame-level predictions across all frames in each video.

Implementation Details We adopt the facebook/dinov2-with-registers-large (Oquab et al. 2023) checkpoint as a frozen backbone and insert DFF-Adapter into the query, value, and dense projections of Transformer blocks. Each DFF-Adapter is configured with rank $r = 16$, scaling factor $\alpha = 32$, a total of 6 LoRA experts, and 4 head adapters.All the images are cropped to 224×224 and the batch size is configured as 24. We train the model for 50 epochs on a single NVIDIA RTX 4090 GPU using the Adam optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. The loss weights in Eq. (10) are set to $\lambda_0 = 10$ and $\lambda_1 = 2$. All experiments follow the default DeepfakeBench settings.

Comparison with SOTA Detection Methods

We compare our approach with several state-of-the-art deepfake detection methods.

Intra-Dataset Evaluation. Following the intra-dataset evaluation protocol proposed in DeepfakeBench, we conduct a comprehensive comparison between our method and existing state-of-the-art deepfake detection approaches on the FF++ dataset. To ensure fairness and consistency, we strictly adhere to the training and testing splits defined in the benchmark. As reported in Table 1 our method achieves a detection AUC of 99.56, demonstrating highly competitive performance and validating its effectiveness under standard evaluation settings.

Cross-Dataset Evaluation. To assess the generalization ability of our method, we perform cross-dataset evaluations following the standardized protocol defined in DeepfakeBench. Specifically, the model is trained on the FF++ dataset (c23 version) and evaluated on several unseen datasets, including CDF-v1, CDF-v2, DFDCP, and DFDC. We report video-level AUC scores to comprehensively measure performance across diverse domains. As shown in Table 1, our method consistently outperforms existing state-of-the-art approaches on multiple challenging benchmarks, achieving an average improvement of 2.41 AUC points.

Cross-Manipulation Evaluation. With the rapid advancement of generative technologies, new forgery techniques continue to emerge at an unprecedented pace. As a result, many existing detection models exhibit suboptimal performance on the recently introduced DF40 dataset. To evaluate the generalization capability of our method against unseen forgery techniques, we train the model on the FF++ dataset and conduct testing on the DF40 benchmark. We evaluate performance across the FF++ domain of DF40, covering three major forgery categories: Face-swapping (FS), Face-reenactment (FR), and Entire Face Synthesis(EFS), spanning a total of 12 diverse manipulation methods. As pre-

		FaceDancer	InSwapper	FSGAN	UniFace
DF40-FS	SBI	78.18	88.52	89.62	89.02
	DINOv2	57.31	64.48	77.24	68.76
	LVLM-DFD	82.97	87.64	93.75	90.61
	Ours	**93.15**	**93.98**	**98.84**	**94.51**
		FOMM	HyperReenact	Wav2Lip	MCNet
DF40-FR	SBI	88.03	65.26	77.06	81.51
	DINOv2	76.67	46.25	56.36	57.40
	LVLM-DFD	93.34	81.56	78.60	83.45
	Ours	**95.36**	**90.28**	**91.65**	**86.29**
		StyleGAN3	StyleGAN-XL	VQGAN	DiT-XL/2
DF40-EFS	SBI	97.91	23.26	91.47	53.59
	DINOv2	85.93	87.70	90.21	74.57
	LVLM-DFD	98.87	100.00	99.99	86.61
	Ours	**99.92**	**100.00**	**100.00**	**98.44**

Table 2: Cross-manipulation evaluation on DF40. All methods are evaluated on different manipulated subsets. Best results are in bold.

sented in Table 2, our method outperforms all existing detection approaches across all categories, demonstrating strong generalization ability in detecting a wide range of unseen forgery techniques.

Analysis

Ablation studies. To evaluate the effectiveness of the proposed Forgery-Aware Multi-Head Router (FAMHR) and Shared-Enhanced Task Fusion (SETF), we perform ablation studies across multiple datasets. As shown in Table 3, introducing FAMHR significantly improves cross-dataset performance by enabling adaptive expert routing over feature subspaces, which enhances the model's ability to capture diverse localized forgery artefacts. Building on this, incorporating SETF consistently boosts performance by transferring fine-grained cues from the auxiliary task to the main authenticity stream. The full model achieves the highest AUC across all benchmarks, demonstrating the complementary benefits of FAMHR and SETF in enhancing generalization under diverse manipulation types.Additional ablation results and technical analyses are available in our extended version posted on arXiv.

