2025 IEEE International Conference on High Performance Computing and Communications (HPCC) | 979-8-3315-6874-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/HPCC67675.2025.00049

Deepfake Detection via 3D Face Reconstruction–Based Image Blending

Junfeng Xu
School of Computer and Cybger Science
Communication University of China
Beijing, China
junfeng@cuc.edu.cn

Weiguo Lin
School of Computer and Cybger Science
Communication University of China
Beijing, China
linwei@cuc.edu.cn

Mingyang Shao
School of Computer and Cybger Science
Communication University of China
Beijing, China
1423406156@qq.com

Wanshan Xu
School of Computer and Cybger Science
Communication University of China
Beijing, China
xws@cuc.edu.cn

Jing Zhou
School of Computer and Cybger Science
Communication University of China
Beijing, China
zhoujing@cuc.edu.cn

Yikun Xu
School of Computer and Cybger Science
Communication University of China
Beijing, China
xyk@cuc.edu.cn

***Abstract*—Deepfake technologies leverage deep learning to generate highly realistic videos involving face swapping and expression transfer, often exceeding the threshold of human visual perception. This poses serious challenges to social governance and digital security, highlighting the urgent need for reliable forgery detection methods. Training detection models without using real forgeries is considered a promising strategy to improve generalization. These approaches simulate diverse forgery traces to generate synthetic training data. However, most existing methods rely on 2D image manipulation and fail to capture 3D forgery characteristics such as geometric distortion, expression mismatch, and texture anomalies—leading to poor performance on reconstruction-based forgeries. To address this problem, we propose a Reconstruction-Blended Image (RBI) generation method based on 3D Morphable Models (3DMM). By perturbing facial shape and expression parameters, this approach produces training samples that better reflect 3D reconstruction artifacts. When combined with traditional Self-Blended Images (SBI), the hybrid training strategy enhances the model's ability to detect a wider range of forgeries. Experiments show that this method improves AUC by 10.79% on challenging cases like Face2Face forgeries. In summary, our 3D face reconstruction-based generation strategy significantly enhances the generalization and robustness of forgery detection models, offering a practical solution to emerging deepfake threats.**

***Index Terms*—Deepfake, AGI Detection, CV.**

I. Introduction

In 2022, a forged video depicting Ukrainian President Volodymyr Zelensky falsely calling for surrender shocked global audiences online, triggering a political crisis that forced the Ukrainian government to initiate emergency public relations measures. This landmark event revealed the comprehensive threat posed by deepfake technology to modern systems of social governance. As a typical application of AI-generated content (AIGC), deepfake leverages deep neural networks to achieve facial replacement, expression manipulation, and voice cloning. The realism of the generated content has surpassed the visual discrimination threshold of the human eye.

Deepfake technology not only enables the creation of fake audio and video of political figures—affecting public emotions and cognition—but is also exploited in financial fraud and identity theft. These attacks jeopardize corporate economic security and expose the vulnerabilities of traditional identity verification methods. Moreover, deepfakes present significant threats to personal privacy and reputation. Many individuals become targets of malicious forgeries without their knowledge, with their likeness and speech altered to disseminate harmful information, leading to defamation or humiliation. On social media platforms in particular, fake videos often spread far faster than the truth can be clarified, causing immeasurable psychological harm to victims.

Existing forgery detection methods still face significant bottlenecks in terms of cross-domain generalization. This challenge has given rise to the use of synthetic data to train detection models. The core idea is to construct a generalized detection model by simulating the common features of forgery traces, thus overcoming the technical limitations of real forgery samples. For example, mainstream Simulated Blending Images (SBI) methods simulate diverse and generalizable forgery traces to improve the overall performance of detection models. However, SBI exhibits limited performance on some forgery data based on 3D facial reconstruction. This is primarily because, during source image perturbation, SBI struggles to simulate reconstructed skin textures and suffers from insufficient geometric distortion. Additionally, 3D face reconstruction-based face swapping or expression reenactment methods often involve complex facial expression reconstruction and transfer, which tend to produce unnatural distortions. SBI methods, which rely only on affine transformations of the source image, fail to adequately simulate the forgery traces arising from expression mismatches.

