2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Unlocking CLIP for Generalized Deepfake Detection with Dynamic Mixture-of-Adapters

Jialong Liu, Guanghui Li(✉) and Chenglong Dai

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, Jiangsu, China
6223114017@stu.jiangnan.edu.cn
{ghli,chenglongdai}@jiangnan.edu.cn

Abstract. The rapid development of deepfake has raised significant security and ethical concerns, requiring robust and generalizable detection methods. In this work, we propose a novel framework for deepfake detection that leverages the power of large-scale pre-trained vision-language models, specifically the Contrastive Language–Image Pre-training (CLIP) model. Our approach fine-tunes the CLIP image encoder for deepfake detection by introducing a Dynamic Mixture-of-Adapters (MoA) architecture, which consists of multiple lightweight, domain-specific adapter modules that are dynamically activated based on input images. To further improve cross-domain performance, we introduce three auxiliary regularization terms for fine-tuning: attention alignment and similarity regularization, which enforce consistency in feature extraction, and cached domain regularization, which preserves domain-specific prototypes. The proposed framework effectively balances domain-specific adaptation and generalization, addressing critical challenges in generalized deepfake detection. Extensive experiments on benchmark datasets, including FaceForensics++, CelebDF, DFDC, DFD, and DiFF, show that our method performs well in both in-domain and cross-domain deepfake detection tasks.

Keywords: Deepfake Detection, Dynamic Mixture-of-Adapters, Contrastive Language–Image Pre-training (CLIP).

1 Introduction

In recent years, the emergence of Artificially Intelligent Generated Content (AIGC) has garnered significant attention from both academia and industry. Among these technologies, deepfake stands out as one of the most notable advancements in the domain of AIGC, primarily thanks to its ability to manipulate faces with high fidelity and capability. This technique has evolved from traditional image editing methods based on pixel manipulation to modern methods powered by deep learning. Early deepfake generation techniques leveraged models such as Variational Autoencoders (VAE) [1,2] and Generative Adversarial Networks (GAN) [3,4,5]. Although these models achieved substantial improvements over traditional image editing methods, they faced challenges in generating realistic human faces that could deceive the human eye, limiting their large-scale applicability. Recently, diffusion-based models [6,7,8] have further improved the generation of realistic deepfake images and videos.

Such deepfake content, when disseminated on the internet, has caused significant damage to personal reputation and privacy. Furthermore, manipulated content has been used as a propaganda tool. For example, works by Bounareli et al. [9], Hsu et al. [10], and Koujan et al. [11] showcased methods for synthesizing realistic facial reenactments. As deepfake videos become increasingly realistic, deepfake detection techniques have also advanced. However, many recent works perform well on training datasets but experience a sharp decline when confronted with unseen types of deepfake forgeries. The biggest challenge in deepfake detection is improving generalization, which requires the ability of detection methods to perform effectively on unseen or novel deepfake techniques and datasets. Deepfake detection models trained on specific datasets or techniques often struggle to adapt to new types of manipulations, leading to reduced detection accuracy in real-world scenarios. Therefore, developing detectors with good generalization ability is a critical challenge that researchers in the field of deepfake detection must address.

To address these challenges, we propose a novel framework that leverages the power of large-scale pre-trained vision-language models, specifically the Contrastive Language–Image Pre-training (CLIP) model [12], to enhance the generalization of deepfake detection. Specifically, we introduce a dynamic Mixture-of-Adapters (MoA) architecture, which augments CLIP's Vision Transformer (ViT) backbone with lightweight and domain-specific adapter modules. These adapters are dynamically selected and activated based on the domain characteristics of the input, enabling the model to adapt to a wide range of forgery techniques while preserving its pre-trained generalization capabilities. In addition, we incorporate attention alignment and similarity regularization to enforce consistent feature extraction and cached domain regularization mechanism to enhance generalization against domain changes. We leverage the intrinsic capabilities of CLIP to align high-level semantic information with low-level visual features, which is crucial for detecting subtle artifacts in generated fake content. The contributions of this work are summarized as follows:

- A novel deepfake detection framework is proposed that integrates CLIP with a Dynamic Mixture-of-Adapters (MoA) architecture, improving generalization to diverse forgery techniques.
- We integrate attention alignment and similarity regularization to ensure uniform feature extraction, along with a cached domain regularization approach to improve adaptability across varying domains.
- We extensively evaluate the proposed method on multiple benchmark datasets, demonstrating its superior performance in both in-domain and cross-domain deepfake detection tasks.

2 Related Work

2.1 Deepfake generative models

Deepfake synthesis primarily relies on three types of face image generation models: VAE, GAN, and diffusion models. VAE models revolutionized the mapping between the latent space of autoencoders and the RGB feature space. They incorporated high-dimensional Gaussian distributions and enabled image generation through interpolation. GAN models achieved high-quality generation by training a discriminator and generator in an adversarial framework. Diffusion models excelled in generating high-quality images due to their ability to iteratively model data distributions, particularly for tasks involving large-scale complex distributions. Different generative models produce outputs with unique patterns and features. This domain divergence complicates detection, especially as hybrid forgeries—combining outputs from multiple models become more prevalent. These developments necessitate sophisticated and multifaceted forensic approaches to tackle the multi-model and multi-domain challenges of modern deepfake generation, further intensifying the challenges of achieving generalization and robust detection.

