Adapting Pretrained Large-Scale Vision Models for Face Forgery Detection

Lantao Wang and Chao Ma(✉)

MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University,
Shanghai, China
{lantao11,chaoma}@sjtu.edu.cn

Abstract. In the evolving digital realm, generative networks have catalyzed an upsurge in deceptive media, encompassing manipulated facial imagery to tampered text, threatening both personal security and societal stability. While specialized detection networks exist for specific forgery types, their limitations in handling diverse online forgeries and resource constraints necessitate a more holistic approach. This paper presents a pioneering effort to efficiently adapt pre-trained large vision models (LVMs) for the critical task of forgery detection, emphasizing face forgery. Recognizing the inherent challenges in bridging pretraining tasks with forgery detection, we introduce a novel parameter-efficient adaptation strategy. Our investigations highlight the imperative of focusing on detailed, local features to discern forgery indicators. Departing from conventional methods, we propose the Detail-Enhancement Adapter (DE-Adapter), inspired by 'Unsharp Masking'. By leveraging Gaussian convolution kernels and differential operations, the DE-Adapter enhances detailed representations. With our method, we achieved state-of-the-art performance with only 0.3% network adjustment. Especially when the number of training samples is limited, our method far surpasses other methods. Our work also provides a new perspective for the Uni-Vision Large Model, and we call on more fields to design suitable adapting schemes to expand the capabilities of large models instead of redesigning networks from scratch.

Keywords: Face Forgery Detection · Parameter-efficient Tuning · Pretrained Large-Scale Vision Models · Uni-Vision Large Model

1 Introduction

In the modern digital landscape, the development of generative networks [7, 13] has led to a surge in forged media online. Such deceptive content, ranging from manipulated facial imagery [2,12,20,24,43] to tampered text [41], poses significant threats to personal security and societal harmony.

To counter this issue, researchers typically devise dedicated detection networks based on their expertise in specific forgery types. Although effective for specific cases, this method struggles when confronted with the myriad of

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14557, pp. 71–85, 2024.
https://doi.org/10.1007/978-3-031-53302-0_6

forgery media found online. Moreover, the resource-intensive nature of deploying specialized systems for each forgery media type renders this approach impractical. Therefore, we turn our attention to pre-trained large vision models (LVMs) [28,42], which possess vast semantic knowledge learned from large-scale images and perform exceptionally across various downstream tasks [9,33]. We aim to exploit the power of LVMs to develop a unified solution, avoiding the need to deploy multiple specialist networks.

This paper tackles the novel and critical task of *parameter efficiently adapting pre-trained large vision models for forgery detection*, with a particular emphasis on the widely influential face forgery detection task. Given the significant gap between the pre-training task and forgery detection, this challenge is formidable and non-trivial. While direct detection of facial forgery by LVMs without additional training has poor performance, fine-tuning LVMs for this purpose is computationally demanding due to the numerous parameters involved. Although partial fine-tuning offers a potential solution, it presents its own set of difficulties, such as deciding the optimal tunable parameters ratio.

More importantly, our insights into forgery detection indicate that identifying forgery markers requires a keen focus on detailed features that encapsulate local nuances. This perspective stands in stark contrast to existing parameter-efficient tuning (PET) methods for LVMs [3,5,22], which may not fully address the intricacies of forgery. Inspired by the traditional image processing technique 'Unsharp Masking' [38], we propose a lightweight Detail-Enhancement Adapter (DE-Adapter). By employing Gaussian convolution kernels and differential operations, the DE-Adapter captures detail-enhanced representations which are then integrated into the original representation, thereby 'sharpen' detail information. Additionally, our method is versatile and compatible with a wide range of LVM architectures, including both Convolutional networks and Transformers.

The main contributions of this work are summarized as follows:

- We establish a new paradigm, namely solving the forgery detection problem by maximizing the use of pre-trained knowledge from LVMs. Numerous studies have already proven that LVMs possess strong capabilities; hence, we advocate for more domains to design suitable schemes. The idea is to utilize LVMs to solve problems rather than redesigning networks entirely.
- We propose a lightweight Detail-Enhanced Adapter (DE-Adapter). By employing Gaussian convolution kernels and differential operations, the DE-Adapter captures detail-enhanced representations which are then integrated into the original representation, thereby 'sharpen' detail information, the DE-Adapter empowers LVMs to excel at detecting face forgeries.
- Extensive experiments have proven the effectiveness of our method. We can make LVMs achieve state-of-the-art results with only a very small number of parameters trained, especially when the number of training datasets is limited, our method far surpasses other methods. Moreover, our method is applicable to LVMs with different structures, including Convolutional networks and Transformers.

2 Related Work

2.1 Face Forgery Detection

Face forgery detection is a critical task in computer vision and image processing, with the objective of identifying manipulated or forged facial images or videos. Typically, face forgery detection methods rely on human prior knowledge when formulating model architectures. The prior knowledge typically includes visual cues that are crucial to identifying the differences between real and fake images, such as noise statistics [14], spatial domain [25,44], and frequency information [4,21,31,35]. Zhou *et al.* [45] attempted to add a side branch to the image classification backbone, focusing on local noise patterns under the assumption that these patterns differ between real and fake images. Zhao *et al.* [44] redesigned the spatial attention module to enhance the network's capability of extracting subtle forgery traces from local regions. Qian *et al.* [35] and Miao *et al.* [31] proposed frequency-aware models utilizing Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) tools to extract frequency details. While these methods have demonstrated admirable performance, they have not fully harnessed the potential of LVMs. These methods also require extensive data for training from scratch, which can be a limitation.

