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

MLVP-Net: Deepfake Detection Based on Multi-Level Visual Perception

Kai Li, Shaochen Jiang, Liejun Wang, Songlin Li, Chao Liu, and Sijia He

College of Computer Science and Technology, Xinjiang University
Xinjiang Multimodal Intelligent Processing and Information Security Engineering Technology Research Center

Abstract. The negative impacts of Deepfake technology have attracted widespread attention in the multimedia forensics community. Due to the insufficient diversity of existing datasets, models tend to overly rely on forgery-specific features, resulting in poor generalization. To address this issue, we propose a multi-level visual perception network (MLVP-Net), which explores local, spatial, and semantic consistency from different perspectives to improve detection accuracy. Specifically, we first introduce a Multi-scale Spatial Perception Module (MSPM) that effectively captures both long-range and local information through parallel cascaded Hybrid State Space (HSS) blocks and multi-kernel convolution operations. Then, we present a Detail Feature Enhancement Module (DFEM), which employs multiple differential convolutions for multi-directional perception, enabling the model to sense and weight details from different directions. Finally, we propose a Content-Adaptive Attention Module (CAAM), which enriches contextual information by fusing multi-level features while guiding the model to focus on more useful information through combing channel and spatial attention mechanisms. Extensive experiments demonstrate that our MLVP-Net significantly outperforms all comparison methods across five benchmark datasets in Deepfake detection.

Keywords: Deepfake Detection, Efficientnet, State Space Model, Differential Convolution Content, Adaptive Attention.

1 Introduction

With the continuous development of information technology, Deepfake technology has gained widespread attention in recent years. This technology enables the creation of highly realistic forged media, particularly images, videos, and audio, which has raised serious concerns in fields such as security, politics, entertainment, and social media. As Deepfake technology advances, the authenticity of forged content becomes increasingly difficult to distinguish, posing significant threats to public trust, privacy protection, and national security. As a result, the detection of Deepfake content has become a critical task in the field of multimedia forensics. In recent years, numerous methods have been proposed to address this challenge. For example, Guo et al. [1] proposed

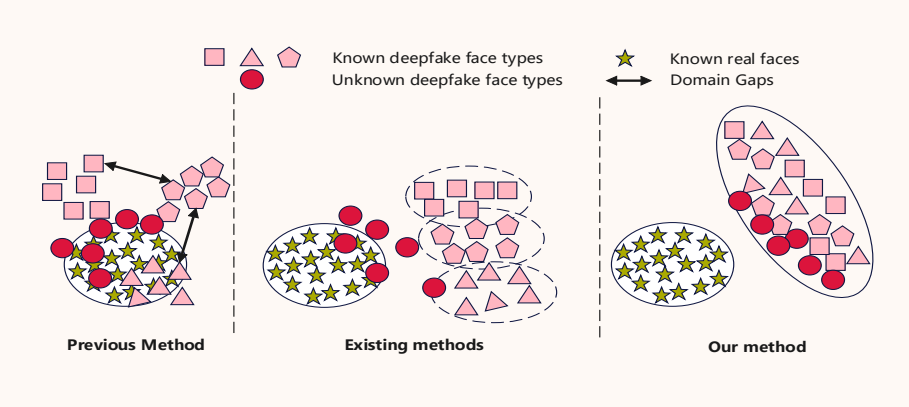

Fig. 1. (a) Previous methods exhibit limited capability in learning forgery-specific features, which makes it difficult to distinguish unknown forgery types from authentic facial samples. (b) Existing methods demonstrate relatively strong feature learning abilities and can effectively separate unknown forgeries from real faces; however, they tend to develop a rigid reliance on specific forgery patterns. (c) Our method learns more generic and robust forgery representations through multi-level visual perception, without being constrained to fixed forgery types, thus exhibiting superior generalization to unseen forgeries.

LDFnet to tackle the issue of achieving an optimal balance between detection accuracy and model complexity. Li et al. [2] introduced BLR-Net, which improves object detection accuracy by jointly guiding feature representation through localization and boundary information. Guo et al. [3] also proposed AdapGRnet, which enhances Deepfake detection accuracy by complementarily fusing spatial and residual domain features. Yan et al. [4] expanded the forgery space by modeling the internal and inter-feature variations of forgery characteristics in latent space, aiming to overcome the generalization barriers faced by Deepfake detection. Additionally, Guo et al. [5] proposed SFIConv, which simulates the manipulation traces left by Deepfakes, enhancing the traditional backbone network's ability to capture space-frequency interactions and thereby improving detection accuracy. Zhou et al. [6] proposed an unsupervised domain adaptation method for fine-grained open-set Deepfake detection, which enhances the model' s generalization ability to unknown Deepfake classes through adaptive clustering and pseudo-label generation.

Although existing methods have achieved remarkable success, several issues still need to be addressed. First, the lack of data diversity lead to overfitting, where the model tends to memorize specific forgery patterns rather than learning generalized features. Second, an overreliance on local forgery features or features specific to certain forgery techniques results in a limited ability to handle new forgery methods. Lastly, the similarity between real and fake features is high, and forgery techniques often involve fine manipulation of facial details. If the model fails to effectively capture these subtle differences, distinguishing authenticity becomes extremely challenging.