2.2 Generalized deepfake detection methods

To enhance the generalization capability of the deepfake detection model, various strategies have been employed to expose it to a broader range of forgery types. Yu et al. [13] utilized data augmentation during training and introduced a common learning method. This involved training a dedicated forgery feature extractor across multiple forgery datasets to improve the model's ability to detect previously unseen forgery techniques. Similarly, Wang et al. [14] explored data augmentation for GAN fingerprint analysis. They devised a method that separates the fingerprints and content of GAN-generated images using an autoencoder-based GAN fingerprint extractor and applies random perturbations to these fingerprints. The resulting manipulated images effectively simulate outputs from diverse GANs, thereby enhancing the detectors' capacity to generalize across different GAN-based synthesis methods.

Other works focus more on forensic features, such as visual consistency, and motion coherence in spatial or frequency domains. Fei et al. [15] addressed feature anomalies in forged faces caused by common blending operations in face forgery techniques. They introduced a weakly supervised second-order local anomaly learning module, which leverages deep feature maps to detect localized anomalies, thereby improving generalization capability. Dai et al. [16] used explicit constraints to distinguish forged regions from genuine ones, constructing horizontal and vertical triplet sets with adjacent vectors to enhance universal feature extraction. Additionally, Luo et al. [17] proposed a detection method based on latent reconstruction errors for diffusion-generated images, incorporating feature refinement techniques to significantly improve generalization performance.

3 Background and Preliminaries

In this section, we provide an overview of the key concepts and methodologies that underpin our proposed framework, including the CLIP model, adapter-based fine-tuning, and domain adaptation techniques.

Contrastive Language–Image Pre-Training (CLIP). The CLIP model [12] represents a significant milestone in multimodal learning, demonstrating the ability to align visual and textual representations through contrastive learning. Trained on a large-scale dataset of 400 million image-text pairs, CLIP learns a shared embedding space where semantically similar images and text are closely aligned. Its architecture consists of two main components: a Vision Transformer (ViT) for image encoding and a transformer-based text encoder for natural language processing. By leveraging this shared embedding space, CLIP achieves state-of-the-art performance on a wide range of zero-shot and few-shot tasks, making it an ideal candidate for transfer learning in domain-specific applications.

The ViT within CLIP processes images by dividing them into non-overlapping patches, which are linearly projected into a fixed-dimensional embedding space. These patch embeddings are then passed through multiple transformer layers, where self-attention mechanisms capture long-range dependencies and contextual information. The pre-trained ViT weights encode rich visual features that are highly transferable, enabling robust performance on downstream tasks even with minimal fine-tuning.

Fine tuning method based on adapters. Adapter modules [18], a lightweight and modular approach to fine-tuning large models, have become an efficient alternative to full model adaptation. Instead of updating all parameters of a pre-trained model, adapters introduce small task-specific layers that are inserted between the existing transformer layers. These layers learn task-specific mappings while the original pre-trained weights remain frozen, preserving the model's generalization capabilities. The modular nature of adapters allows for efficient parameter sharing across tasks, making them particularly well-suited for scenarios involving multiple domains or data distributions. In the context of deepfake detection, adapter-based methods offer several advantages. By isolating task-specific adaptations within dedicated modules, they mitigate the risk of negative transfer, where conflicting domain-specific features degrade overall performance.

Domain adaptation for deepfake detection. Domain adaptation techniques aim to bridge the gap between training and testing distributions, a critical challenge in deepfake detection due to the diversity of forgery techniques and real-world data. Existing methods often rely on data augmentation [19], feature alignment [20], or reconstruction-based approaches [21] to enhance generalization. However, these methods typically assume a fixed domain structure and struggle with unseen or hybrid forgeries that combine characteristics from multiple generative models.

Our proposed framework builds on these foundations by introducing a dynamic domain adaptation mechanism through the MoA architecture. By dynamically routing inputs through domain-specific adapters, the model effectively captures both shared and domain-specific features, enabling robust performance across a wide range of forgery techniques.

4 Methodology

4.1 Motivation

Deepfake detection poses unique challenges due to the subtle and diverse artifacts introduced by different generative models, as well as the rapid evolution of these models. Large pre-trained vision-language models like CLIP offer strong generalization capabilities on many vision tasks, but their direct application to deepfake detection is limited, as they are not optimized to capture the fine-grained discrepancies that distinguish real from generated content. The domain gap between real and deepfake images is often characterized by subtle differences in texture, lighting, and semantic consistency, making it difficult for pre-trained models to adapt effectively. Furthermore, fully fine-tuning these models is computationally expensive and poses a risk of overfitting, particularly in scenarios with limited annotated data. To this end, we propose a framework that enhances CLIP for deepfake detection by introducing a Mixture of Adapters. As shown in Fig. **1**, our method incorporates lightweight adapters designed to capture domain-specific deep-fake features, and learns shared deepfake features to improve the model's generalization ability. Our method is designed to bridge the domain gap and adapt to the diverse characteristics of multiple-domain deepfake, ensuring generalizable detection performance.