2.2 Parameter-Efficient Transfer Learning

Over the past few years, transfer learning has surged in prominence, leading to an increased dominance of large-scale foundation models within the field of deep learning. In this context, PET has garnered attention due to its effectiveness and efficiency. Current PET methods can be categorized into three groups. Firstly, Adapter [5] inserts a trainable bottleneck block into the LVMs for downstream tasks adaptation. Secondly, Prompt Learning [22] adds several trainable tokens to the input sequences of LVM blocks. Lastly, Learning weight Decomposition [17,18] breaks down the learning weight into low-rank metrics, training only the low-rank portion. Moreover, researchers have begun applying the PET paradigm to specific task that may not initially seem suitable for LVMs. For example, Pan *et al.* [33] proposed a new Spatio-Temporal Adapter (ST-Adapter) to adapt LVMs, lacking temporal knowledge, to dynamic video content reasoning. Huang *et al.* [19] introduced ensemble adapters in the Vision Transformer (ViT) for robust cross-domain face-anti-spoofing. In these instances, PET not only matches full fine-tuning performance but also enables LVM adaptation to incompatible downstream tasks. In this work, we address the challenging problem of adapting LVMs for face forgery detection, which demands fine-grained local features rather than category-level differences inherent in the original LVMs.

3 Methodology

3.1 Preliminary: Adapter

We first define the face forgery detection task as an image classification problem. The LVM can be separated into a backbone and a classifier. To efficiently use the LVM for downstream tasks, a straightforward approach is to freeze the backbone during training and insert a trainable lightweight module into it. This plugged-in module is designed to learn task-specific features, enhancing the original representations. Essentially, it modulates the original hidden features [15].

Formally, given an image $x \in \mathbb{R}^{H \times W \times 3}$ as input, the backbone extracts the image features as $h_l \in \mathbb{R}^{H' \times W' \times C}$, where h_l is the output of the l-th block in the backbone, Δh_l denotes the task-specific representation, (H, W) and (H', W') represents the size of the input image and the features respectively, C denotes the channel dimension.

$$h_{l+1} \leftarrow h_l + \alpha \cdot \Delta h_l, \tag{1}$$

where α is a scale factor. After that, the classifier is applied on the image feature to output the prediction.

To introduce task-specific representations, the simplest solution is the basic Adapter [5]. When constructing Δh, the adapter module is designed as a bottleneck structure. It includes a down-projection layer with parameters $W_{down} \in \mathbb{R}^{C \times c'}$, an up-projection layer with parameters $W_{up} \in \mathbb{R}^{c' \times C}$ and an activation function. Here, c' represents the middle dimension and satisfies $c' \ll C$. The task-specific feature learning process can be expressed as:

$$\Delta h_l = f(h_l \cdot W_{down}) \cdot W_{up}, \tag{2}$$

where $f(\cdot)$ denotes the activation function.

The basic Adapter [5] is lightweight and flexible, facilitating the design of different structures to achieve optimal task-specific representations. We refer to this architecture in our design, adhering to two principles: (1) parameter efficiency to minimize the cost of parameters used for face forgery detection, and (2) suitability for the task at hand. The design should enable the LVM to effectively extract detail information relevant to face forgery detection.

3.2 The Design of Our Detail-Enhanced Adapter(DE-Adapter)

Given our understanding of face forgery, we recognize the critical importance of the detail information. A pretrained LVM is proficient at extracting category semantic information, enabling high performance on classification tasks like ImageNet [6]. However, without incorporating detailed information, fine-tuning the LVM with existing PET methods does not yield satisfactory results, as validated by Table 2. Thus, introducing the detail information is pivotal to adapting the LVM for face forgery detection with minimal parameter fine-tuning.

By incorporating the detail information, a type of task-specific representation, into the backbone to enhance the original representation, this process resembles the ‘sharpening’ operation in traditional image processing. Drawing inspiration from the ‘Unsharp Masking’ technique [38] used in conventional image processing, we obtain detail-enhanced representations through Gaussian convolution kernels combined with differential operations. This processed detail-enhanced representation is then introduced into the original representation to achieve a ‘sharpening’ effect.

