LGSFNet: A Local–Global Semantic Fusion Network for robust cross-domain deepfake detection

Zulkaif Sajjad[1]
zulkaifsajjad123@outlook.com
Junaid Mir[1]
junaid.mir@uettaxila.edu.pk
Shah Nawaz[2]
shah.nawaz@jku.at
Furqan Shaukat[1]
furqan.shoukat@uettaxila.edu.pk
Muhammad Haroon Yousaf[3]
haroon.yousaf@uettaxila.edu.pk

[1] Department of Electrical Engineering
University of Engineering and
Technology Taxila, Pakistan

[2] Institute of Computational Perception
Johannes Kepler University Linz
Linz, Austria

[3] Department of Computer Engineering
University of Engineering and
Technology Taxila, Pakistan

Abstract

Deepfakes, enabled by recent advances in generative models, pose significant ethical, societal, and security risks. Although many detection methods achieve strong intra-dataset performance, they often degrade on low-quality or cross-domain data due to compression artifacts and unseen manipulations. To address this, we introduce LGSFNet, a robust deepfake detection framework that fuses local and global forgery semantics in a dual-path architecture. The design integrates a Spatial Resolution Adapter (SRA) to extract local low-level features and a novel Local Semantic Fusion Adapter (LSFA) to inject these cues into the DINOv3 transformer backbone for multi-stage feature fusion with parameter-efficient training. Experiments on FaceForensics++ demonstrate state-of-the-art results across all four manipulation types, achieving up to 99.98% AUC. Cross-corpora evaluations on Celeb-DF, DFD, and DFDC further highlight strong generalization, with improvements of up to +11.2% AUC over prior methods. A t-SNE visualization confirms discriminative representation of forgery features, while ablation studies validate that three LSFA modules achieve the best trade-off between performance and complexity. Overall, LGSFNet provides a robust, efficient, and generalizable solution for detecting low-quality and unseen deepfakes, moving toward reliable real-world deployment. The source code can be accessed using the link: https://github.com/zulkaifsajjad/LGSFNet

1 Introduction

Recent advances in deep generative models have enabled the synthesis of hyper-realistic facial images and videos that are often indistinguishable to the human eye. When misused,

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

these techniques facilitate misinformation and fabricated content across entertainment, social, and political domains [19, 20, 31, 32]. Known as deepfakes, such media raise serious ethical and security concerns. Common approaches include face swapping, face synthesis, attribute manipulation, expression transfer, and eye or lip-sync synthesis. Powered largely by generative adversarial networks (GANs) and autoencoders [7, 8, 25, 37], these methods allow users with minimal expertise to generate highly realistic manipulations. As a result, distinguishing authentic from synthetic content has become increasingly difficult, with innovation in generation techniques continually outpacing detection methods and sustaining the ongoing contest between forgers and forensic defenders.

To counter the growing threat of deepfakes, researchers have proposed detection methods that can be broadly categorized into spatial-based and frequency-based approaches. Spatial-based methods operate in the image domain, aiming to capture low-level forgery semantics. For instance, various studies have analyzed local textures [2, 10, 45, 47] to highlight appearance discrepancies between authentic and manipulated faces. Visual artifacts such as blending boundaries, which frequently arise from face forgery operations, have also been leveraged for detection [14]. In addition, a few studies [42, 46] investigate patch diffusion and path inconsistencies to model the correlation patterns between local features in real and synthetic content. Conversely, frequency-based approaches exploit spectral representations to identify forgery artifacts. For example, some methods [5, 6] extract the high-frequency components of the Discrete Fourier Transform (DFT) to characterize differences in spectral distribution between genuine and forged images. Similarly, F3-Net [27] employs local frequency statistics to capture forgery cues and introduces specialized designs to detect low-quality manipulations.

Existing detection methods perform well under intra-dataset evaluation, where training and test sets share similar distributions, but their effectiveness declines in cross-domain settings. This challenge is compounded by diverse image and video compression techniques [21, 24, 28] widely used on social media, which obscure the subtle artifacts left by manipulation. In highly compressed videos, cues such as texture, lighting, and boundary inconsistencies are often blurred, making detection difficult. Therefore, developing robust and generalizable deepfake detection methods is critical to counter the growing sophistication of generative forgeries. To address these challenges, and inspired by recent work [13, 26, 33], combining Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) has proven effective for distinguishing authentic from manipulated media. CNNs capture local spatial cues [23, 38] such as texture irregularities, edge inconsistencies, and pixel-level artifacts introduced during forgery, while ViTs model long-range dependencies and global semantics [38], enabling analysis of facial structure, contextual coherence, and temporal consistency. Their fusion thus yields a more robust framework, with CNNs providing fine-grained details and ViTs capturing broader contextual relationships.