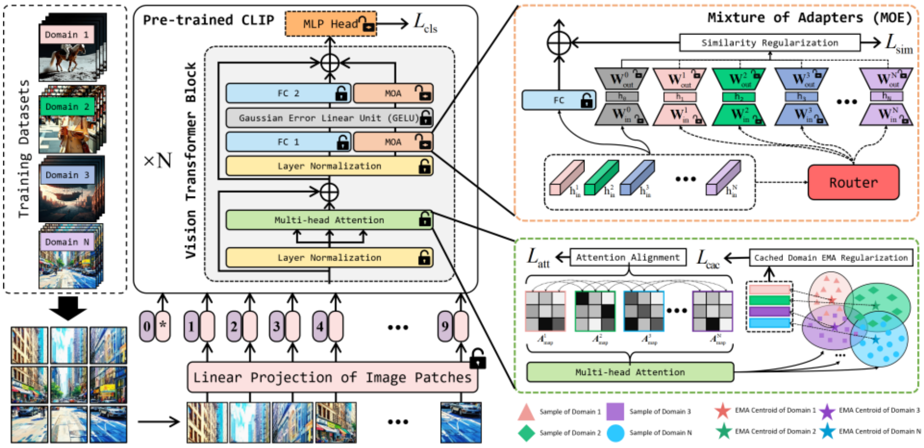

Fig. 1. We propose a CLIP image encoder with Mixture of Adapters (MoA) for multi-domain deepfake detection, where domain-specific adapters extract features and a router coordinates their collaboration.

4.2 Overview

The proposed method extends the pre-trained CLIP model by integrating several key components: Vision Transformer with MoA, attention alignment, cached domain regularization, and similarity regularization. Let the dataset consist of multiple deepfake domains $D_i, i = 1,..,N$, where each domain $D_i = \{x_{i,j}, y_{i,j}\}, j = 1,...,n_i$ contains n_i samples, with input images $x_{i,j} \in \mathbb{R}^{H \times W \times C}$ and corresponding labels $y_{i,j} \in \{0,1\}$. Note that the datasets include the real domain. During training, input deepfake images from multiple domains are divided into patches and passed through a linear projection layer, followed by a ViT backbone enhanced with the MoA. Domain-specific deepfake images are processed by separate adapters, enabling the model to dynamically adapt to different domains while preserving a shared model structure. Additionally, images from all domains are processed by a shared adapter, denoted as Adapter 0.

The training process alternates between two iterative steps as shown in Fig. **2**. In the first step, domain-specific adapters are trained to promote domain-invariant attention patterns and regularize the feature space and encourage domain-invariant and discriminative representations. In the second step, the parameters of the adapters are frozen, and a router composed of an MLP and softmax is trained to dynamically route inputs to the most suitable adapters, improving the model's ability to process multi-domain inputs.

The framework finally produces real/fake predictions by the MLP head for each input image, and the training process optimizes a combined loss function that integrates classification objectives and other regularizations, ensuring robust feature learning and generalization across all domains.

4.3 Vision Transformer with Mixture of Adapters

The ViT in CLIP image encoder serves as the backbone for image feature extraction. An input image $x \in \mathbb{R}^{H \times W \times C}$ is divided into M patches, with each patch of size $P \times P \times C$. These patches are flattened and linearly projected into a d-dimensional embedding space, followed by the addition of positional embeddings. Formally, the patch embeddings are computed as:

$$h_0 = [x_1\mathbf{E}; x_2\mathbf{E}; \ldots; x_M\mathbf{E}] + \mathbf{P} \tag{1}$$

Where $x_i \in \mathbb{R}^{P \times P \times C}$ represents the i-th image patch, $\mathbf{E} \in \mathbb{R}^{(P^2 C) \times d}$ is the learnable projection matrix, and $\mathbf{P} \in \mathbb{R}^{M \times d}$ denotes the positional embeddings. These embeddings are then processed by a stack of L transformer layers, where each layer comprises a Multi-Head Self-Attention (MHSA) module and two Feed-Forward Network (FFN):

$$h_{l+1} = \text{TransformerBlock}(h_l), \quad l = 0, \ldots, L-1 \tag{2}$$

To adapt CLIP to domain-specific deepfake detection, each transformer block is augmented with MoA. MoA integrates multiple lightweight domain-specific adapters into the model, where each adapter is exclusively responsible for processing deepfake images from a specific domain. For a given input hidden state h_{in}, the MoA module processes the features using the adapter corresponding to the input's domain, and the adapter's output is combined with the original FFN output as follows:

$$h_{\text{out}} = h_{\text{FFN}} + \text{Adapter}_k(h_{\text{in}}) \tag{3}$$

where h_{FFN} is the output of the original feed-forward network, and Adapter_k corresponds to the k-th domain-specific adapter. Each adapter is a lightweight feed-forward layer designed to process domain-specific features and is defined as:

$$\text{Adapter}_k(h_{\text{in}}) = h_{\text{in}} \mathbf{W}_{\text{in}}^k \mathbf{W}_{\text{out}}^k \tag{4}$$

where $\mathbf{W}_{\text{in}}^k \in \mathbb{R}^{d \times d_a}$ and $\mathbf{W}_{\text{out}}^k \in \mathbb{R}^{d_a \times d}$ are the learnable weights of the k-th adapter, d_a is the hidden dimension of the adapter. The adapters allow the model to handle the unique characteristics of each deepfake domain while maintaining a shared transformer backbone.