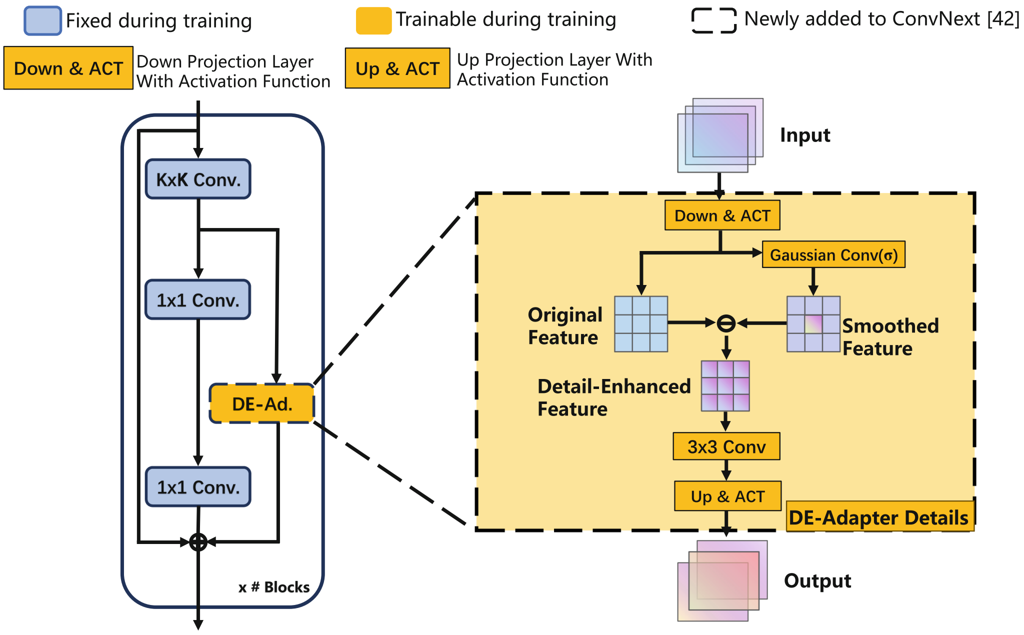

Fig. 1. The architecture of our proposed Detail-Enhanced Adapter. Taking the ConvNext [42] network as an example, we obtain detail-enhanced representations through Gaussian convolution kernels combined with differential operations. This processed detail-enhanced representation is then introduced into the original representation to achieve a 'sharpening' effect.

A classical 'Unsharp Masking' techniques [38] can be described by the equation:

$$\hat{I} = I + \lambda(I - G(I)) \tag{3}$$

where $\hat{I}$ denotes the enhanced image, I denotes the original image, an unsharp mask is represented by $(I - G(I))$ where G denotes a Gaussian filter and an amount coefficient by λ controls the volume of enhancement achieved at the output.

Indeed, the Eq. 3 bears a striking resemblance in Eq. 1. Intuitively, we can treat $(I - G(I))$ part as a task-specific representation. However, in contrast to traditional 'Unsharp Masking' technique, we need to extract detail information of varying scales from the feature map generated by the preceding block to enrich the original representation of the deeper block. This process requires more than just simple operations on the image. Consequently, we cannot merely employ a static Gaussian convolution kernel but must dynamically adjust the kernel based on the depth of the block.

We design a circular concentric Gaussian Convolution (GC) based on the one-dimensional (1D) Gaussian function:

$$f_{1d}(d) = A \cdot exp(-\frac{(d-\mu)^2}{2\sigma^2}) \tag{4}$$

where $A = (\sqrt{2\pi}\sigma)^{-1}$ represents the coefficient terms, d represents the distance between points and the center μ within the sample grids, σ determines the distribution of the convolutional kernel, μ denotes the extreme point, where the function takes on the highest value. For simplicity, we put μ at the center of the kernel. We set σ as a trainable parameter to accommodate multi-scale features generated by blocks at varying depths.

Furthermore, to prevent extreme values and a lack of receptive field, we apply Max normalization to the mask:

$$G(d) = \frac{f_{1d}(d)}{\underset{d \in D}{max} f_{1d}(d)} \tag{5}$$

where D represents a set of distances from the center of the kernel. With Max normalization, the modified Gaussian function always reaches its maximum value of 1 at the center point.

To better extract detail-enhanced features, we incorporate a 3*3 convolutional layer. Simultaneously, to maintain the lightweight nature of the model as much as possible, the input feature h is first downsampled before performing the 'Unsharp Masking' operation and then upsampled again after completion to restore its original shape. Taking the ConvNext [42] network as an example, our proposed Detail-Enhanced Adapter (DE-Adapter) is depicted in Fig. 1. The computation process can be formulated as follows:

$$\Delta h_l = f(CONV(h_l^d - G_\sigma(h_l^d)) \cdot W_{up}) \tag{6}$$

where $CONV$ denotes the 3*3 convolutional layer, $h_l^d = f(h_l \cdot W_{down})$ represents the original feature after downsampling, $G_\sigma(\cdot)$ denotes the Gaussian Convolution with a trainable parameter σ.

3.3 The DE-Adapter Integration Scheme

When integrating the proposed DE-Adapter module into the backbone, the integration scheme also plays a pivotal role, influencing knowledge transfer performance. To explore an effective integration scheme, both the position adaptation in the LVM and the insertion form of the adapter warrant careful consideration [15]. The position determines which layer the hidden representation h is to be adapted in the LVM, while the insertion form decides how to set the input to the DE-Adapter to compute the task-specific representation Δh.