Comparison with Fine-Tuning Strategies. We conduct a comprehensive evaluation of the proposed DFF-Adapter for deepfake detection by comparing it with several representative fine tuning strategies : (1) Linear Probing (LP), which freezes the backbone and trains only a linear classifier; (2) LoRA fine-tuning, which inserts low-rank adaptation matrices for efficient updates; (3) MoE-FFD (Kong et al. 2024), a recent approach that integrates LoRA and Adapter layers with a mixture-of-experts routing mechanism for forgery detection; and (4) our method. As reported in Table 4, DFF-Adapter achieves superior performance across diverse experimental settings, demonstrating its effectiveness in capturing manipulation-specific features and strong generalization ability to unseen forgeries.This superior performance is attributed to its ability to fully exploit DINOv2's strength in local feature representation, enabling the model to capture fine-grained forgery artifacts more effectively.

DINOv2	FAMRH	SETF	CDF-v2	DFDC	CDF-v1	DFDCP
✓			61.93	52.83	68.46	55.54
✓	✓		89.56	86.24	85.53	87.11
✓	✓	✓	**95.26**	**89.96**	**96.14**	**91.57**

Table 3: Ablation study results on cross-dataset evaluation.

Fine-Tuning Paradigms	CDF-v2	DFDC	CDF-v1	DFDCP
LP	66.63	72.54	56.45	65.90
LoRa	85.14	79.95	85.14	81.49
MoE-FFD	80.40	76.52	65.75	87.60
ours	**95.26**	**89.96**	**96.14**	**91.57**

Table 4: Comparison with fine-tuning strategies on cross-dataset evaluation.

Identity Constrained Training. As concerns over biometric privacy continue to rise, it is becoming increasingly difficult to obtain large-scale face datasets, particularly in terms of acquiring a sufficient number of identities and training images. To rigorously evaluate the generalization capability of our proposed DFF-Adapter-DINO detector under such data-scarce conditions, we conducted a series of experiments with progressively fewer training identities. Specifically, we randomly sampled 10, 30, and 50 identities from the FF++ training set, which correspond to approximately 1%, 3%, and 5% of the total available identities, respectively. As shown in Table 5, our method achieves decent performance with only 50 identities. Even with as few as 10 identities, it maintains competitive accuracy. These results demonstrate the generalization strength and practical relevance of our method in few-identity settings, indicating that DFF-Adapter successfully harnesses the visual priors of DINOv2 for generalizable forgery detection.

Comparison with Pre-trained Large Models. As shown in Table 6, we conducted a comparative study of representative pre-trained vision models, including CLIP and DINO, and evaluated their performance using AUC and Equal Error Rate (EER). Models marked with * denote those integrated with DFF-Adapter. CLIP ,trained on image–text pairs, emphasizes semantic alignment but lacks sensitivity to fine-grained visual artifacts. DINO improves upon CLIP by capturing local structures through self-distillation. To fully leverage the rich visual priors encoded in DINOv2, we propose DFF-Adapter. Our method jointly optimizes authenticity prediction and forgery-type classification, enabling more discriminative and manipulation-sensitive representations through multi-head routing and shared knowledge transfer. Built on this synergistic design, our method achieves the highest AUCs and lowest EERs across all evaluated benchmarks.

Training Identities	CDF-v2	DFDC	CDF-v1	DFDCP
10	81.97	**82.63**	81.62	82.22
30	81.94	79.16	90.49	81.25
50	**87.28**	82.37	**91.26**	**86.16**

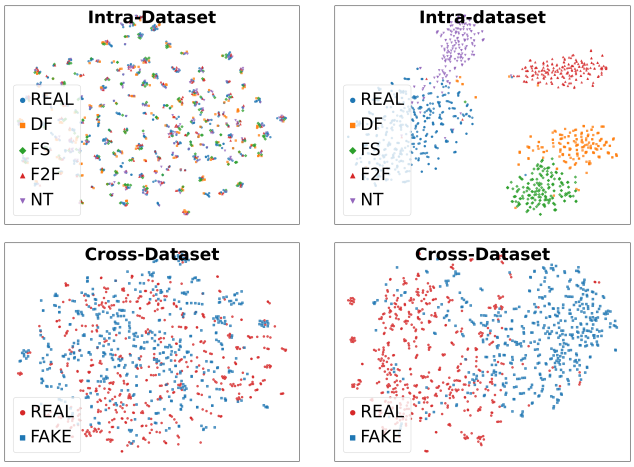
Table 5: Cross-dataset evaluation under identity constrained training.