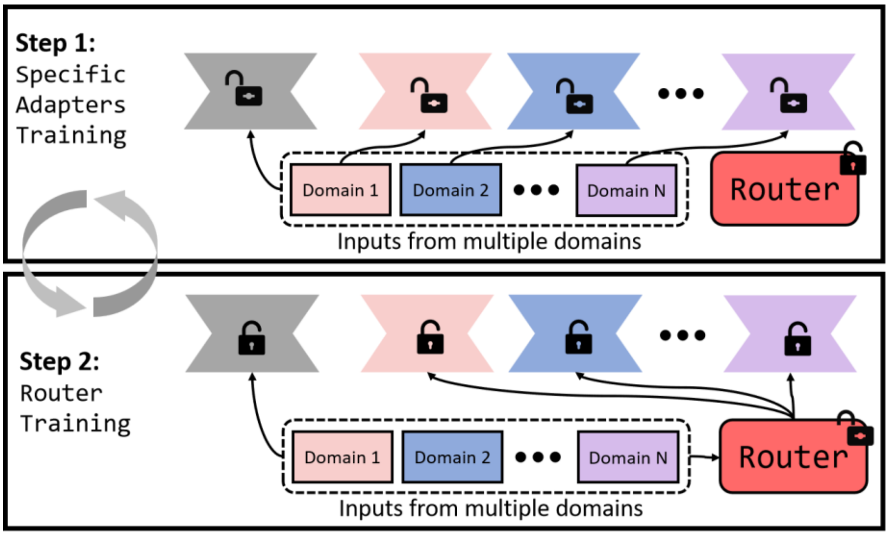

Fig. 2. The two-step training process of the proposed method.

The training process alternates between two iterative steps to ensure effective learning from multiple domains:

Step 1: Adapter Training. In this step, the domain-specific adapters are trained while the rest of the model is frozen. Note that the common adapter takes inputs of all domains. The model learns to capture domain-specific features while preserving generalization across domains.

Step 2: Router Training. After the domain-specific adapters are trained, their parameters are frozen, and the router is optimized to dynamically route input images to the most suitable adapters. The router consists of a lightweight multi-layer perceptron (MLP) followed by a softmax layer, which computes routing weights α_k for each adapter based on the input's hidden state h_{in}. These weights determine the contribution of each adapter to the final output. Specifically, the router computes:

$$\alpha = \text{softmax}(\text{Router}(h_{\text{in}})) \tag{5}$$

where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^{\top}$. The final output of the MoA module is computed as a weighted sum of the adapter outputs:

$$h_{\text{out}} = h_{\text{FFN}} + \text{Adapter}(h_{\text{in}}) + \sum_{k=1}^{N} \alpha_k \cdot \text{Adapter}_k(h_{\text{in}}) \tag{6}$$

The router enables the model to dynamically aggregate domain-specific features from multiple adapters, improving its ability to handle multi-domain inputs efficiently.

4.4 Optimization Objectives

The proposed method employs a unified loss function to ensure accurate classification, domain-invariant representation, and robust cross-domain generalization. The total loss consists of four components: classification, attention alignment, cached domain regularization, and similarity regularization. These components work together to align attention patterns, maintain compact feature spaces, and encourage consistent embeddings across domains.

Classification Loss. To ensure accurate real/fake classification, we employ a standard cross-entropy loss. Let $\hat{y}$ denote the true label (real or fake) and $p(y \mid x)$ denote the predicted probability for input x. The classification loss is defined as:

$$L_{\text{cls}} = -\frac{1}{|D|}\sum_{x \in D} \hat{y} \log p(y \mid x) \tag{7}$$

where D represents inputs. This loss serves as the foundation for distinguishing real and fake images.

Attention Alignment Loss. To encourage consistent feature extraction across domains, we introduce the attention alignment, which aligns the attention maps generated by the model for different domains. Let $A_{\text{map}}^{i} \in \mathbb{R}^{M \times M}$ denote the attention map for domain i, where M is the number of patches. The loss is formulated as:

$$L_{\text{att}} = \sum_{i,j} \| A_{\text{map}}^{i} - A_{\text{map}}^{j} \|_F^2 \tag{8}$$

where $\| \cdot \|_F^2$ denotes the Frobenius norm, and i, j iterate over all domain pairs. By minimizing L_{att}, the model learns to focus on similar regions of the image across domains, promoting domain-invariant attention patterns and improving cross-domain generalization.

Cached Domain Regularization Loss. To enhance domain invariance, we propose cached domain regularization, which maintains a prototype μ_i for each domain i and regularizes the feature space to encourage compact and domain-aligned representations. The prototype is updated using an exponential moving average (EMA):

$$\mu_i^{(t)} = (1-\lambda)\mu_i^{(t-1)} + \lambda \cdot \frac{1}{|D_i|}\sum_{x \in D_i} f(x) \tag{9}$$

where $\mu_i^{(t)}$ is the prototype for domain i at iteration t, D_i is the set of samples in domain i, $f(x)$ represents the model's embedding for input x, and λ is the EMA update rate. The regularization loss is then defined as:

$$L_{\text{cac}} = \sum_{i} \| f(x) - \mu_i \|_2^2 \tag{10}$$

where μ_i corresponds to the prototype of the domain for input x. This loss encourages embeddings to cluster around their respective domain prototypes, reducing domain-specific variability in the feature space.

Similarity Regularization Loss. To further enhance consistency across domains, we introduce a similarity regularization that encourages embeddings from different domains to be similar. For two inputs x_i and x_j from different domains, their embeddings $f(x_i)$ and $f(x_j)$ are aligned through:

$$L_{\text{sim}} = \frac{1}{|P|} \sum_{(x_i, x_j) \in P} \| f(x_i) - f(x_j) \|_2^2 \tag{11}$$

where P is the set of all domain-pair combinations. By minimizing L_{sim}, the model learns to produce consistent embeddings across domains, improving robustness against domain shifts.