We propose LGSFNet, a Local–Global Semantic Fusion Network that exploits complementary forgery semantics extracted from images. The network is organized into N stages. The first stage incorporates a pretrained DINOv3 backbone [34], which remains frozen during training, alongside our proposed Local Semantic Fusion Adapter (LSFA) with trainable parameters to facilitate efficient adaptation. In addition, the head module integrates the embedding layer of DINOv3 to generate learnable embeddings and a Spatial Resolution Adapter (SRA) to capture fine-grained local forgery semantics. By adopting a multi-stage fusion strategy, LGSFNet effectively combines local artifact details with global contextual representations, thereby enhancing the detection of low-quality and highly compressed manipulated videos.

The key contributions of this work are as follows:

1. A dual-path architecture that combines the Spatial Resolution Adapter (SRA) with DINOv3 embeddings to capture fine-grained local forgery semantics.
2. A novel Local Semantic Fusion Adapter (LSFA) that integrates global high-level semantics with local low-level features at multiple stages of DINOv3, enabling a more generalizable representation for detecting low-quality manipulated video frames.
3. An efficient integration strategy for LSFA that balances accuracy and model complexity, achieving parameter-efficient training with only 20% of the parameters being trainable.

2 Related Work

CNN and ViT-based Hybrid Approaches: Many studies have integrated specialized modules into CNN- and ViT-based architectures to enhance the modeling of long-range dependencies. For instance, a spatiotemporal inconsistency learning strategy [9] extracts generalizable forgery cues across both spatial and temporal domains. Similarly, a dynamic inconsistency learning method [10] detects subtle temporal artifacts in deepfake videos, capturing both global semantics and dynamic inconsistencies through a two-branch architecture. Although effective in cross-dataset evaluation, this approach struggles with low-quality videos and incurs a high computational cost due to its two-branch design and frame-level processing, which limits its suitability for real-time and resource-constrained environments. Other works have explored alternative strategies. A triplet network guided by depth maps [17] improves feature separation by combining depth estimation with triplet loss; however, its performance is limited in scenarios where depth signals are unreliable, such as compressed or occluded video frames. An efficient capsule network [12] has also been proposed to detect shallow and deep-fake facial forgeries by modeling part-whole relationships in facial images, incorporating the activation function of the max feature map to improve robustness and reduce complexity. More recently, a framework leveraging a pre-trained ViT with dual-level forgery modeling was introduced in the DeepFake Adapter [33]. This method integrates globally aware bottleneck adapters with locally aware spatial adapters to enable lightweight adaptation, achieving strong results on benchmark datasets such as FF++ and DFDC. Similarly, a multi-scale framework [41] employs high-frequency feature extraction and fusion modules to capture generalizable forgery semantics from the frequency domain. Despite their contributions, these methods remain less effective on low-quality video frames, where critical forgery cues are often suppressed or lost due to compression artifacts.

3 Methodology

3.1 Architecture Overview

The proposed architecture of LGSFNet is depicted in Fig. 1. Given an image $I \in R^{H\times W\times C}$ with spatial resolution $H \times W$ and number of channels $C = 3$, it is fed into the encoder, which consists of a head and intermediate stages. The input image I is processed in parallel by the DINOv3 embedding layer and the Spatial Resolution Adapter (SRA) module in the head part of the proposed encoder.