With the above thoughts in mind, we propose a novel multi-level visual perception network (MLVP-Net), as illustrated in Fig. 1. Our architecture enhances the discrimination capability against unknown forgery techniques by mapping features into different feature spaces to perform multi-level visual perception, thereby learning the common characteristics shared across various forgeries. Specifically, we introduce a Multi-scale Spatial Perception Module (MSPM) that effectively captures long-range dependencies and local details within images through parallel cascaded Hybrid State Space (HSS) blocks and multi-kernel convolution operations. Subsequently, we present a Detail Feature Enhancement Module (DFEM), which utilizes multi-directional differential convolutions to perceive and amplify subtle inconsistencies from multiple directions, thereby increasing sensitivity to forgery artifacts. Finally, we propose a Content-Adaptive Attention Module (CAAM), which dynamically adjusts and emphasizes critical information by integrating channel, spatial, and pixel attention mechanisms, thereby enriching the representation of contextual information. The proposed network not only significantly improves the accuracy of authenticity discrimination for facial images, but also effectively enhances robustness against unseen forgery attacks. The main contributions of this paper include:

– We present a novel MLVP-Net that significantly enhances Deepfake detection accuracy by integrating multi-scale spatial perception, detail feature enhancement, and content-adaptive attention mechanisms. Extensive experiments conducted on five widely recognized benchmark datasets show that MLVP-Net significantly outperforms existing state-of-the-art methods in detection accuracy.
– We propose a DFEM that utilizes multi-directional differential convolutions to perceive and amplify subtle inconsistencies from multiple directions. This approach increases sensitivity to forgery artifacts and enhances the model's ability to detect Deepfakes effectively.
– We propose a CAAM that integrates channel, spatial, and pixel attention mechanisms to perform dynamic adaptive adjustments. This enables effective fusion of multi-level features and guides the model to focus on more informative and relevant information.
– We propose a MSPM that effectively captures long-range dependencies and local details within images through parallel cascaded HSS blocks and multi-kernel convolution operations, thereby enhancing the robustness of feature extraction.

2 Methodology

2.1 Overview

The structure of our proposed MLVP-Net is illustrated in Fig. 2. We utilize EfficientNet [7] as the backbone network to extract multi-level features from input images. To fully leverage the advantages of features at different levels while minimizing the interference of redundant information, we selectively use the features from the third, fifth, and final stages, denoted as f1, f2, and f3, respectively. Subsequently, we input f1 and f2 into the

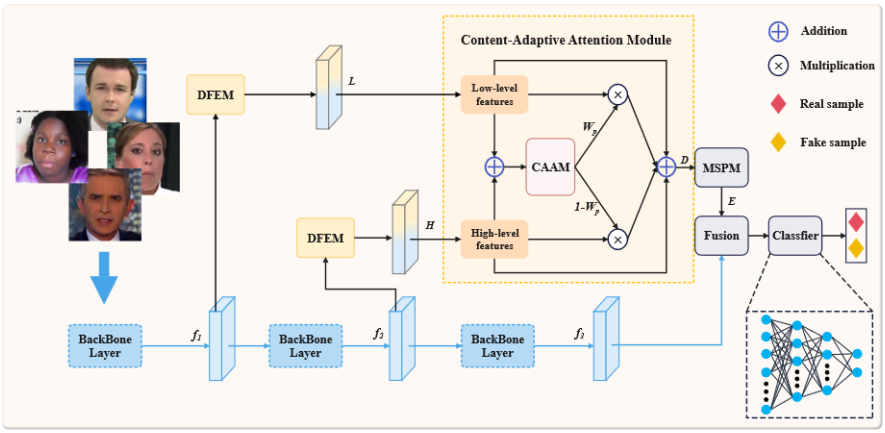

Fig. 2. The overall architecture of MLVP-Net.

Detail Feature Enhancement Module (DFEM) to enhance their high-frequency features, resulting in enhanced features L and H. Then, L and H are input as low-level and high-level features into the Content-Adaptive Attention Module (CAAM), respectively, which produces pixel-level weights Wp. These weights Wp are used to adaptively adjust the spatial and channel information, resulting in feature D. Finally, MSPM is applied to project feature D into a one-dimensional space, capturing sufficient contextual information and enhancing sensitivity to forged details for accurate prediction.