The final loss function combines all components:

$$L = \beta_1 L_{\text{cls}} + \beta_2 L_{\text{att}} + \beta_3 L_{\text{cac}} + \beta_4 L_{\text{sim}} \tag{12}$$

where $\beta_1, \beta_2, \beta_3, \beta_4$ are hyperparameters that control the contribution of each loss term. This joint objective ensures accurate classification, domain-invariant attention, compact feature spaces, and consistent embeddings, enabling robust multi-domain deepfake detection.

5 Experiments

5.1 Datasets and Implementation Details

We use FaceForensics++ (FF++) dataset [22] that consists of four subsets: DeepFake (DF), Face2Face (F2F), FaceSwap(FS), and NeuralTextures (NT), and four other datasets: CelebDF (CDF-v1, CDF-v2) [23], Deepfake Detection Challenge (DFDC) [24], DFD [25], and DiFF [26] to evaluate our proposed method.

We employ the CLIP pre-trained ViT as the backbone for feature extraction. The backbone is initialized with publicly available pre-trained weights and remain frozen during training to preserve generalization capabilities. Task-specific modules, including the MoA and the classification head, are updated during training. The MoA modules have a hidden dimension of 256 and use the Gaussian Error Linear Unit (GELU) as the activation function. We optimize the trainable components using the AdamW optimizer with a learning rate of 5×10^{-4}, weight decay of 0.01. A cosine annealing learning rate scheduler with 500 warmup steps and a minimum learning rate of 1×10^{-6} is applied for smooth convergence. Gradient clipping with a maximum norm of 1.0 is used to stabilize training. The training process employs a batch size of 64 for 50 epochs.

The final loss function combines cross-entropy classification loss (L_{cls}, $\beta_1 = 1.0$), attention alignment loss (L_{att}, $\beta_2 = 1.0$), cached domain regularization loss (L_{cac},

$\beta_3 = 0.5$), and similarity regularization loss (L_{sim}, $\beta_4 = 0.1$). The performance of our method is evaluated using Area Under the Receiver Operating Characteristic Curve (AUC). AUC measures the model's ability to distinguish between real and fake inputs independently of the decision threshold by TPR (True Positive Rate) and FPR (False Positive Rate). A higher AUC indicates better discrimination performance.

5.2 Comparative experiments

In the cross-dataset deepfake detection experiments, as shown in Table 1, our method demonstrates superior performance across multiple datasets, significantly outperforming existing methods, including LSDA, SPSL, and FFD. Specifically, our model achieves AUC scores of 0.906, 0.875, 0.783, 0.872, and 0.908 on Celeb-DF v1, Celeb-DF v2, DFDC, DFDC-P, and DFD, respectively. These results surpass the previous best-performing method, LSDA, which achieved AUC scores of 0.867, 0.830, 0.736, 0.815, and 0.880 on the same datasets. This highlights the strong generalization ability of our method, particularly on high-quality deepfake datasets such as Celeb-DF v2 and diverse datasets like DFDC.

Table 1. Performance of cross-domain detection. All methods are trained on FF++ (all 4 subsets) and evaluated on other datasets (AUC).

Method	CDFv1	CDFv2	DFDC	DFDCP	DFD	Avg
CLIP[12]	0.711	0.727	0.728	0.726	0.824	0.744
F3Net[27]	0.777	0.735	0.702	0.735	0.798	0.749
X-ray[28]	0.709	0.679	0.633	0.694	0.766	0.696
FFD[29]	0.784	0.744	0.703	0.743	0.802	0.755
SPSL[30]	0.815	0.765	0.704	0.741	0.812	0.767
SRM[31]	0.793	0.755	0.700	0.741	0.812	0.760
Recce[32]	0.768	0.732	0.713	0.734	0.812	0.752
UCF[33]	0.779	0.753	0.719	0.759	0.807	0.763
LSDA[34]	0.867	0.830	0.736	0.815	0.880	0.826
Ours	**0.906**	**0.875**	**0.783**	**0.872**	**0.908**	**0.868**

Our method achieves an average AUC of 0.868 across all datasets, representing a 4.2% improvement over LSDA (0.826). Other methods, such as SPSL and FFD, exhibit average AUC of 0.767 and 0.755, respectively, showing a notable gap compared to our method. This significant performance gain can be attributed to our model's architecture, which effectively combines pre-trained features with task-specific modules and leverages multiple regularization techniques, to enhance generalization. Moreover, our model achieves an AUC of 0.908 on DFD, demonstrating its robustness in detecting high-resolution manipulated videos.

Table 2 presents the cross-dataset evaluation results on the DiFF dataset, which includes four categories: DiFF-T2I, DiFF-I2I, DiFF-FS, and DiFF-FE, evaluated using AUC (%). Our method consistently surpasses the baselines (CLIP and Xception) in both in-domain and cross-domain testing scenarios, demonstrating superior generalization. For instance, when trained on DiFF-T2I, our method achieves an average AUC of 0.855, surpassing Xception (0.828) and CLIP (0.755). Similarly, training on DiFF-I2I results in an AUC of 0.845, outperforming Xception (0.789) and CLIP (0.741). Our model also exhibits strong cross-category generalization. For example, when trained on DiFF-T2I, it achieves an AUC of 0.897 on DiFF-I2I and 0.773 on DiFF-FS. In category-specific evaluations, our method achieves stronger performance, such as an AUC of 0.977 on DiFF-FE (compared to 0.933 for Xception and 0.915 for CLIP). These results highlight the model's ability to generalize across diverse forgery types while avoiding overfitting. Overall, our method achieves the best average AUC across all training scenarios, demonstrating its generalization. Fig. **3** shows the t-Distributed Stochastic Neighbor Embedding (t-SNE) graphs on different datasets. Although only trained on the FF++ dataset, our model is still able to distinguish between real and various forged datasets, demonstrating its strong generalization ability.