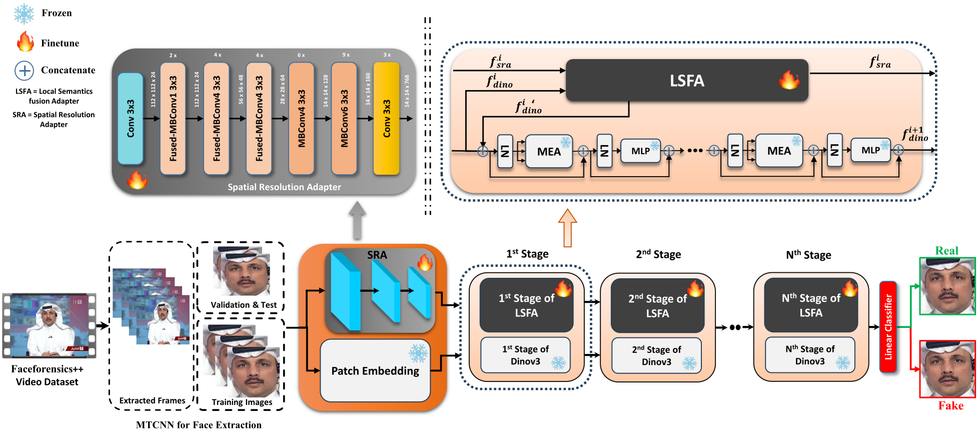

Figure 1: Block diagram of the proposed LGSFNet architecture.

A DINOv3 embedding layer divides the image into $P \times P$ non-overlapping patches and flattens them into sequential patches $I_p \in R^{K\times(P^2\cdot C)}$, where $K = H \cdot W/P^2$ is the total number of patches. These flattened patches are projected into D-dimensional embeddings and added with a positional embedding $f^1_{dino} \in R^{\;(P^2\cdot C)\times D}$ to retain the positional information.

Inspired by EfficientNetV2 [35], a lightweight SRA module depicted in Fig. 1 is composed of six sequential blocks: the first three are Fused-MBConv blocks, followed by two MBConv blocks, and concluding with a TripleConv block. The module combines Fused-MBConv and MBConv blocks to leverage fast feature extraction. Fused-MBConv merges expansion and depthwise convolution into a single 3×3 operation, with less computational cost, making them suitable for the initial SRA layers. On the other hand, MBConv blocks, first introduced in [30], utilize depthwise separable convolutions with an expansion–projection mechanism, enabling rich forgery feature representations. At last, the TripleConv block contains three consecutive 3×3 convolutional layers, each followed by Batch Normalization and ReLU activation, and it is designed to align the output channel dimension with the embedding dimension. By employing this strategy, the SRA module helps to extract high-quality, low-level feature maps. Each feature map is then projected into a standard embedding dimension D using a projection layer. The resulting vector from a projection layer is a unified feature representation $f^1_{sra} \in R^{B\times\left(\frac{H}{P}\cdot\frac{W}{P}\right)\times D}$. This efficient design of the SRA module enables the rich, local low-level features, effectively capturing fine-grained spatial details.

3.1.1 Transformers with N Stages

The extracted features f^1_{dino} and f^1_{sra} from the head part of the encoder are passed through the 1^{st} Stage of the encoder block. A pre-trained DINOv3-base backbone is utilized, comprising a total of L blocks, where each block consists of a Memory-Efficient Attention (MEA) and a Multi-Layer Perceptron (MLP) layer. N stages are formed by evenly grouping L blocks, with each stage containing L/N blocks of the DINOv3 and a single Local Semantic fusion

Adapter (LSFA) module for integration. We develop the LSFA interaction component for the SRA module, which facilitates the engagement of these features (e.g., f^1_{sra} in the 1^{st} Stage) with features from both the beginning of DINOv3 blocks at that Stage (e.g., f^1_{dino} and f^1_{sra} in the 1^{st} Stage).

The Multi-Head Cross Attention (MHCA) mechanism in the LSFA module, at the start of each DINOv3 stage, facilitates progressive adaptation. The MHCA layer incorporates local low-level CNN features into the DINOv3, allowing it to process input with insights from local patterns. This feedback loop fosters iterative alignment between high-level (DINOv3) and low-level (SRA) features, enhancing generalization to fine details. Thus, this fusion of local features from the SRA module at the start of each stage of DINOv3 is a purposeful design for synchronized local-global understanding.