This paper proposes a Reconstruction-Blended Images (RBI) generation method based on 3D face reconstruction. While maintaining the same processing pipeline for target im-

979-8-3315-6874-0/25/$31.00 ©2025 IEEE
DOI 10.1109/HPCC67675.2025.00049

ages as in SBI, RBI introduces a reconstruction and perturbation mechanism based on the 3D Morphable Model (3DMM) to the source image, thereby enriching its transformation process. By mixing the two types of data generated by SBI and RBI for training, the detection model can learn a more comprehensive set of forgery traces, ultimately improving its overall detection performance.

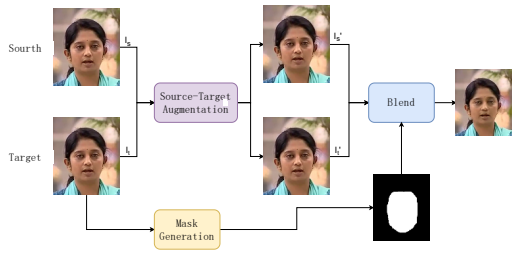

Fig. 1. Overview of generating an SBI.

II. Related Work

Deepfake Detection

Image feature extraction via neural networks has become the mainstream approach. Typical models like XceptionNet [24] and EfficientNet [25] directly classify images using deep architectures. However, these general-purpose backbones struggle to capture forgery-specific cues, so more targeted strategies are needed.

To strengthen feature awareness, Dang et al. [23] introduced an attention mechanism that steers the network toward likely tampered regions, improving detection accuracy. Zhao et al. [17] extended this idea with multi-attention modules, allowing fine-grained feature extraction from several facial subregions and greatly boosting generalization.

Frequency-domain information is another powerful cue for forgery detection. Masi et al. [4] proposed a two-branch network that separately extracts spatial features and high-frequency details, then fuses them to highlight subtle inconsistencies. Li et al. [5] designed an adaptive frequency-domain module that applies discrete cosine transform (DCT) on image patches, enabling the network to learn forgery artifacts across different frequency bands. Wei et al. [6] further combined frequency-aware components with local frequency statistics in their F3-Net, enhancing both robustness and accuracy.

Temporal cues from video sequences also aid detection. Guera et al. [7] were among the first to use a CNN for frame-level features and an RNN to model temporal changes, demonstrating the value of time-domain analysis. Saikia et al. [8] incorporated optical flow as an auxiliary temporal feature within a CNN-RNN framework. Cozzolino et al. [9] observed that forged videos often break the link between identity traits and biometric consistency, so they combined metric learning with adversarial training to build a robust spatiotemporal detector.

A. Training Data Synthesis.

While methods that combine spatial, frequency, and temporal features for forgery detection have made progress in uncovering subtle traces, they still face limitations and fail to significantly improve model generalization. This is mainly because they rely on existing forged datasets, which often contain limited examples and use fixed forgery techniques. For instance, many datasets focus on techniques like GANs [3] or Autoencoders [2], which generate high-quality images but still have inherent limitations. To improve model generalization, several methods have been proposed that don't rely on real forged samples. These methods use self-synthesized data to simulate forgery traces, enabling the model to capture these features and identify forged videos.

Li et al. [11] argued that early Deepfake algorithms generated low-resolution images and used affine transformations to merge faces, causing detectable artifacts. They simulated these artifacts using Gaussian blur and affine transformations, but with improved generation techniques, these artifacts became harder to detect. Li et al. [12] found that most face-swapping methods involve a face blending step, where synthesized faces are merged into the target background. This blending creates inherent differences in the image's boundaries. They proposed the Face-X-Ray method to generate negative samples by blending two real faces based on facial keypoints, which helped improve model learning of blending features. Zhao et al. [13] expanded on Face-X-Ray by adding image enhancements like Gaussian blur and affine transformations to increase the diversity of synthetic images, boosting the model's generalization. Shiohara et al. [14] showed that using more generic, harder-to-detect fake samples improves generalization. They introduced the Self-Blended Images (SBI) method, which blends source and target images by adjusting their spatial and frequency domain features. This approach reduced overfitting to specific forgery methods, improving the model's robustness and generalization in real-world applications. Zhou et al. [15] applied self-blending to temporal sequences, further disrupting the spatiotemporal features of real videos to simulate forged ones, enhancing generalization. Guan et al. [16] added a gradient regularization term during training to help the model find stable network weights that are insensitive to changes in shallow feature statistics. This, combined with SBI data, led to significant performance improvements.