2.2 Multi-scale Spatial Perception Module

Forgery techniques typically focus on refining specific local facial regions to make the manipulated content appear more convincing. A lack of sensitivity to subtle forgery details increases the risk of misclassification. To address these challenges, we propose a MSPM to capture rich and sufficient contextual information while enhancing sensitivity to subtle forgery details. As illustrated in Fig. 3. Specifically, the input feature D is fed into both the global and local branches of the MSPM module to extract rich global contextual information and refined local representations, respectively. We first introduce the global branch. To accommodate the complexity of the Deepfake task, we adopt a cascaded hybrid state-space module strategy to ensure the acquisition of sufficient contextual information:

$$G = B_n^{\mathrm{HSS}}\left(B_{n-1}^{\mathrm{HSS}}\left(\cdots B_2^{\mathrm{HSS}}\left(B_1^{\mathrm{HSS}}(X_i)\right)\right)\right) \qquad (1)$$

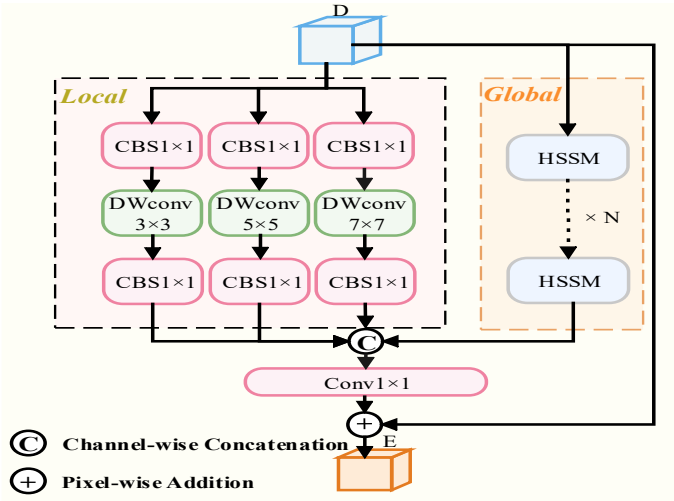

Fig. 3. Multi-scale Spatial Perception Module (MSPM)

Fig. 4 illustrates the principle of HSSM, which is a Mamba-based method with hybrid encoding. It utilizes a Zigzag scanning strategy to address the spatial continuity issue in the scanning process [8], offering excellent long-range modeling capability and linear computational complexity. The details are as follows:

HSSM utilizes a Zigzag scanning strategy to perform an initial global scan of the input features. The input features are then divided into four equal patches, with local scanning performed on each patch. After merging the scanning results, they are concatenated with the global scan encoding to produce the original encoding. By applying various encoding operations to the original encoding, such as Inverse Encoding, Rotation Encoding, and Transposition Encoding, we enhance the module's multi-layer visual perception capability and robustness. These encoding results are then projected into a 1D space to model long-range dependencies. Finally, the spatial structure of the features is restored using a decoding matrix, yielding the output results.

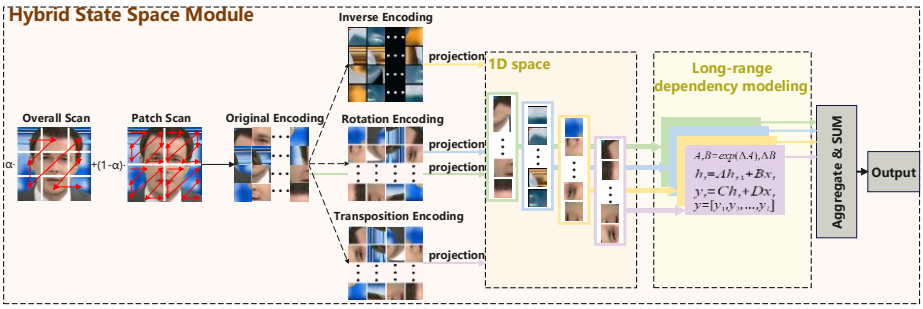

Fig. 4. Hybrid State Space Module (HSSM)

It should be further noted that why HSSM is effective in capturing long-range dependencies is explained by the following equations:

$$h_t = e^{A\Delta t} h_{t-1} + \left(I - e^{A\Delta t} \right) \mathbf{B} x_t \qquad (2)$$

where A denotes the state transition matrix, and B denotes the projection matrix. h_t denotes the state at the t-th time step, h_{t-1} denotes the state at the previous time step, and x_t denotes the fixed-length one-dimensional sequence encoding input at the t-th time step. I is the identity matrix, and A is a negative definite matrix. Δt denotes the time step, describing the interval between two time steps. Its value determines whether the current state h_t is more influenced by the current input x_t or the previous state h_{t-1}. As a result, the current state h_t is determined by both the current input and the historical states across different time steps. While traditional convolutions focus solely on local features within a single state, HSS demonstrates outstanding performance in capturing long-range dependencies, reducing reliance on specific patterns, and significantly enhancing the model's generalization capability.

We then introduce the local branch. The details are as follows: We perform efficient local feature extraction using multi-scale depthwise separable convolutions, and combine BatchNorm2d and Silu to mitigate the vanishing gradient problem and enhance the network's nonlinear modeling capability.

$$L = \mathrm{CBS}_{1\times1}\left(\mathrm{DWBS}_{k\times k}\left(\mathrm{CBS}_{1\times1}(D) \right) \right) \qquad (3)$$

The global features are fused with local features at different scales, and D is used as a residual connection to alleviate the vanishing gradient problem, resulting in the final output of MSPM:

$$X_o = \mathrm{C}_{1\times1}\left(\mathrm{Concat}\left(G, \mathbf{L}_{k_3}, \mathbf{L}_{k_5}, \mathbf{L}_{k_7} \right) \right) + D \qquad (4)$$

where $\mathrm{CBS}_{1\times1}$ and $\mathrm{DWBS}_{k\times k}$ denotes the standard convolution and depthwise separable convolution, respectively, they both include a BatchNorm layer and a SiLU activation function. Concat(∗) refers to concatenation along the channel dimension.