Specifically, the interaction in the i^{th} Stage begins with a Multi-Head Cross Attention (MHCA) operation between f^i_{sra} and the features from the beginning of DINOv3 f^i_{dino}, as shown in Fig. 1. During this process, the normalized DINOv3 features $\widehat{f^i_{dino}}$ serves as the query while the normalized SRA features $\widehat{f^i_{sra}}$ are used as both the key and value as follows,

$$f^{i'}_{dino} = f^i_{dino} + \text{MHCA}\left(\widehat{f^i_{dino}}, \widehat{f^i_{sra}}, \widehat{f^i_{sra}}\right) \tag{1}$$

where $f^{i'}_{dino}$ are the features after the interaction of the LSFA module. These features are added element-wise with f^i_{dino} and then fed back into the DINOv3 blocks of the i^{th} Stage resulting in f^{i+1}_{dino} features. This interaction process injects the low-level features from the SRA module into the forward process of DINOv3 blocks. f^i_{sra} represents the same low-level features that will interact with the updated features from the DINOv3 blocks f^{i+1}_{dino} in the subsequent stage. Consequently, the encoded features will be further enhanced during the fusion process at the start of each stage.

After the extraction of f^{N+1}_{dino} features through N stages of the encoder block, the features are fed-forwarded to the Linear classifier (LC) and compute the cross-entropy loss.

$$\mathcal{L} = H\big(\text{LC}(f^{N+1}_{\text{dino}}), y\big) \tag{2}$$

where LC is the linear classifier, f^{N+1}_{dino} are the features extracted from the encoder, and y are labels form the corresponding samples, and $H(\cdot)$ is the cross-entropy function. We train SRA and all the LSFA modules with the cross-entropy loss function $\mathcal{L}$ in an end-to-end manner. In summary, by incorporating the high-level forgery features from the pre-trained DINOv3 that interact with local low-level features from SRA modules, our model based on the adaptation of low-level forgery semantics at multiple stages of DINOv3, could exploit better generalizable forgery representations.

4 Experiments

4.1 Benchmark Datasets

FaceForensics++ (FF++) [29] : FF++ dataset is used as a standard benchmark for deepfake detection. It contains 1,000 real YouTube videos and 4,000 fake videos generated by four distinct categories of manipulated techniques: Deepfakes, Face2Face, Neural Textures, and FaceSwap. We train our model separately on each of the four categories of fake videos, as well as on a combined faceforensics++ dataset.

Celeb-DF [16] : Celeb-DF is a well-known dataset for developing and evaluating deepfake detection models. It consists of 590 real videos and 5,639 manipulated videos collected from YouTube. The fake videos are generated by using advanced synthesis methods, which enhance the quality and reduce the visible artifacts typically found in deepfake videos.
Deepfake Detection Challenge (DFDC) [1] : DFDC contains 1,131 real videos and 4,119 manipulated videos that have been generated using several unknown manipulation methods. The faces in these videos may be partially real and partially forged, and most state-of-the-art algorithms struggle to detect whether the frame is real or fake.
DeepfakeDetection (DFD) [4] : DFD is collected by Google/Jigsaw, which has 363 real videos and 3,068 fake videos of 28 consented individuals of various genders, ages and ethnic groups. The details of the synthesis algorithm are not disclosed, but it is likely to be an improved implementation of the basic DeepFake maker algorithm.

4.2 Implementation Details:

The proposed LGSFNet framework is implemented using PyTorch 3.5.0. version. We only use the FaceForensics++ with 40% compression (low-quality) dataset for training and evaluation, and the other three datasets for cross-corpora evaluation to check the generalization of our method. Firstly, faces are extracted from the FaceForensics++ video dataset using a multi-task cascaded convolutional network (MTCNN) [43], ensuring only facial sections are analyzed. Images are resized to 384×384 pixels for consistency. Training includes multiple augmentations, such as flips and small-angle rotations, to enhance the model's robustness against real-world variations. The training occurred on a system equipped with an NVIDIA RTX 3090 GPU, which also included 24 GB of VRAM memory. An initial learning rate of 2e-5, a batch size of 32 within 30 epochs, and the Adam optimizer are used in training. The linear classifier includes only a single linear layer to classify the forgery or real faces. The training procedure utilized a learning rate scheduler that automatically adjusted the learning rate by evaluating validation accuracy. Standard-defined training, validation, and test sets of all datasets with 70%, 15%, and 15% splits, respectively, are used in this study.