Model	Arch.	FF++		CDF-v2		DFDC	
		AUC	EER	AUC	EER	AUC	EER
CLIP	ViT-B/16	86.01	22.85	76.27	30.89	77.20	30.06
CLIP	ViT-L/14	91.30	17.14	80.16	29.17	76.76	30.46
CLIP*	ViT-B/16	99.29	**0.71**	85.19	22.47	83.19	24.82
CLIP*	ViT-L/14	97.14	2.57	89.10	15.28	84.75	23.32
DINOv2	ViT-B/16	79.07	28.57	65.38	40.45	68.97	36.58
DINOv2	ViT-L/14	84.84	23.57	66.63	35.39	72.54	33.15
DINOv2*	ViT-B/16	**99.61**	2.32	93.35	14.61	82.00	25.49
DINOv2*	ViT-L/14	99.56	1.79	**95.26**	**12.65**	**89.96**	**17.96**

Table 6: Comparison with pre-trained large models on FF++, CDF-v2, and DFDC. Performance is reported as AUC (%) and EER (%).

t-SNE Feature Visualization. To illustrate the effectiveness of DFF-Adapter in learning discriminative and generalizable representations, we conduct a t-SNE visualization of the learned feature distributions under both intra-dataset and cross-dataset settings. Specifically, we compare the features extracted from a vanilla DINOv2 model (fine-tuned only at the last layer) and our full DFF-Adapter model. The visualization is performed on the FF++ dataset (intra-dataset) and the unseen CDF-v2 dataset (cross-dataset).

As shown in Figure. 3, in the intra-dataset setting (FF++), the DINOv2 baseline fails to distinguish between real and fake samples, as its t-SNE plot exhibits a highly entangled distribution with no clear separation between authentic and manipulated faces. This indicates that simply fine-tuning the final layer is insufficient to extract meaningful features for forgery detection. In contrast, our DFF-Adapter yields well-separated and compact clusters in the feature space—not only achieving clear boundaries between real and fake samples, but also effectively separating different types of forgeries. This demonstrates that the auxiliary forgery-type classification task enables the authenticity branch to benefit from manipulation-specific cues, resulting in more discriminative and semantically structured feature representations.In the cross-dataset visualization on CDF-v2, the difference becomes even more pronounced. The baseline DINOv2 model shows entangled and irregular feature clusters, highlighting its poor generalization to out-of-distribution samples. Meanwhile, our DFF-Adapter maintains clear separation between real and fake samples, and still preserves identifiable structure among different forgery types.



(a) DINOv2 **(b) OURS**

Figure 3: T–SNE visualizations on two datasets: top row shows intra-dataset (FF++), bottom row shows cross-dataset (CDF-v2).

Conclusion

In this work, we introduced DFF-Adapter-DINO, which enriches a frozen DINOv2 backbone with DeepFake Fine-Grained Adapters. By jointly optimizing authenticity detection and forgery-type classification, and by sharing fine-grained cues through a dedicated shared adapter, DFF-Adapter-DINO leverages both global semantics and local artifact patterns.Across five deepfake benchmarks, our method attains the best overall performance and maintains strong generalization under challenging cross-dataset and cross-manipulation scenarios. Taken together, these results demonstrate that DFF-Adapter provides an efficient way to leverage the rich priors of DINOv2 for face forgery detection.

In future work, we will enhance the detection performance of our method on forged data generated by the latest models and design efficient algorithms to tackle real-world scenarios.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grant numbers, U23B2023 and 62472199, Guangdong Key Laboratory of Data Security and Privacy Preserving under Grant 2023B1212060036, the basic and Applied Basic Research Foundation of Guangdong Province (2025A1515011097), and the Outstanding Youth Project of Guangdong Basic and Applied Basic Research Foundation (2023B1515020064). This work is also supported by Engineering Research Center of Trustworthy AI, Ministry of Education.)

References

Ba, Z.; Liu, Q.; Liu, Z.; Wu, S.; Lin, F.; Lu, L.; and Ren, K. 2024. Exposing the deception: Uncovering more forgery

clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 719–728.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4113–4122.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Cheng, J.; Yan, Z.; Zhang, Y.; Luo, Y.; Wang, Z.; and Li, C. 2024. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37: 21979–21998.