Combining design dimension from these two perspectives, we meticulously design and assess five different integration schemes. Using the inverted residual blocks of ConvNext [29] as an example, Fig. 2 illustrates the proposed five variants of integration designs. The DE-Adapter can be flexibly inserted into every block in ConvNet, introducing only a minimal number of parameters. From our empirical studies, we found that: (1) Compared to the sequential architecture, the parallel architecture is more adept at extracting task-specific features. (2) When adapting convolutional networks on tasks with substantial domain shifts, the radical mismatch of the receptive field in Δh and h might result in inferior transfer performance. Therefore, we select the inverse residual parallel approach

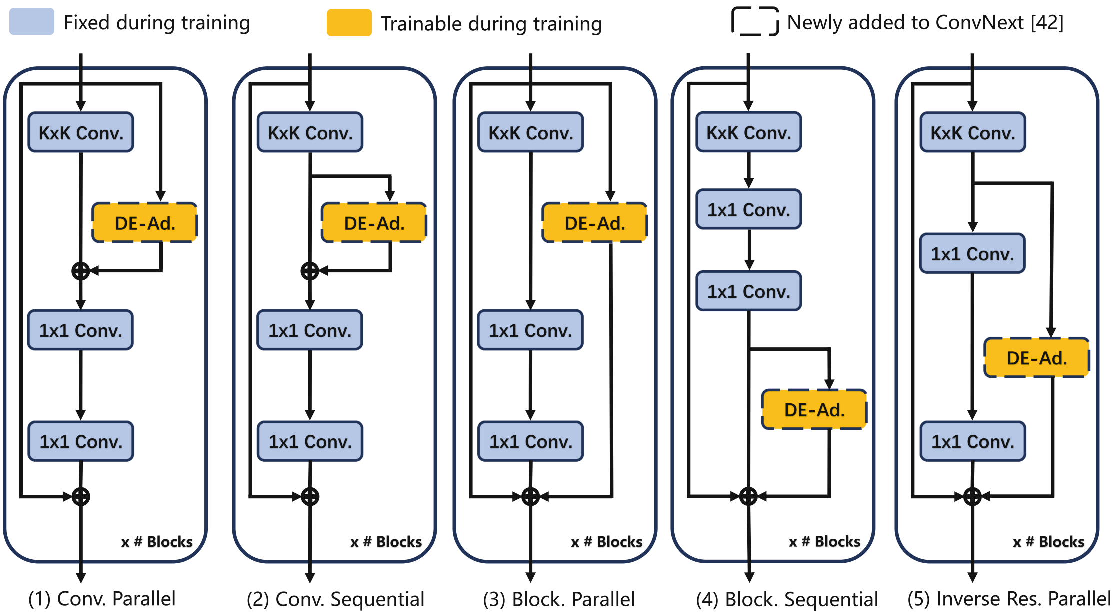

Fig. 2. Illustrations of five integrating designs of DE-Adapter to ConvNext. The schemes differ regarding the position of the modified representation and corresponding insertion form. DE-Ad. denotes the proposed DE-Adapter.

that simultaneously extracts task-specific features and maintains the same receptive field. Our ablation study in Table 5 further verifies this analysis.

Besides, it has been experimentally proven that other LVMs such as Vision Transformer [10] and even Swin Transformer [28] are also applicable.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. In this research, we utilize three extensively recognized deepfake datasets, including FaceForensics++ (FF++) [36], Celeb-DeepFake (Celeb-DF) [26] and DFDC [8]. The FF++ dataset is the most frequently employed dataset in this field, encompassing 1,000 original videos and 4,000 corresponding fake videos. The content within FF++ is compressed into two distinct versions: high quality (C23) and low quality (C40). The Celeb-DF dataset incorporates 590 real videos and employs the advanced DeepFake algorithm [26] to generate a substantial collection of 5,639 high-quality forgery videos. The DFDC dataset presents a unique challenge, housing an extensive array of 128,154 facial videos. These videos, originating from 960 diverse subjects, have been subjected to a variety of manipulations and perturbations, adding to the complexity of the dataset. For a rigorous comparative analysis, we report the Accuracy (ACC) and the Area Under the Receiver Operating Characteristic Curve (AUC), both of which are critical metrics in this field.

Implementation Details. Both our method and the re-implemented approaches are built on PyTorch [34]. All experiments were conducted using

Table 1. Comparison with State-of-the-art Forgery Detector Models. **Bold** and underline refer to the top and second result separately. # Params refers to the amount of tuning parameters. The symbol $*$ indicates the result of reproduction.