MSPM enriches contextual information by capturing long-range dependencies to enhance the global feature representation capability. Additionally, we capture fine-grained local information through multi-scale convolutions.

2.3 Detail Feature Enhancement Module

Advanced Deepfake forgeries often manifest as subtle artifacts embedded in high-frequency edge and texture details, making them difficult to detect. While standard convolution effectively captures pixel-level intensity, it tends to smooth local regions and overlook inter-pixel variations, leading to detail loss. To mitigate this, we propose a Detail Feature Enhancement Module (DFEM) that integrates multi-directional differential convolution for gradient-aware feature extraction with vanilla convolution (VC) to jointly enhance and preserve fine-grained details.

Previous studies [9-12] have demonstrated the effectiveness of differential convolutions in enhancing gradients and capturing intrinsic detailed patterns, particularly in edge detection and anti-spoofing tasks. As shown in Fig. 5, the DFEM consists of four

parallel differential convolutions: Central Difference Convolution (CDC), Angular Difference Convolution (ADC), Horizontal Difference Convolution (HDC), Vertical Difference Convolution (VDC), and VC. The Gradient-Enhanced Image clearly preserves richer detail information than that produced by vanilla convolution.

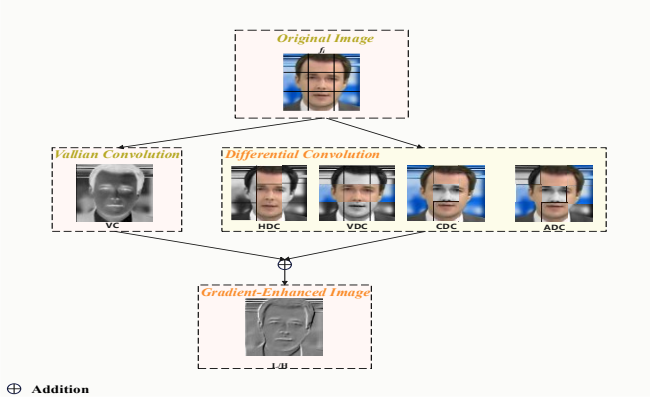

Fig. 5. Detail-Feature Enhance Module(DFEM)

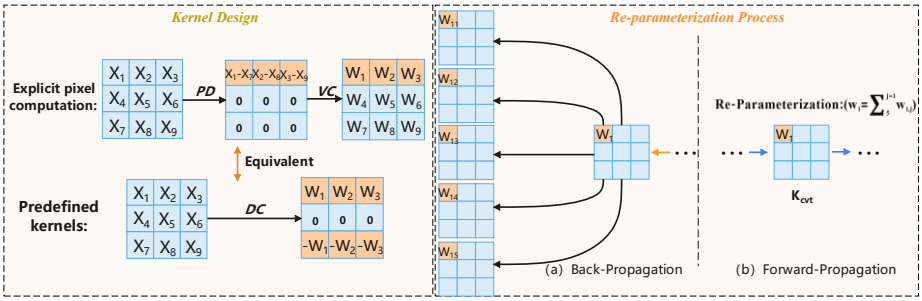

Fig. 6. Kernel Design and Re-parameterization Process

Fig. 6 illustrates the convolutional kernel design process using differential convolution, with vertical differential convolution as an example. PD denotes Pixel Difference, VC denotes Vanilla Convolution, and DC denotes Differential Convolution. x_i and w_i denotes the pixel values and convolutional kernel weights, respectively. Explicit pixel computation first captures inter-pixel differences via *PD*, followed by *VC* to model local variations. Predefined kernels, exemplified by vertical differential convolution, encode gradient priors into convolutional layers through fixed kernels, enhancing the network's ability to capture gradient information. Specifically, *DC* avoids explicit pixel difference computation while achieving the same effect. Similarly, we can design several other types of differential convolutions.

Fig. 6 also illustrates our reparameterization process. We apply the reparameterization technique to simply sum the weights k_i and biases b_i of the five types of convolutions to construct an equivalent convolution, thereby achieving the same effect as five

parallel convolutions with a single equivalent convolution. The specific process is as follows:

$$F_{\text{out}} = F_{\text{in}} * \left(\sum_{i=1}^{5} k_i \right) + \sum_{i=1}^{5} b_i = F_{\text{in}} * K_{\text{cvt}} + b_{\text{cvt}} \qquad (5)$$

Notably, during backpropagation, the computational graph automatically tracks the gradient flow, ensuring proper allocation for updating each convolution's weights and biases. In forward propagation, the structure is re-parameterized into a single convolution with multi-directional perception, significantly reducing computational overhead. This unified representation enables efficient inference and accelerates the overall training process. Consequently, DFEM enhances multi-level detail perception while maintaining high computational efficiency.