Dai, Y.; Fei, J.; and Huang, F. 2024. IDGuard: Robust, General, Identity-Centric POI Proactive Defense Against Face Editing Abuse. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11934–11943. IEEE.

Dai, Y.; Fei, J.; Huang, F.; and Chang, C. H. ???? Robust Secure Swap: Responsible Face Swap With Persons of Interest Redaction and Provenance Traceability. In *Forty-second International Conference on Machine Learning*.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020a. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020b. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Fu, X.; Yan, Z.; Yao, T.; Chen, S.; and Li, X. 2025a. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3040–3048.

Fu, X.; Yan, Z.; Yao, T.; Chen, S.; and Li, X. 2025b. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3040–3048.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5039–5049.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022a. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, J.; Liao, X.; Gao, D.; Tsutsui, S.; Wang, Q.; Qin, Z.; and Shou, M. Z. 2024. Delocate: Detection and localization for deepfake videos with randomly-located tampered traces. *arXiv preprint arXiv:2401.13516*.

Hu, J.; Liao, X.; Liang, J.; Zhou, W.; and Qin, Z. 2022b. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 951–959.

Jeong, Y.; Kim, D.; Ro, Y.; and Choi, J. 2022. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1060–1068.

Kong, C.; Luo, A.; Bao, P.; Yu, Y.; Li, H.; Zheng, Z.; Wang, S.; and Kot, A. C. 2024. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*.

Kundu, R.; Balachandran, A.; and Roy-Chowdhury, A. K. 2025. TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data. *arXiv preprint arXiv:2503.15867*.

Larue, N.; Vu, N.-S.; Struc, V.; Peer, P.; and Christophides, V. 2023. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21011–21021.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5001–5010.

Li, Y.; and Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3207–3216.

Lin, K.; Lin, Y.; Li, W.; Yao, T.; and Li, B. 2025a. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5262–5270.

Lin, L.; Santosh, S.; Wu, M.; Wang, X.; and Hu, S. 2025b. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3503–3515.

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772–781.

Luo, A.; Kong, C.; Huang, J.; Hu, Y.; Kang, X.; and Kot, A. C. 2023. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19: 1168–1182.

Masood, M.; Nawaz, M.; Malik, K. M.; Javed, A.; Irtaza, A.; and Malik, H. 2023. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4): 3974–4026.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2024. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Page-Caccia, L.; Ponti, E. M.; Su, Z.; Pereira, M.; Le Roux, N.; and Sordoni, A. 2023. Multi-head adapter routing for cross-task generalization. *Advances in Neural Information Processing Systems*, 36: 56916–56931.

Pei, G.; Zhang, J.; Hu, M.; Zhang, Z.; Wang, C.; Wu, Y.; Zhai, G.; Yang, J.; Shen, C.; and Tao, D. 2024. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

Pellicer, A. L.; Li, Y.; and Angelov, P. 2024. Pudd: Towards robust multi-modal prototype-based deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3809–3817.

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

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18720–18729.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5052–5060.

Thakral, K.; Ranjan, R.; Singh, A.; Jain, A.; Vatsa, M.; and Singh, R. 2025. ILLUSION: Unveiling truth with a comprehensive multi-modal, multi-lingual deepfake dataset. In *The Thirteenth International Conference on Learning Representations*.

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

Wang, Y.; Yu, K.; Chen, C.; Hu, X.; and Peng, S. 2023a. 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*, 7278–7287.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; and Li, H. 2023b. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4129–4138.

Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; and He, R. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22658–22668.

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024a. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8984–8994.

Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; et al. 2024b. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37: 29387–29434.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023a. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22412–22423.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023b. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*.

Yan, Z.; Zhao, Y.; Chen, S.; Guo, M.; Fu, X.; Yao, T.; Ding, S.; Wu, Y.; and Yuan, L. 2025. 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*, 12615–12625.

Yu, P.; Fei, J.; Gao, H.; Feng, X.; Xia, Z.; and Chang, C. H. 2025. Unlocking the Capabilities of Large Vision-Language Models for Generalizable and Explainable Deepfake Detection. *arXiv preprint arXiv:2503.14853*.

Zhai, Y.; Luan, T.; Doermann, D.; and Yuan, J. 2023. Towards generic image manipulation detection with weakly-supervised self-consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22390–22400.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15023–15033.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022a. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision*, 391–407. Springer.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022b. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision*, 391–407. Springer.