4.3 Intra-domain subset evaluation of FF++:

To assess the effectiveness of the proposed LGSFNet framework, an intra-dataset evaluation is conducted on the FF++ dataset, covering its four manipulation types: DeepFakes (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT). The evaluation is performed on the low-quality (40% compression) version of FF++, which poses additional challenges due to compression artifacts. For a comprehensive comparison, we choose Xception [3], ResNet-50 [11], and EfficientNet-B4 [35] as baseline results. Table 1 reports the results, where both training and testing are carried out on the same manipulation type. LGSFNet consistently outperforms state-of-the-art (SOTA) methods across all four forgeries. In particular, the framework achieves 99.82% AUC on DF, 99.62% on FS, and 99.98% on F2F, improving upon the latest SOTA methods such as TripletNet [17] and Shao et al. [33] by 1–2% AUC. For the NT manipulation, which is the most challenging due to its complex texture synthesis, LGSFNet attains 98.94% AUC, surpassing competitive methods including CapsuleNet [12]. These results highlight the robustness and effectiveness of LGSFNet, particularly the benefit of integrating local and global forgery semantics through the LSFA and SRA modules.

For further validation, the proposed LGSFNet is trained on the combined FF++ dataset and evaluated separately on each manipulation type, as presented in Table 2. This setup

Method / Year	DF	FS	F2F	NT
Xception [3], 2017	96.78	94.64	91.07	87.14
S-MIL-T [15], 2020	97.14	96.07	91.07	86.79
ADD-Net [39], 2020	90.36	80.00	78.21	69.29
DSANet [44], 2020	97.86	95.36	93.57	92.50
STIL [9], 2021	98.21	97.14	92.14	91.78
SIM [10], 2022	99.28	97.86	95.71	94.28
CapsuleNet [12], 2023	98.61	99.51	99.68	95.14
TripletNet [17], 2023	99.13	97.64	96.53	85.10
Shao et al. [33], 2025	99.65	99.20	97.61	94.30
Ours	**99.82**	**99.62**	**99.98**	**98.94**

Table 1: Model trained and tested on sub-datasets of FF++ separated by four types of forgeries: F2F, FS, NT, and DF. Best results are highlighted in bold.

Method / Year	F2F	FS	NT	DF
ResNet-50 [11], 2016	93.76	93.30	83.43	93.34
Xception [3], 2017	96.92	95.85	94.00	97.47
EfficientNet-B4 [36], 2019	97.41	97.10	90.87	97.02
F3-Net [27], 2020	96.56	94.14	93.15	97.67
SRM [22], 2021	96.49	97.59	92.66	97.64
UCF [40], 2023	97.12	97.46	91.99	97.40
Wei *et al.* [41], 2024	99.15	99.36	96.23	**99.29**
Lin *et al.* [18], 2024	98.37	97.97	95.06	98.86
Ours	**99.81**	**99.92**	**98.10**	99.18

Table 2: Model trained on combined FF++ and evaluated on intra-domain on sub-datasets of FF++ separated by four types of forgeries: F2F, FS, NT, and DF. Best results are highlighted in bold.

provides more balanced learning across different forgeries. Even under this more challenging setting, LGSFNet continues to demonstrate superior performance, achieving 99.81% AUC on F2F, 99.92% on FS, and 98.10% on NT. The consistent improvements across both single-type and combined-type training strategies indicate that LGSFNet not only excels in intra-type evaluation but also exhibits strong generalization capability when exposed to diverse manipulation types.

4.4 Cross-dataset Evaluation

To further assess the generalization ability of LGSFNet on unseen forgeries with larger variations, cross-corpora experiments are conducted where the training is performed on the C40 version of FF++ and testing data originate from three widely used benchmarks: Celeb-DF, DFD, and DFDC. The results, summarized in Table 3, show that LGSFNet consistently outperforms SOTA approaches by a significant margin in terms of AUC. In particular, the framework achieves 86.38% on Celeb-DF, 84.03% on DFD, and 64.41% on DFDC, surpassing Wei et al. [41] by approximately +9.4%, +1.0%, and +1.9%, and outperforming Lin et al. [18] by +11.2%, +3.5%, and +2.2% on the respective datasets. Compared to earlier baselines