2.4 Content-Adaptive Attention Module

Deepfake detection requires rich and effective semantic information. Many studies combine low-level features and high-level features from backbones to obtain multi-level information. Since a single pixel in high-level features originates from a pixel region in low-level features, it indicates that they possess different receptive fields. Previous works [13], [14], [15] employed simple element-wise addition for feature fusion. Subsequently, Wu et al. [16]. proposed adjusting the fusion ratio through self-learned weights, which offers greater flexibility compared to addition. However, the aforementioned methods still fail to adequately address the issue of receptive field mismatch. Research on Feature Attention Modules (FAM) [17] employs independent channel attention and spatial attention mechanisms to guide the model in focusing on more informative features during the encoding process.

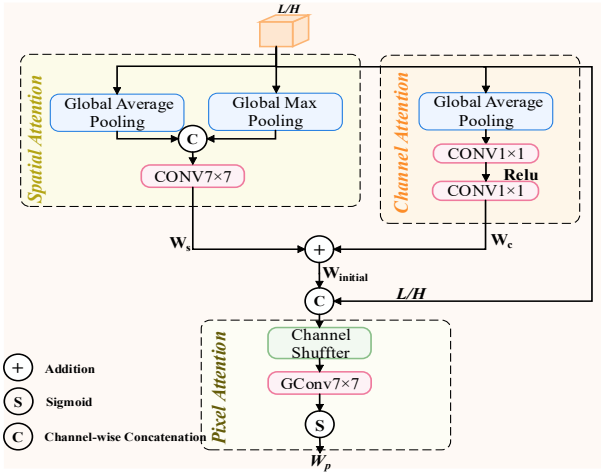

Fig. 7. Content-Adaptive Attention Module (CAAM)

Inspired by the aforementioned studies, we designed the CAAM, which combines channel attention and spatial attention to guide the model in focusing on effective semantic information. Besides, In Fig. 2, we developed an adaptive feature fusion mechanism based on CAAM, which effectively integrates features with different receptive fields while preserving rich semantic information. The design of CAAM is illustrated in Fig. 7, The details are as follows:

We first obtain the channel-level weights and spatial-level weights separately. It can be formulated as:

$$W_c = \mathrm{C}_{1\times1}\left(\mathrm{ReLU}\left(\mathrm{C}_{1\times1}\left(X^c_{\mathrm{GAP}}\right)\right)\right) \qquad (6)$$

$$W_s = \mathrm{C}_{7\times7}\left(X^s_{\mathrm{GAP}} \oplus X^s_{\mathrm{GMP}}\right) \qquad (7)$$

where $C_{k\times k}(*)$ denotes a convolution layer with a kernel size of k, $\oplus$ denotes concatenation along the channel dimension, X^s_{GMP}, X^s_{GMP} and X^c_{GAP} denotes global average pooling over the spatial dimension, global max pooling over the spatial dimension, and global average pooling over the channel dimension, respectively. To limit the model complexity, the first 1×1 convolution reduces the number of channels from C to C_r, and the second 1×1 convolution restores the number of channels back to C. To limit the model complexity, the first 1×1 convolution reduces the number of channels from C to C_r, and the second 1×1 convolution restores the number of channels back to C.

Next, we integrate spatial weights and channel weights and perform adaptive per-channel refinement guided by the input features, in order to enhance the model's focus on effective semantic information. The specific process is as follows:

$$W_p = \mathrm{Sigmoid}\left(\mathrm{GC}_{7\times7}\left(\mathrm{CS}\left(X \oplus W_{\mathrm{initial}}\right)\right)\right) \qquad (8)$$

where $W_{initial}$ is obtained by fusing W_c and W_s following the broadcasting principle. X denotes the input features, $\oplus$ denotes concatenation along the channel dimension. CS(*) denotes channel shuffling, $GC_{k\times k}(*)$ refers to a grouped convolution with a kernel size of k × k. The number of channels in X and W is C. The group number is set to C, ensuring that the generated spatial importance weights are based on different channels. W_p denotes the pixel-level weights.

We achieve adaptive fusion of low-level and high-level features by modulating the features with pixel-level weights generated by CAAM. Additionally, input features are added via skip connections to mitigate the gradient vanishing problem. The specific process is as follows:

$$FS = \mathrm{C}_{1\times1}\left(W_p \cdot F_L + \left(\mathbf{E} - W_p\right) \cdot F_H + \left(F_H + F_L\right)\right) \qquad (9)$$

where F_L, F_H, and F_S denotes low-level features, high-level features, and the fused features, respectively. **E** denotes an all-ones matrix with the same size as W_p. $C_{1\times1}$ denotes a 1×1 convolution.

Our CAAM performs adaptive adjustments by combining spatial-level and channel-level information, emphasizing effective semantic information. In addition, the efficient fusion mechanism based on CAAM preserves rich semantic information from different levels.

3 Experiments

The following section outlines the details of our experiments.

3.1 Datesets

We tested on five datasets to evaluate the generalization capability of our MLVP-Net, FaceForensics++ [18], DeepFake Detection Challenge(DFDC) [19], CelebDF-V2(CELEDFv2) [20], DeeperForensics-1.0(DFR) [21], WildDeepfake(WDF) [22].