Table 2. Performance of cross-domain detection on DiFF (AUC)

Method	Train Set	Test Set				Avg
		DiFF-T2I	DiFF-I2I	DiFF-FS	DiFF-FE	
CLIP[1]	DiFF-T2I	0.928	0.769	0.637	0.687	0.755
Xception[35]		0.963	0.863	0.744	0.722	0.828
Ours		**0.976**	**0.897**	**0.773**	**0.753**	**0.855**
CLIP[1]	DiFF-I2I	0.726	0.902	0.612	0.722	0.741
Xception[35]		0.769	0.912	0.727	0.750	0.789
Ours		**0.814**	**0.968**	**0.786**	**0.792**	**0.845**
CLIP[1]	DiFF-FS	0.577	0.726	0.902	0.594	0.690
Xception[35]		0.605	0.771	0.929	0.585	0.722
Ours		**0.768**	**0.844**	**0.941**	**0.669**	**0.800**
CLIP[1]	DiFF-FE	0.570	0.599	0.680	0.915	0.691
Xception[35]		0.582	0.595	0.729	0.933	0.715
Ours		**0.765**	**0.691**	**0.738**	**0.977**	**0.798**

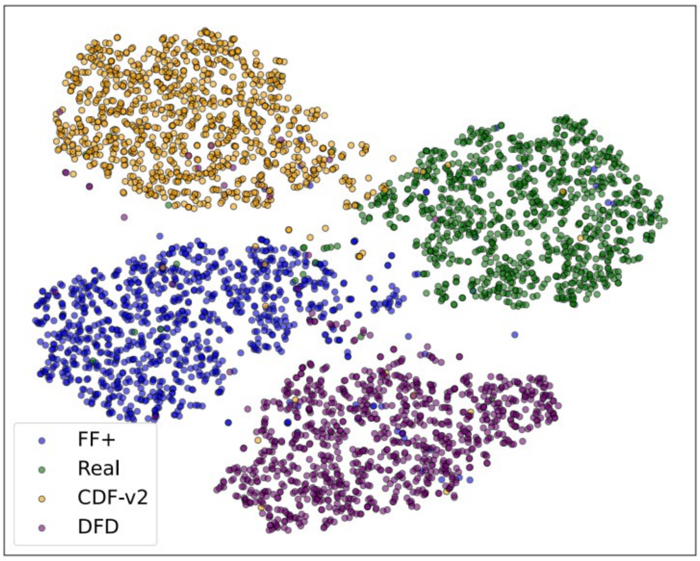

Fig. 3 t-SNE visualization on different datasets.

5.3 Ablation Studies

The effect of Losses. Table 3 highlights the contributions of similarity regularization (L_{sim}), attention alignment (L_{att}), and cached domain regularization (L_{cac}). Without L_{sim} , the AUC on DFDC drops from 0.783 to 0.763, showing that L_{sim} enhances cross-domain feature consistency. Removing L_{att} reduces the AUC on DFD from 0.908 to 0.876, demonstrating its role in aligning attention maps for stable feature extraction. L_{cac} has the most significant impact, as the DFDC AUC decreases from 0.783 to 0.738 without it, indicating its importance in combating domain shifts by enforcing domain-invariant representations. When all three losses are combined, the model achieves the best results across datasets (CDF-v2: 0.875, DFDC: 0.783, DFD: 0.908), proving their complementary effects in improving robustness and generalization.

These results demonstrate that each loss addresses a specific challenge: L_{sim} enhances similarity across domains, L_{att} ensures consistent attention mechanisms, and L_{cac} mitigates domain shifts. These losses form a robust regularization framework that achieves superior performance in cross-domain deepfake detection.

Table 3. Ablation studies on the effect of three losses.

L_{sim}	L_{att}	L_{cac}	CDF-v2	DFDC	DFD
×	√	√	0.866	0.763	0.879
√	×	√	0.854	0.761	0.876
√	√	×	0.870	0.738	0.861
√	√	√	0.875	0.783	0.908

The effect of Different Backbones. Table 4 evaluates the impact of different CLIP ViT backbones. ViT-B/16 achieves the best average AUC (0.855), with strong performance on CDF-v2 (0.875) and DFD (0.908). Its smaller patch size allows it to capture fine-grained details critical for distinguishing real and fake images, while its moderate capacity avoids overfitting. ViT-L/14 achieves the highest DFD AUC (0.915) but underperforms on DFDC (0.779), likely due to overfitting caused by its larger capacity. ViT-B/32 performs the worst (average AUC: 0.820), as its larger patch size reduces its ability to capture subtle local forgery artifacts.

Table 4. The effect of using various backbones.

BackBone	CDF-V2	DFDC	DFD	Avg
CLIP ViT-B/32	0.849	0.750	0.862	0.820
CLIP ViT-B/16	0.875	0.783	0.908	0.855
CLIP ViT-B/14	0.862	0.779	0.915	0.852

The results suggest that ViT-B/16 achieves the optimal balance between feature resolution and generalization. Its ability to capture high-resolution details while maintaining computational efficiency makes it well-suited for cross-domain tasks where both subtle artifacts and diverse forgery types must be detected effectively.