Method	#Params.	Input	FF++(C23)		FF++(C40)		Celeb-DF		DFDC	
			ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MesoNet [1]	2.8	256	83.10	–	70.47	–	–	–	–	–
Multi-task [32]	26.08	256	85.65	85.43	81.30	75.59	–	–	–	–
SPSL [27]	–	299	91.51	95.32	81.57	82.82	–	–	–	–
Face X-ray [25]	–	299	–	87.40	–	61.60	–	–	–	–
Xception [36]	20.81	299	95.73	96.30	86.86	89.31	97.90	99.73	78.87	89.39
RFM [39]	–	299	95.69	98.79	87.06	89.83	97.96	99.94	80.83	89.75
Add-Net [46]	–	299	96.78	97.74	87.59	91.01	96.93	99.55	78.71	89.85
F^3-Net [35]$*$	41.99	299	96.52	98.11	86.43	91.32	95.95	98.93	76.17	88.39
MultiAtt [44]	18.82	380	97.61	99.29	87.69	91.41	97.92	99.89	76.81	90.32
M^{2}TR [40]$*$	40.12	320	93.22	97.84	86.09	87.97	98.76	99.02	–	–
F^{2}Trans-B [31]$*$	128.01	224	96.59	99.24	87.21	89.91	98.79	99.23	81.32	89.12
DE-Adapter(Ours)	**0.38**	224	**98.01**	**99.31**	**88.92**	**93.10**	**98.93**	**99.91**	**81.59**	**90.37**

4 Nvidia GeForce 3090 GPUs. During training, we utilized random horizontal flipping as a form of data augmentation. We employed the AdamW optimizer [30] with an initial learning rate of 1e-3 and a weight decay of 1e-3. Additionally, a step learning rate scheduler was used to adjust the learning rate over time.

4.2 Main Results and Analysis

Comparison with State-of-the-Art Forgery Detector Models. Table 1 reveals that our method not only delivers state-of-the-art results across all datasets, but does so with a mere 0.38M training parameters - significantly less than other methods by at least one to two orders of magnitude. These findings suggest that in the era of large models, fully exploiting the robust capabilities of LVMs may be more beneficial than redesigning network structures.

Comparison with Previous PET Methods. Table 2 compares our method with the widely used and effective PET approach [5,22], as well as three baselines: Full Finetune (FT), Partially Tuning (PT), and Linear Probing (LP). The results underscore the efficacy of our proposed DE-Adapter, corroborating our analysis that for efficient fine-tuning of LVMs to solve face forgery detection task, merely adjusting representation mapping is insufficient. It is crucial to incorporate detailed information.

Evaluation on Limited Sample Training Dataset. Limited Sample Training Datasets, a common real-world scenario often overlooked by the community, was used to evaluate our model's data utilization efficiency. We restricted the detector to use only a tiny fraction of the training set (like 1/512, 1/256, etc.). Three state-of-the-art detectors (Multiatt [44], F^3-Net [35], and M^{2}TR [40]) were chosen as baselines.

Table 2. Comparison with PET methods, including Fine-Tuning (FT), Linear Probing (LP), Partially Tuning (PT), Bias Tuning (Bias), visual prompt tuning (VPT) [22] and Adaptformer [5]. **Bold** refers to the top result.

Method	#Params.	FF++(C23)		FF++(C40)		Celeb_DF		DFDC	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
FT	87.69	96.67	98.91	89.44	94.32	98.34	99.03	82.95	91.48
LP	0.01	68.21	54.19	58.76	50.79	84.64	55.44	61.07	51.39
PT	8.46	77.35	58.96	65.08	58.30	85.46	65.40	68.90	66.28
VPT [22]	0.02	71.91	55.31	66.51	52.51	83.89	66.26	65.73	56.43
Bias [3]	0.13	84.65	84.08	80.89	73.58	94.18	86.07	74.69	80.54
AdaptFormer [5]	0.09	86.72	83.19	81.11	76.25	95.18	89.27	71.15	79.69
DE-Adapter	0.38	+1.34	+0.40	−0.52	−1.22	+0.59	+0.88	−1.36	−1.11
(Ours)		**98.01**	**99.31**	**88.92**	**93.10**	**98.93**	**99.91**	**81.59**	**90.37**

Table 3. Generalization across datasets in terms of AUC (%) by training on FF++. **Bold** and underline refer to the top and second result separately.

Method	Celeb-DF	DFDC
RFM [39]	57.75	65.63
Add-Net [46]	62.35	65.29
F^3-Net [35]	61.51	64.59
MultiAtt [44]	67.02	67.79
DE-Adapter (Ours)	**67.12**	**68.84**

As shown in Fig. 3, our method significantly outperforms these small expert networks built on human prior knowledge when training data is limited, with the gap reaching over 40% at most. These results validate our analysis - fully leveraging LVMs' pre-training knowledge while avoiding network structure rebuilds drastically reduces the need for large training data volumes. In the era of large models, it becomes possible to achieve high performance with minimal training data.

Generalization Across Datasets. We conduct experiments on evaluating the generalization performance to unknown forgeries. Specifically, we train the models on the FF++ dataset and evaluate their performance on Celeb-DF and DFDC. The results, shown in Table 3, demonstrate that despite introducing only a minimal number of learnable parameters, our proposed DE-Adapter outperforms the baselines in terms of generalization. This result further validates the effectiveness of DE-Adapter and underscores its potential for robust generalization.