3.2 Evaluation Metrics

Evaluation Metrics: We use accuracy and the Area Under the ROC Curve (AUC) as the criteria to assess the model's ability to distinguish between real and fake images. Higher values of these metrics indicate stronger performance in detecting forgery.

3.3 Training Parameters

Training Parameters: We use MTCNN [23] for face detection and alignment, resizing the images to 224 × 224 pixels to ensure the model focuses on face-related tasks. For each video, 20 frames are extracted. The Adam optimizer is employed with parameters ($\beta1 = 0.9$, $\beta2 = 0.999$). The learning rate is set to 0.0001 and decays by a factor of 0.5 after each epoch. All models are trained for 20 epochs.

Table 1. Intra-dataset and cross-dataset evaluation. All models were trained on faceforensics++. '*' indicates intra-dataset evaluation. '†' indicates cross-dataset evaluation.

Method	FF++C23*		DFDC†		DFR†		CELEDFV2†		WDF†		AVG	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MAT [24]	93.72	97.26	61.96	65.88	65.59	83.78	70.79	73.87	65.03	73.33	65.84	74.22
RECCE [25]	**95.95**	**98.37**	61.08	65.60	59.67	88.58	71.43	74.12	58.52	70.29	62.68	74.65
F3-Net [26]	93.15	96.74	65.20	67.76	65.09	84.13	72.99	74.03	63.55	72.13	66.71	74.51
SFIC [27]	92.98	93.32	64.83	67.96	56.67	78.82	68.76	70.36	63.01	69.92	63.32	71.77
CADDM [28]	93.96	97.45	64.90	68.72	**67.48**	87.05	68.75	66.83	62.00	71.00	65.78	73.40
IFFD [29]	92.62	94.36	63.61	69.63	67.18	81.78	71.68	72.07	65.78	65.82	67.06	72.33
DFGAZE [30]	79.44	90.90	57.09	**73.54**	53.09	78.09	70.72	78.82	67.02	69.08	61.98	74.88
UMFC [31]	91.61	95.48	63.71	68.64	56.72	83.44	69.63	70.68	64.62	69.81	63.67	73.14
Ours	94.73	97.96	**67.55**	73.24	67.05	**91.54**	**76.05**	**81.51**	**67.40**	**76.28**	**69.51**	**80.64**

3.4 Generalizability evaluation

We evaluate the generalizability of our model by comparing it with recent state-of-the-art (SOTA) methods. The comparison experiments include in-dataset evaluation, cross-dataset evaluation, and cross-operation evaluation.

1) Intra-dataset Evaluation: In Table 1, the best results are highlighted in bold, while the second-best results are underlined. From the in-dataset comparison conducted on

FF++C23, RECEE is identified as the best detector, while our MLVP-net achieves slightly lower performance but still outperforms most other methods.

2) Cross-dataset Evaluation: In Table 1, Cross-dataset performance serves as an evaluation of generalization from the perspective of domain shift, highlighting the model's robustness in real-world deployment scenarios. To verify the effectiveness of MLVP-Net, we conducted cross-dataset evaluations on four benchmark datasets. DFGAZE achieves a slightly higher AUC than our model on the DFDC dataset by 0.3%. Our model performs slightly worse than CADDM and IFFD on the DFR dataset in terms of ACC, with a maximum gap of no more than 0.43%. Apart from these cases, our model consistently outperforms the compared methods across all datasets, demonstrating impressive generalization capability.

3) Multi-source Cross-manipulation Evaluation: In Table 2, Cross-manipulation evaluation assesses the model's generalization ability from the perspective of attack type, emphasizing its robustness against unseen forgery attacks. we select three out of the four different forgery techniques in FF++ as the training set, with the remaining one reserved as the testing set. Although our model does not consistently achieve the highest accuracy (ACC) across all forgery types, it consistently outperforms other methods in terms of AUC. Since AUC reflects the model's discriminative ability across all decision thresholds and is less sensitive to dataset-specific distributions, it provides a more reliable measure of generalization. In contrast, ACC can fluctuate due to changes in the test set distribution or threshold sensitivity. Therefore, the consistently superior AUC achieved by our model indicates strong generalization capability against unseen manipulation methods.

Table 2. Multi-source cross-manipulation evaluation in terms of acc (%) and auc (%).

Method	NT		F2F		FS		NDEEP		AVG	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MAT	56.61	66.19	62.51	69.47	51.88	53.98	80.13	88.32	62.78	69.49
F3-Net	57.85	63.29	**63.85**	70.00	49.68	54.04	82.31	89.50	63.42	69.21
CADDM	55.36	64.62	59.56	68.25	50.85	52.90	70.39	84.92	59.04	67.67
SFIC	**58.70**	63.72	63.53	71.56	50.72	55.81	75.08	90.04	62.01	70.28
OURS	55.78	**66.27**	63.22	**72.10**	**52.70**	**57.00**	**83.76**	**91.92**	**63.87**	**71.82**

Table 3. Comparison with ablation methods on the ff++, dfdc, dfr, celedfv2 and wdf datasets in acc and auc.