Method / Year	FF++	Celeb-DF	DFD	DFDC
ResNet-50 [11], 2016	91.06	64.78	72.94	53.38
Xception [3], 2017	95.93	69.37	78.08	56.87
EfficientNet-B4 [36], 2019	95.63	67.80	76.81	56.59
F3-Net [27], 2020	95.64	67.62	80.51	55.96
SRM [22], 2021	96.30	68.08	77.57	58.22
UCF [40], 2023	96.17	70.48	75.68	55.20
Wei *et al.* [41], 2024	98.58	76.94	83.02	62.55
Lin *et al.* [18], 2024	97.68	75.19	80.56	62.18
Ours	**99.36**	**86.38**	**84.03**	**64.41**

Table 3: Model trained on combined FF++ and Cross-corpora evaluation on Celeb-DF, DFD, and DFDC. Best results are highlighted in bold.

such as ResNet50 [11] and Xception [3], the improvements are even more pronounced, often exceeding 15–20% absolute gains on the challenging cross-domain benchmarks. These results clearly demonstrate that our LGSFNet framework, regularized by generalizable high-level forgery semantics from the pre-trained DINOv3 backbone and enhanced by the LSFA module for fusion of local low-level forgery cues, achieves stronger generalization across diverse datasets. This highlights the robustness of our approach for practical deepfake detection scenarios, where training and testing conditions often differ significantly.

4.5 Visualization of Features

To assess the representation capability of the proposed architecture, which integrates the dual-path design with the efficient fusion of the LSFA module into DINOv3, a t-SNE (t-distributed stochastic neighbor embedding) visualization is generated from the extracted features of each manipulation type (Fig. 2). In the 2D t-SNE map, each dot corresponds to an image, with green representing real samples and red representing fake samples. Well-separated clusters reflect effective feature discrimination, whereas overlapping regions suggest weaker generalization. For this analysis, 1,500 real and 1,500 fake samples are randomly selected from the test set. The results demonstrate that LGSFNet achieves strong generalization, maintaining a clear separation between real and fake clusters. Minor overlaps are observed in the NT manipulation, likely due to its subtle texture variations, which slightly reduce separability. Overall, the framework exhibits robust discriminative embedding learning, with most samples forming distinct and well-preserved spatial clusters.

4.6 Ablation Study

To further validate the design of LGSFNet, an ablation study is conducted by varying the number of LSFA modules and reporting performance across multiple datasets (Table 4). As outlined earlier, the LSFA is designed to complement the SRA, which extracts local low-level semantics from forgery images in parallel with the DINOv3 embedding layer. While the SRA enriches the input with localized features, the LSFA injects these low-level cues into the DINOv3 transformer by fusing them at the beginning of each stage.

The results indicate that using two LSFA modules achieves strong performance on FF++, but underperforms on cross-dataset evaluations, particularly on DFDC. Increasing the num-

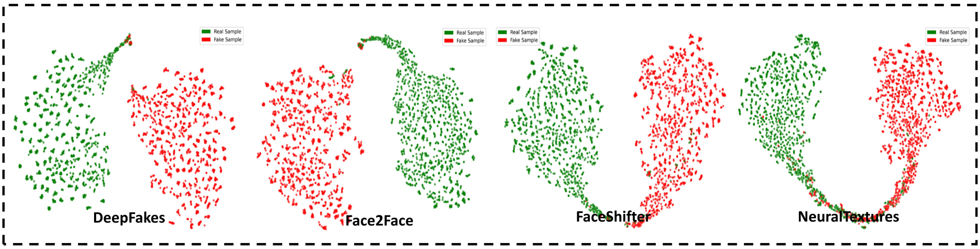

Figure 2: T-SNE features visualization on sub-datasets of FF++ separated by four types of forgeries: F2F, FS, NT, and DF.

LSFA Modules	FF++	Celeb-DF	DFD	DFDC
2	99.36	87.88	81.18	61.38
3	99.36	86.38	84.03	64.41
6	99.22	89.72	84.32	64.89

Table 4: Ablation study on the number of LSFA modules and evaluation on different datasets.

ber to three LSFA modules provides the most favorable trade-off between accuracy and complexity, yielding superior results on Celeb-DF (86.38%) and competitive gains on DFD (84.03%) and DFDC (64.41%). In contrast, increasing to six LSFA modules slightly improves generalization on Celeb-DF, but reduces performance on FF++ and DFDC, likely due to overfitting and higher model complexity.