Universality of DE-Adapter. We evaluate the universality of DE-Adapter by applying it to various backbones pre-trained with various strategies. Specifically, we use Vision Transformer [10], Swin Transformer [28], ConvNext Model [29] as

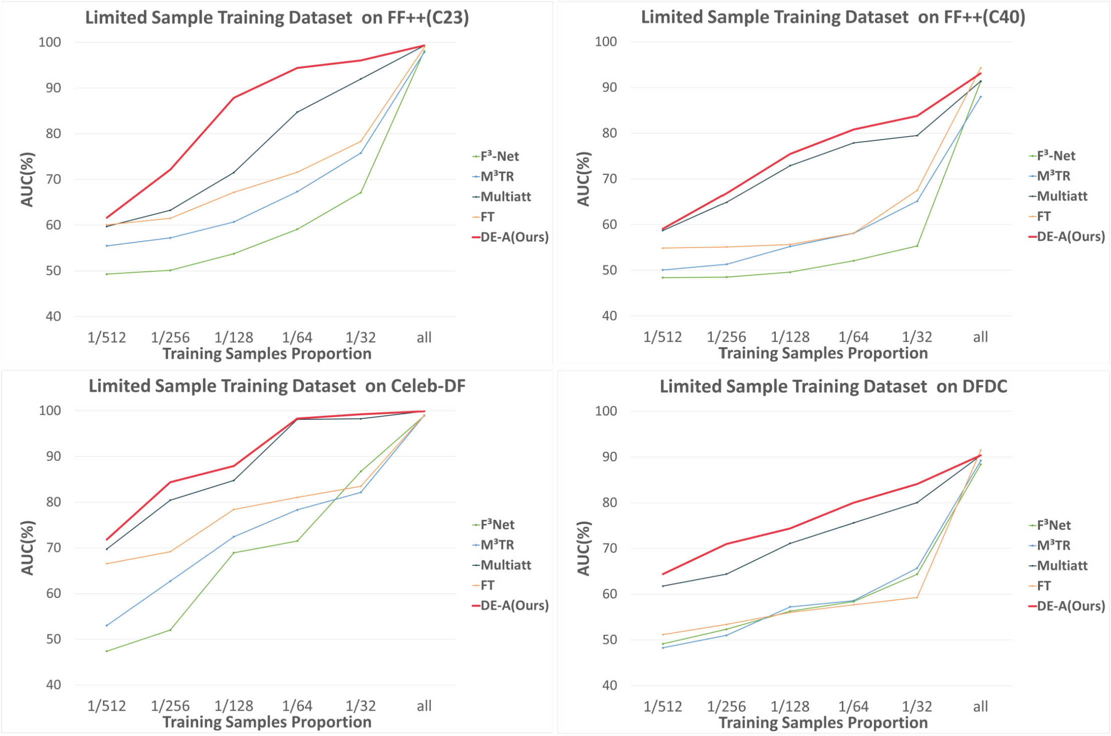

Fig. 3. Results of Limited Sample Training setting on various datasets. Our method significantly outperforms these small expert networks built on human prior knowledge when training data is limited, with the gap reaching over 40% at most. In the era of large models, it becomes possible to achieve high performance with minimal training data.

the backbone models. These models are trained on ImageNet-21K dataset [6] with supervised or self-supervised training. MAE [16] is adapted for self-supervised training. Table 4 presents the experiment results of different backbones and pre-training methods. We utilized the results of full fine-tuning as a benchmark for comparison. Our DE-Adapter performs comparably to full fine-tuning across all datasets, which not only attests to its ability to harness the potential of LVMs with a minimal number of parameters but also showcases the universality of our method regarding different structures and pre-training methods of LVMs.

4.3 Ablation Study

We provide an ablation study on each component of the DE-Adapter, including the adapter architectures and integration schemes. We use ConvNext V2 [42] as the backbone and conduct experiments on the FF++(C40) dataset.

Effects of Detail-Enhanced Operation. We try various Detail-Enhanced Operations in the adapter module. The results are shown in Table 5(a). Compared to the vanilla convolution baseline, a gain of +16.19% and +7.53% is obtained by using our Gaussian Mask. We have also tried other methods that extract detail information, such as LBP [23] and SRM [11]. However, the results are not satisfactory. We argue that these operations destroy the original semantic features.

Effects of Adapting Schemes. Table 5(b) showcases comparisons among various integration schemes. In general, the parallel approach outperforms the sequential one. We argue that the sequential method can not facilitate the extraction of task-specific features by the adapting module. Among the parallel approaches, both the convolution and block parallel methods can lead to a mismatch in the receptive fields due to differing convolution kernel sizes on either side of the parallel. Conversely, the inverse residual parallel emerges as the optimal choice due to its unique design that aligns the receptive fields.

Table 4. Comparison of DE-Adapter (DE-A) and full FineTuning (FT) with various backbones and different pre-training. CN denotes ConvNext backbones. Sup. denotes supervised pre-training.