The effect of Adapter Structures. Table 5 investigates the effect of varying the hidden dimension of the Mixture-of-Adapters (MoA) modules. The 128-dimension adapter achieves the best performance, with an average AUC of 0.860 and the highest scores on CDF-v2 (0.875) and DFD (0.908). This configuration balances expressiveness and computational efficiency, enabling the model to effectively capture domain-specific features. The smaller 64-dimension adapter underperforms (average AUC: 0.853), as its limited capacity restricts the model's ability to represent complex forgery patterns. Conversely, the 256-dimension adapter achieves the lowest average AUC (0.827), with substantial overfitting evident on DFDC (AUC: 0.740).

Table 5. The effect of using various adapter structures.

Adapter	CDF-v2	DFDC	DFDCP	DFD	Avg
64	0.869	0.771	0.876	0.897	0.853
128	0.875	0.783	0.872	0.908	0.860
256	0.863	0.740	0.851	0.855	0.827

These findings emphasize that a carefully tuned adapter size is critical. The 128-dimension configuration provides sufficient capacity to capture diverse forgery types without introducing redundancy or overfitting, making it the most effective design choice for cross-domain deepfake detection.

The effect of Number of adapters. Table 6 evaluates the impact of varying the number of adapters (N) in the MoA framework on the model's performance across FF++, CDF-v2, and DFD. As the number of adapters increases, the model's ability to capture diverse domain-specific features improves, leading to better generalization performance. With N=2 (DeepFake + FaceSwap), the model achieves an average AUC of 0.835. Although it performs well on FF++ (0.899), its generalization to cross-domain datasets such as CDF-v2 (0.824) and DFD (0.783) is limited due to insufficient domain coverage. Increasing to N=3 (DeepFake + FaceSwap + NeuralTextures) results in a significant improvement, raising the average AUC to 0.859. This indicates that the inclusion of a third adapter helps the model better adapt to more complex forgery types. Finally, with N=4, where adapters cover all FF++ subsets, the model achieves the best performance with an average AUC of 0.926, excelling on FF++ (0.995), CDF-v2 (0.875), and DFD (0.908).

Table 6. Ablation studies on the effect of the number of adapters.

N	Train set	Method	FF++	CDF-v2	DFD	Avg
2	DF+FS	CLIP	0.899	0.824	0.783	0.835
		Ours	0.907	0.828	0.805	0.847
3	DF+FS+NT	CLIP	0.960	0.821	0.796	0.859
		Ours	0.974	0.843	0.821	0.879
4	FF++(all)	CLIP	0.982	0.727	0.824	0.844
		Ours	0.995	0.875	0.908	0.926

These results demonstrate that increasing the number of adapters allows the model to handle a broader diversity of forgery techniques, improving its ability to extract domain-specific features while maintaining robust generalization across unseen datasets. However, this improvement plateaus once the adapters fully cover the training data, as observed with N=4. Fig. **4** presents the heatmaps of our model on the DiFF dataset, demonstrating its ability to distinguish between forged facial regions and background.

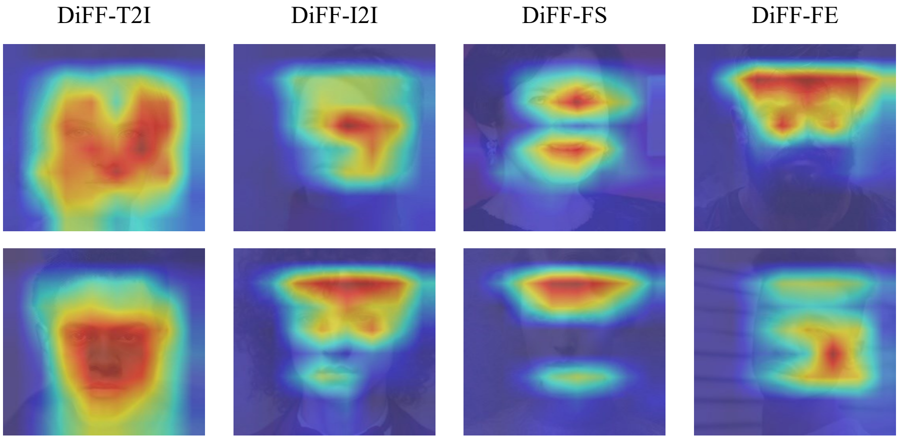

Fig. 4. Heatmaps on the DiFF dataset.

6 Conclusion

In this work, we propose a novel framework that bridges the gap between domain-specific adaptation and cross-domain generalization by enhancing the pre-trained CLIP model with a dynamic MoA architecture. The lightweight adapter modules enable the model to dynamically adapt to diverse forgery techniques while preserving CLIP's inherent generalization capabilities. Extensive evaluations demonstrate that the proposed framework outperforms recent research across multiple datasets, excelling in both in-domain and cross-domain scenarios. Notably, the modular design of MoA achieves this performance with minimal computational overhead, making the approach scalable and practical for real-world applications. Our work provides a significant step forward in digital media forensics and establishes a foundation for future research into robust and generalizable detection frameworks. Our future work will focus on extending the framework to other generative modalities, such as audio and text, while further enhancing computational efficiency to enable deployment in resource-constrained environments.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62372214) and the Science and Technology Program of Suzhou (SGC2021070).