Method	FF++C23*		DFDC†		DFR†		CELEDFV2†		WDF†		AVG	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Baseline	89.93	92.14	61.78	65.72	58.88	78.83	66.92	72.10	62.84	65.70	62.61	70.59
Baseline /M	94.51	97.71	64.44	67.79	59.24	88.59	73.37	75.96	66.53	74.02	65.90	76.59
Baseline /M C	93.45	97.48	65.81	70.48	64.92	91.01	75.31	80.12	**67.50**	75.13	68.39	79.19
Baseline /M D	93.79	97.40	66.75	71.86	64.50	89.22	74.81	78.66	64.94	73.51	67.75	78.31
Baseline /M D C	**94.73**	**97.96**	**67.55**	**73.24**	**67.05**	**91.54**	**76.05**	**81.51**	67.40	**76.28**	**69.51**	**80.64**

3.5 Ablations

To verify the reliability of our proposed method, we conducted extensive ablation experiments on the five dataset.

As shown in Table 3, we denote the Multi-scale Spatial Perception Module (MSPM) as M, the Detail Feature Enhancement Module (DFEM) as D, and the Content-Adaptive Attention Module (CAAM) as C. After applying MSPM to the baseline network, the detection accuracy (ACC) and generalization performance (AUC) are significantly improved by 3.29% and 6% on average, respectively. Building upon this, we further introduce CAAM and DFEM individually, both of which bring varying degrees of improvement in accuracy and generalization, demonstrating the effectiveness of each module in our network. Notably, when MSPM, CAAM, and DFEM are combined, the model achieves the best performance in authenticity classification.

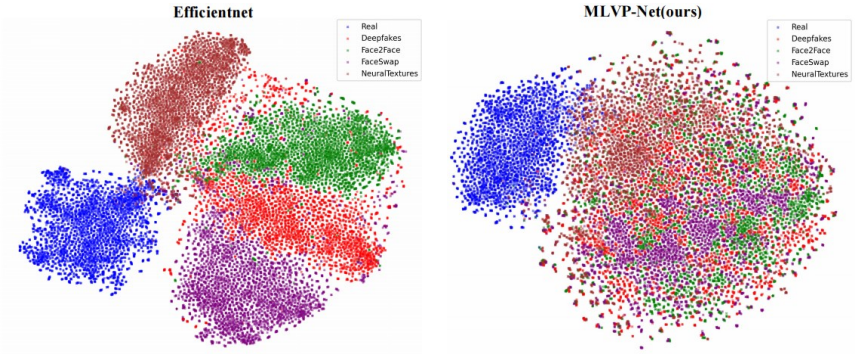

Fig. 8. t-SNE visualization

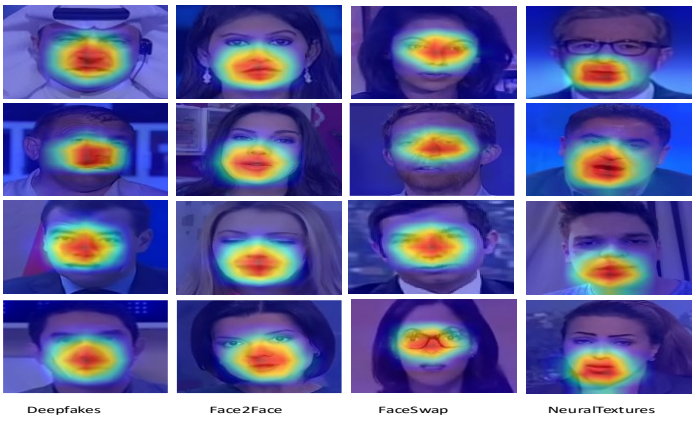

Fig. 9. Grad-CAM++ visualization

3.6 Visualization

In this section, we present of feature distributions and saliency maps to illustrate the advantages of MLVP-Net.

4) T-SNE Feature Distribution: We employ T-SNE [32] to visualize and analyze both the baseline and our model on the FF++ dataset, As illustrated in Fig. 8. The baseline network demonstrates its ability to effectively distinguish between different types of forgeries in the embedding space, indicating that it can learn distinct features specific to each forgery type. However, this tendency to overfit such unique characteristics may lead to rigid reliance, thereby weakening its ability to discern authenticity when encountering unseen forgery types.

In contrast, our proposed MLVP-Net leverages multi-level visual perception to learn common patterns of forgery from multiple perspectives. Evidently, our network tends to group four different types of forgeries into a single cluster in the embedding space, rather than maintaining clear boundaries between them. This reveals a key principle behind our significant improvement in generalization: learning shared characteristics of forgeries helps break rigid reliance.

5) saliency maps: We employed Grad-CAM++ to visualize the manipulated facial regions that are of interest to our network, thus demonstrating the effectiveness of our model. As illustrated in Fig. 9. Our model accurately localizes the manipulated regions and exhibits strong activations even in the absence of ground-truth annotations, while showing minimal responses in non-facial structural areas that are irrelevant to authenticity classification.