Pre-train	Backbone	Method	#Param.	FF++(C23)		FF++(C40)		Celeb-DF		DFDC	
				ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Sup.	ViT-B	FT	86.13	**93.76**	**98.16**	85.24	90.36	97.86	99.47	**81.65**	**90.12**
		DE-A(Ours)	0.33	93.75	97.62	**85.78**	**89.53**	**98.17**	**99.67**	80.92	89.11
	Swin-B	FT	86.75	**97.10**	**99.48**	**88.97**	**94.59**	**98.95**	**99.95**	**82.76**	**89.77**
		DE-A(Ours)	0.43	96.61	98.97	87.51	92.30	98.05	98.63	81.17	89.59
	CN-T	FT	27.94	95.62	98.49	85.93	88.68	**97.81**	**98.71**	**81.16**	**89.54**
		DE-A(Ours)	0.12	**96.34**	**99.07**	**86.59**	**92.51**	97.79	98.53	80.10	89.10
	CN-B	FT	87.57	**97.22**	**99.43**	**88.89**	**94.41**	98.11	99.68	**81.63**	89.74
		DE-A(Ours)	0.32	96.70	99.18	88.09	93.89	**98.74**	**99.83**	81.59	**90.44**
MAE	ViT-B	FT	85.80	96.29	98.51	**87.85**	**93.12**	**98.37**	99.39	**81.06**	**89.01**
		DE-A(Ours)	0.32	**96.12**	**98.76**	87.22	92.59	98.03	**99.58**	80.56	88.15
	CNV2-T	FT	27.99	94.43	97.12	**87.23**	88.12	**98.37**	**99.34**	79.35	**89.50**
		DE-A(Ours)	0.12	**95.81**	**97.46**	86.03	**91.77**	97.91	99.12	**81.51**	89.12
	CNV2-B	FT	87.69	96.67	98.91	**89.44**	**94.32**	98.34	99.03	**82.95**	**91.48**
		DE-A(Ours)	0.32	**98.01**	**99.31**	88.92	93.10	**98.93**	**99.91**	81.59	90.37

Table 5. Ablation study on each component of DE-Adapter.

(a) Effects of convolution types

Detail-Enhanced Operation Type	AUC
Linear	76.19
Vanilla 3 × 3 Conv.	85.57
LBP Conv. [23]	84.15
SRM Conv. [11]	87.17
Gaussian Mask (Ours)	**93.10**

(b) Effects of adapting schemes

Adapting Scheme	AUC
Conv. Parallel	91.53
Conv. Sequential	89.72
Block Parallel	91.54
Block Sequential	90.90
Inverse Residual Parallel	**93.10**

Visualization. To gain deeper insights into the decision-making mechanism of our approach, we utilize Grad-CAM [37] for visualization on FF++ as displayed in Fig. 4. It is noticeable that the baseline method (an Adapter with standard 3×3 convolution) is highly vulnerable to high-frequency noise beyond the facial area in DeepFakes (DF), NeuralTextures (NT), and Real scenarios. In contrast, our method generates distinguishable heatmaps for authentic and forged faces where the highlighted regions fluctuate according to the forgery techniques, even though it solely relies on binary labels for training. For example, the heatmaps for both DeepFakes (DF) and FaceSwap (FS) concentrate on the central facial area, whereas that for Face2Face (F2F) identifies the boundary of the facial region. These results corroborate the effectiveness of the proposed DE-Adapter from a decision-making standpoint.

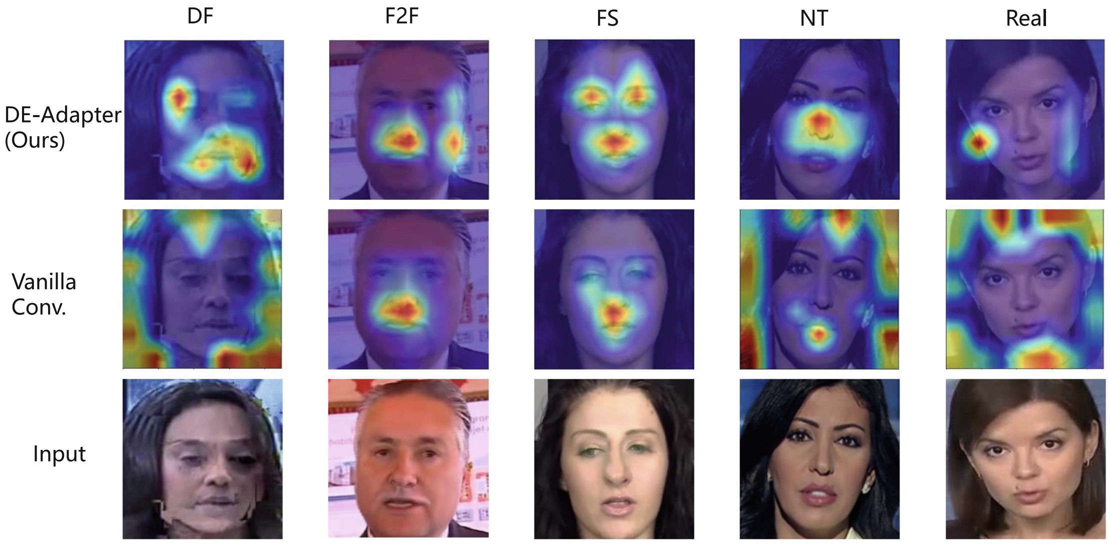

Fig. 4. The Grad-CAM visualization on the FF++ dataset. The first row and the second row display the proposed method result and the baseline result, respectively. It is noticeable that the baseline method is highly vulnerable to high-frequency noise beyond the facial area in DF, NT, and Real scenarios. In contrast, our method generates distinguishable heatmaps for authentic and forged faces where the highlighted regions fluctuate according to the forgery techniques, even though it solely relies on binary labels for training. For example, the heatmaps for both DF and FS concentrate on the central facial area, whereas that for F2F identifies the boundary of the facial region. These results corroborate the effectiveness of the proposed DE-Adapter from a decision-making standpoint.