Based on these findings, the final architecture employs three LSFA modules, as this configuration consistently demonstrates robust cross-dataset generalization while maintaining manageable model complexity.

5 Conclusion

In this paper, we introduced LGSFNet, a robust and scalable deepfake detection framework built on the DINOv3 architecture. To enhance cross-scale feature interaction, we integrated the Spatial Resolution Adaptor (SRA) and LSFA modules, enabling efficient fusion of low-level forgery semantics with high-level features across network stages. Extensive evaluations on multiple benchmarks demonstrate that LGSFNet achieves state-of-the-art performance on FF++, Celeb-DF, DFD, and DFDC, while requiring only 20% trainable parameters. These results highlight the effectiveness and practicality of our approach for real-world deployment. Future work will explore extending LGSFNet to detect multi-modal forgeries, such as audio-visual deepfakes and cross-domain manipulations.

6 Acknowledgment

This work is supported by the UK Engineering and Physical Sciences Research Council [Grant number: EP/Y002288/1].

References

[1] Ben Pflaum Jikuo Lu Russ Howes Menglin Wang Cristian Canton Ferrer Brian Dolhansky, Joanna Bitton. The deepfake detection challenge dataset, 2020.

[2] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1081–1088, May 2021. doi: 10.1609/aaai.v35i2.16193. URL https://ojs.aaai.org/index.php/AAAI/article/view/16193.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[4] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google & jigsaw. *arXiv preprint arXiv:1901.08971*, 2019.

[5] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

[6] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

[7] FaceApp. Faceapp, 2025. URL https://www.faceapp.com/. Mobile application software.

[8] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. High-fidelity and arbitrary face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16115–16124, 2021.

[9] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021.

[10] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 744–752, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Hafsa Ilyas, Ali Javed, Khalid Mahmood Malik, and Aun Irtaza. E-cap net: an efficient-capsule network for shallow and deepfakes forgery detection. *Multimedia Systems*, 29 (4):2165–2180, 2023.

[13] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Hybrid transformer network for deepfake detection. In *Proceedings of the 19th international conference on content-based multimedia indexing*, pages 8–14, 2022.

[14] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[15] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1864–1872, 2020.

[16] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[17] Buyun Liang, Zhongyuan Wang, Baojin Huang, Qin Zou, Qian Wang, and Jingjing Liang. Depth map guided triplet network for deepfake face detection. *Neural Networks*, 159:34–42, 2023.

[18] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16815–16825, 2024.

[19] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

[21] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11006–11015, 2019.

[22] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.

[23] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023.

[24] Lagsoun Abdel Motalib, Oujaoura Mustapha, and Hedabou Mustapha. Compression-aware hybrid framework for deep fake detection in low-quality video. *IEEE Access*, 2025.

[25] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[26] Georgios Petmezas, Vazgken Vanian, Konstantinos Konstantoudakis, Elena EI Almaloglou, and Dimitris Zarpalas. Video deepfake detection using a hybrid cnn-lstm-transformer model for identity verification. *Multimedia Tools and Applications*, pages 1–20, 2025.

[27] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[28] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G Anderson, and Lubomir Bourdev. Learned video compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3454–3463, 2019.

[29] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[31] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[32] Rui Shao, Tianxing Wu, and Ziwei Liu. Robust sequential deepfake detection. *International Journal of Computer Vision*, pages 1–18, 2025.

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

[34] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[35] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[36] Mingxing Tan, Q Efficientnet Le, et al. Rethinking model scaling for convolutional neural networks. In *Proceedings of the International conference on machine learning, Long Beach, CA, USA*, volume 15, 2019.

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

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[39] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[40] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22412–22423, 2023.

[41] Wei Ye, Xinan He, and Feng Ding. Decoupling forgery semantics for generalizable deepfake detection. *arXiv preprint arXiv:2406.09739*, 2024.

[42] Baogen Zhang, Sheng Li, Guorui Feng, Zhenxing Qian, and Xinpeng Zhang. Patch diffusion: a general module for face manipulation detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3243–3251, 2022.

[43] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, Oct 2016. ISSN 1070-9908. doi: 10.1109/LSP.2016.2603342.

[44] Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. V4d: 4d convolutional neural networks for video-level representation learning. *arXiv preprint arXiv:2002.07442*, 2020.

[45] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[46] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[47] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2929–2939, 2021.