4 Conclusion

This paper proposes a deepfake detection model, MLVP-Net, based on multi- level visual perception. Our method introduces an innovative integration across spatial, channel, and gradient dimensions, combined with feature space projection and updating, enabling the model to better focus on the shared characteristics of forgery techniques, thereby enhancing its adaptability and generalization capabilities. Additionally, through effective feature encoding, our model not only efficiently integrates the semantic information from different levels of features but also guides the model to focus on the information most relevant to the deepfake detection task, significantly improving its generalization performance. This addresses the common challenge of limited generalization faced by current deepfake detection models. While our work represents just one step toward advancing deepfake detection techniques, we believe it makes a substantial contribution to improving the adaptability and scalability of deepfake detection tools.

Acknowledgments. This work was supported by the Xinjiang Uygur Autonomous Region Tianshan Talent Program Project under Grant (2022TSYCLJ0036).

References

1. Guo, Z., Wang, L., Yang, W., Yang, G., Li, K.: Ldfnet: Lightweight dynamic fusion network for face forgery detection by integrating local artifacts and global texture information. IEEE Transactions on Circuits and Systems for Video Technology 34(2), 1255–1265 (2023)

2. Li, S., Li, X., Li, Z., Ma, H., Sheng, J., Li, B.: Dual guidance enhancing camouflaged object detection via focusing boundary and localization representation. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
3. Guo, Z., Yang, G., Chen, J., Sun, X.: Exposing deepfake face forgeries with guided residuals. IEEE Transactions on Multimedia 25, 8458–8470 (2023)
4. Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B.: 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 (2024)
5. Guo, Z., Jia, Z., Wang, L., Wang, D., Yang, G., Kasabov, N.: Constructing new backbone networks via space-frequency interactive convolution for deepfake detec- tion. IEEE Transactions on Information Forensics and Security (2024)
6. Zhou, X., Han, H., Shan, S., Chen, X.: Fine-grained open-set deepfake detection via unsupervised domain adaptation. IEEE Transactions on Information Forensics and Security (2024)
7. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). pp. 6105–6114. PMLR (2019)
8. Hu, V.T., Baumann, S.A., Gui, M., Grebenkova, O., Ma, P., Fischer, J., Ommer, B.: Zigma: A dit-style zigzag mamba diffusion model. European Conference on Computer Vision (ECCV) (2024)
9. Su, Z., Lin, H., Zhao, M., Zhu, Z.: Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5097–5107. IEEE (2021)
10. Yu, Z., Qin, Y., Zhao, H., Li, X., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5304. IEEE (2020)
11. Yu, Z., Qin, Y., Zhao, H., Li, X., Zhao, G.: Dual-cross central difference network for face anti-spoofing. In: Zhou, Z.H. (ed.) Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI). pp. 1281–1287 (2021). https://doi.org/10.24963/ijcai.2021/177
12. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: Nas-fas: Static-dynamic central difference network search for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(9), 3005–3023 (Sep 2021)
13. Dong, H., et al.: Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2154–2164 (Jun 2020)
14. Ye, T., et al.: Perceiving and modeling density is all you need for image dehazing. arXiv preprint arXiv:2111.09733 (2021)
15. Bai, H., Pan, J., Xiang, X., Tang, J.: Self-guided image dehazing using progressive feature fusion. IEEE Transactions on Image Processing 31, 1217–1229 (2022)
16. Wu, H., et al.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10546–10555 (Jun 2021)
17. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11908–11915 (2020)
18. 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 (ICCV). pp. 1–11. IEEE (2019)

19. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton Ferrer, C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
20. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3207–3216. IEEE (2020)
21. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2889–2898. IEEE (2020)
22. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2382–2390. ACM (2020)
23. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10), 1499–1503 (2016)
24. Zhao, H., Li, Y., Sun, H., Wang, S., Wang, Y.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2185–2194 (2021)
25. Cao, J., Ma, C., Yao, T., et al.: End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4113–4122 (2022)
26. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 86–103 (2020)
27. Guo, Z., Jia, Z., Wang, L., Wang, D., Yang, G., Kasabov, N.: Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. IEEE Transactions on Information Forensics and Security 19, 401–413 (2024)
28. Dong, S., Wang, J., Ji, R., et al.: Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3994–4004 (2023)
29. Hua, Y., Shi, R., Wang, P., Ge, S.: Learning patch-channel correspondence for interpretable face forgery detection. IEEE Transactions on Image Processing 32, 1668–1680 (2023)
30. Peng, C., Miao, Z., Liu, D., Wang, N., Hu, R., Gao, X.: Where deepfakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection. IEEE Transactions on Information Forensics and Security (2024)
31. Ba, Z., Liu, Q., Liu, Z., Wu, S., Lin, F., Lu, L., Ren, K.: Exposing the deception: Uncovering more forgery clues for deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 719–728 (2024)
32. Bai, H., Pan, J., Xiang, X., Tang, J.: Self-guided image dehazing using progressive feature fusion. IEEE Transactions on Image Processing 31, 1217–1229 (2022)