References

1. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv: 1312.6114 (2013)
2. Sohn, K., Lee, H., Yan, X.C.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, pp.3483-3481. MIT (2015)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, Bing., Warde-Farley, D., Ozair, S., Courville, A., Bengio Y.: Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, pp.2672-2680. MIT (2014)
4. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, Timo.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.8110-8119. IEEE (2020)
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4401–4410. IEEE (2019)
6. Blattmann, A., Dockhorn, T., Kulal, S., Wetzstein, G.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv: 2311.15127 (2023)
7. Guo, Y.W., Yang, C.Y., Rao, A.Y., Liang, Z.Y., Wang, Y.H., Qiao, Y., Agrawala, M., Lin, D.H., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv: 2307.04725 (2023)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol.33, pp.6840–6851. Curran Associates (2020)
9. Bounareli, S., Tzelepis, C., Argyriou, V., Patras, I., Tzimiropoulos, G.: Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.7149-7159. IEEE (2023)
10. Hsu, G.S., Tsai, C.H., Wu, H.Y.: Dual-generator face reenactment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.642–650. IEEE (2022)
11. Koujan, M.R., Doukas, M.C., Roussos, A., Zafeiriou, S.: Head2head: Video-based neural head synthesis. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp.16–23. IEEE (2020)
12. Radford, A., Kim, J.W, Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748-8763. PmLR (2021)
13. Yu, P.P., Fei, J.W., Xia, Z.H., Zhou Z.L., Weng J.: Improving generalization by commonality learning in face forgery detection. In: IEEE Transactions on Information Forensics and Security, vol.17, pp.547–558. IEEE (2022)
14. Wang, H.M., Fei, J.W., Dai, Y.S., Leng, L.Y., Xia, Z.H.: General GAN-generated image detection by data augmentation in fingerprint domain. In: 2023 IEEE International Conference on Multimedia and Expo, pp.1187–1192. IEEE (2023)
15. Fei, J.W., Dai, Y.S., Yu, P.P, Shen, T.R., Xia, Z.H., Weng, J.: Learning second-order local anomaly for general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.20270–20280. IEEE (2022)
16. Dai, Y.S., Fei, J.W., Wang, H.M., Xia, Z.H.: Attentional local contrastive learning for face forgery detection. In: International Conference on Artificial Neural Networks, pp.709–721. Springer (2022)
17. Luo, Y.P., Du, J.L., Yan, K., Ding, S.H.: Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.17006–17015. IEEE (2024)

18. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, pp.2790–2799. PMLR (2019)
19. Dong X.Y., Bao, J.M., Chen, D.D, Zhang, T., Zhang, W.M, Yu, N.H., Chen, D., Wen, F., Guo, B.: Protecting celebrities from deepfake with identity consistency transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9468–9478. IEEE (2022)
20. Yu, Y., Liu, X.L., Ni, R.R. Yang, S.Y., Zhao, Yao., Kot, A.C.: PVASS-MDD: Predictive visual-audio alignment self-supervision for multimodal deepfake detection. In: IEEE Transactions on Circuits and Systems for Video Technology, vol.34, pp.6926-6936. IEEE (2023)
21. Shi Z.N., Liu, W.Y., Chen, H.P.: Face reconstruction-based generalized deepfake detection model with residual outlook attention. In: ACM Transactions on Multimedia Computing, Communications and Applications, vol.21, pp.1-19. Association for Computing Machinery (2025)
22. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner. M.: FaceForensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.1–11. IEEE (2019)
23. Li, Y.Z, Yang, X., Sun, P., Qi, H.G., Lyu, S.W.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3207–3216. IEEE (2020)
24. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M.L., Ferrer, C.C.: The Deepfake Detection Challenge (DFDC) dataset. arXiv preprint arXiv: 2006.07397 (2020)
25. Google AI Blog, Contributing data to deepfake detection, https://research.google/blog/contributing-data-to-deepfake-detection-research, last accessed 2025-02-26.
26. Cheng, H., Guo, Y.Y., Wang, T.Y., Nie, L.Q., Kankanhalli, M.: Diffusion facial forgery detection. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp.5939–5948. ACM (2024)
27. Qian, Y.Y., Yin, G.J., Sheng, L., Chen, Z.X., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision, pp.86-103. Springer (2020)
28. Li, L.Z., Bao, J.M., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.N.: Face X-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE (2020)
29. Dang, H., Liu, F., Stehouwer, J., Liu, X.M., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5780-5789. IEEE (2020)
30. Liu, H.G., Li, X.D., Zhou, W.B., Chen, Y.F., He, Y., Xue, H., Zhang, W.M., Yu, N.H.: Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.772-781. IEEE (2021)
31. Luo, Y.C., Zhang, Y., Yan, J.C., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16312-16321. IEEE (2021)
32. Cao, J.Y., Ma, C., Yao, T.P., Chen, S., Ding, S.H., Yang, X.K.: End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4113-4122. IEEE (2022)
33. Yan, Z.Y., Zhang, Y., Fan, Y.B., Wu, B.Y.: UCF: Uncovering common features for generalizable deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.22412-22423. IEEE (2023)
34. Yan, Z.Y., Luo, Y.H., Lyu, S., Liu, Q.S., Wu, B.Y.: Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8984-8994. IEEE (2024)

35. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp.1251-1258. IEEE (2017)