III. Recon Blended Images (RBIs)

This paper proposes RBI, a 3D morphable face model-based blended image generation method. Similar to SBI, the RBI framework consists of three stages: source-target preprocessing, mask generation, and image blending, as illustrated in the workflow diagram. During source image processing, RBI introduces 3D face reconstruction and parameter perturbation to simulate geometric and textural anomalies. This enhancement over SBI methods provides richer training data for detection models, significantly improving their detection performance and cross-domain generalization capabilities.

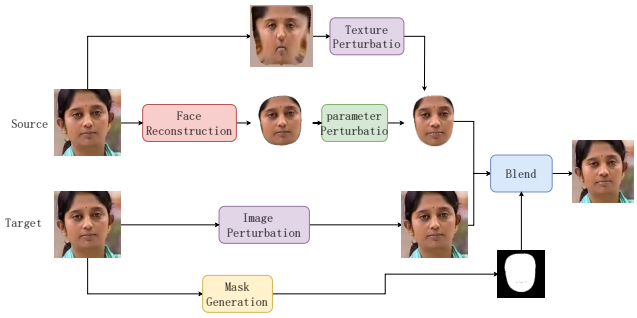

Fig. 2. Overview of generating an RBI.

A. 3D Face Reconstruction and Perturbation

The 3D Morphable Face Model (3DMM) is a facial reconstruction technique that describes 3D facial characteristics through linear combinations of basis vectors. Widely used in computer vision, graphics, facial recognition, and forgery detection, 3DMM operates under the core premise that any 3D face can be represented as a weighted combination of basis vectors capturing shape, texture, and expression variations.

The shape and texture components are formally expressed as:

$$\mathbf{S} = \overline{\mathbf{S}} + \sum_{i=1}^{m} \alpha_i \mathbf{s}_i + \sum_{i=1}^{n} \gamma_i \mathbf{e}_i \tag{1}$$

$$\mathbf{T} = \overline{\mathbf{T}} + \sum_{i=1}^{m} \beta_i \mathbf{t}_i \tag{2}$$

where:

- $\overline{\mathbf{S}}, \overline{\mathbf{T}}$ denote the average shape and texture vectors from the Basel Face Model (BFM) dataset [20]
- $\alpha_i, \beta_i, \gamma_i$ represent shape, texture, and expression parameters respectively
- $\mathbf{s}_i, \mathbf{t}_i, \mathbf{e}_i$ correspond to principal components from BFM (shape/texture) and FaceWarehouse (expression)
- m, n indicate the number of basis vectors for each component

Parameter modifications enable precise control over facial characteristics. For instance, adjusting shape parameters α_i can generate different facial contours (e.g., slender vs round faces).

B. Parameter Prediction and Texture Extraction

This chapter uses the above 3DMM to obtain four types of facial information: shape, texture, expression, and pose. In our implementation, we adopt the weakly supervised 3D face reconstruction algorithm of Deng et al. [21], whose backbone is ResNet-50 followed by a fully connected layer with 257 neurons that outputs a 257-dimensional parameter vector. Specifically, the shape coefficient vector is $\alpha_{\text{id}} \in \mathbb{R}^{80}$, the expression coefficient vector is $\gamma_{\text{exp}} \in \mathbb{R}^{64}$, and the pose parameters form a three-dimensional vector $p \in \mathbb{R}^3$ representing the head's Euler angles. The texture coefficient vector has dimension 80, but in our method we do not directly use these predicted texture parameters; instead, we sample the texture image from the real input.