5 Conclusion

In this paper, we propose a new paradigm to solve the forgery detection problem. Instead of redesigning small expert systems based on prior experience as before, we leverage the powerful capabilities of LVMs to address it. In order for LVMs to better adapt to such a specific downstream task as face forgery detection, inspired by the traditional image processing technique ‘Unsharp Masking’, we propose a lightweight Detail-Enhancement Adapter (DE-Adapter). By employing Gaussian convolution kernels and differential operations, the DE-Adapter captures detail-enhanced representations which are then integrated into the original representation, thereby ‘sharpen’ detail information, which enables LVMs to excel at detecting face forgery. With our method, we achieved state-of-the-art performance with only 0.3% of the training parameter volume. Especially when the number of training samples is limited, our method far surpasses other methods.

Acknowledgement. This work was supported in part by NSFC (62376156, 62322113), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: a compact facial video forgery detection network. In: WIFS (2018)
2. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P.N., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. ACM Trans. Graph **27**, 1–8 (2008)
3. Cai, H., Gan, C., Zhu, L., Han, S.: Tinytl: reduce memory, not parameters for efficient on-device learning. In: NeurIPS (2020)
4. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: AAAI (2021)
5. Chen, S., et al.: Adaptformer: adapting vision transformers for scalable visual recognition. In: NeurIPS (2022)
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: NeurIPS (2021)
8. Dolhansky, B., et al.: The deepfake detection challenge (DFDC) dataset. CoRR (2020)
9. Dong, B., Zhou, P., Yan, S., Zuo, W.: LPT: long-tailed prompt tuning for image classification. CoRR (2022)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)
11. Fridrich, J.J., Kodovský, J.: Rich models for steganalysis of digital images. TIFS **7**, 868–882 (2012)
12. Gao, Y., et al.: High-fidelity and arbitrary face editing. In: CVPR (2021)
13. Goodfellow, I.J., et al.: Generative adversarial networks. CoRR (2014)
14. Han, X., Morariu, V., Larry Davis, P.I., et al.: Two-stream neural networks for tampered face detection. In: CVPR Workshop (2017)
15. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. In: ICLR (2022)
16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
17. He, X., Li, C., Zhang, P., Yang, J., Wang, X.E.: Parameter-efficient fine-tuning for vision transformers. CoRR (2022)

18. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. In: ICLR (2022)
19. Huang, H., et al.: Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13673, pp. 37–54. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19778-9_3
20. Huang, Z., Chan, K.C.K., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: CVPR (2023)
21. Jia, G., et al.: Inconsistency-aware wavelet dual-branch network for face forgery detection. Trans. Biom. Behav. Ident. Sci. **3**, 308–319 (2021)
22. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
23. Juefei-Xu, F., Boddeti, V.N., Savvides, M.: Local binary convolutional neural networks. In: CVPR (2017)
24. Kim, K., et al.: Diffface: diffusion-based face swapping with facial guidance. CoRR (2022)
25. Li, L., et al.: Face x-ray for more general face forgery detection. In: CVPR (2020)
26. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: CVPR (2019)
27. Liu, H., et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: CVPR (2021)
28. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV (2021)
29. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR (2022)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
31. Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., Yu, N.: F^2trans: High-frequency fine-grained transformer for face forgery detection. TIFS (2023)
32. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: BTAS (2019)
33. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: parameter-efficient image-to-video transfer learning. In: NeurIPS (2022)
34. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
35. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: face forgery detection by mining frequency-aware clues. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV 2022. LNCS, vol. 12357, pp. 86–103. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_6
36. Rössler, A., Cet al.: Faceforensics++: learning to detect manipulated facial images. In: ICCV (2019)
37. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IJCV (2020)
38. Shi, Z., Chen, Y., Gavves, E., Mettes, P., Snoek, C.G.M.: Unsharp mask guided filtering. TIP **30**, 7472–7485 (2021)
39. Wang, C., Deng, W.: Representative forgery mining for fake face detection. In: CVPR (2021)
40. Wang, J., et al.: M2tr: multi-modal multi-scale transformers for deepfake detection. In: ICMR (2022)

41. Wang, Y., Xie, H., Xing, M., Wang, J., Zhu, S., Zhang, Y.: Detecting tampered scene text in the wild. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13688, pp. 215–232. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19815-1_13
42. Woo, S., et al.: Convnext v2: co-designing and scaling convnets with masked autoencoders. CoRR (2023)
43. Yao, G., et al.: One-shot face reenactment using appearance adaptive normalization. In: AAAI (2021)
44. Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., Yu, N.: Multi-attentional deepfake detection. In: CVPR (2021)
45. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: CVPR Workshop (2017)
46. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.: Wilddeepfake: a challenging real-world dataset for deepfake detection. In: ACM MM (2020)