We first detect facial landmarks on the input image and crop and resize the face to 224×224. The cropped face is then fed into the parameter-prediction network to produce the 257-dimensional face parameter vector. Using the shape parameters in equation (1), we fit a 3D mesh S with 35,709 vertices. To extract the texture, we project each vertex set $V \in \mathbb{R}^3$ onto the 2D UV plane via cylindrical unwrapping, establishing the mapping between 3D coordinates (x, y, z) and UV coordinates (u, v). We then rasterize the mesh onto the image plane using the PyTorch3D renderer and sample the original image I at each projected UV location to fill in the RGB values, thereby constructing the UV texture map T_{uv}.

C. Parameter & Texture Perturbation

After obtaining the source face's shape, expression, and texture parameters, we apply random perturbations to simulate geometric distortions and expression mismatches relative to the target image. At the same time, texture perturbations emulate differences in color, brightness, and contrast between the source and target.

(1) Shape & Expression Perturbation: We linearly blend the source's shape and expression parameters with those of a randomly selected face to introduce subtle variations. The mixing coefficient is drawn from a uniform distribution, creating diverse shape and expression variants. Formally:

$$\alpha_{\text{dist}} = k\,\alpha_{\text{src}} + (1-k)\,\alpha_{\text{rand}}, \quad k \sim U(0.1, 0.8), \tag{3}$$

$$\gamma_{\text{dist}} = k\,\gamma_{\text{src}} + (1-k)\,\gamma_{\text{rand}}, \quad k \sim U(0.1, 0.8). \tag{4}$$

Here, α_{src} and γ_{src} are the source's shape and expression vectors, while α_{rand} and γ_{rand} are sampled from a random target model. Varying k controls the perturbation magnitude.

(2) Pose Perturbation: We add small uniform noise to each dimension of the pose vector p to simulate head-pose inconsistencies introduced during face reconstruction:

$$\Delta p_i \sim U(0.01, 0.05) \times \text{sgn}, \quad \text{sgn} \in \{-1, 1\},\ i \in \{0, 1, 2\}. \tag{5}$$

Here, Δp_i is the perturbation on the i-th Euler angle, and sgn randomly flips the direction. This mimics slight orientation mismatches between source and target.

(3) Texture Map Perturbation: To reproduce statistical differences between source and target, RBI retains SBI's perturbation strategies on the UV texture map T_{uv}. Specifically, we apply spatial and frequency-domain transforms—such as color shifts, brightness/contrast adjustments, and sharpening—so that the reconstructed face exhibits significant deviations at multiple statistical levels. Table III-C details the specific texture perturbation operations.

Type	Method	Description
Color	Color Shift	Apply random shifts to RGB channel $\Delta\mathrm{RGB} \sim U(-20, 20)$
	Hue & Saturation Adjustment	Perturb hue, saturation, and value in HSV space where $\Delta H, \Delta S, \Delta V \sim U(-0.3, 0.3)$
	Brightness & Contrast Adjustment	Adjust via linear transform $Tr_{\mathrm{uv}} = T_{\mathrm{uv}} \cdot c + b$ where $c \sim U(0.9, 1.1),\ b \sim U(-0.1, 0.1)$
Frequency	Sharpening	Apply Laplacian kernel for sharpening sharpening strength $\alpha \sim U(0.2, 0.5)$
	Downsampling	Randomly downsample to 1/2 or 1/4 size then upsample back to original resolution

TABLE I
TEXTURE PERTURBATION METHODS

D. Mask Generation and Image Blending

We first generate the initial blending mask M in the same manner as SBI. During the process of integrating the reconstructed face into the target image, we must remove the black background from the rendered face. To do so, we use the face model's background mask M_{bg} to distinguish foreground (facial region) from background, and then refine the blending mask M to obtain the final mask image M_{final}:

$$M_{\mathrm{final}} = M \odot \left(1 - M_{\mathrm{bg}}\right) \tag{6}$$

Once the final mask is obtained, we linearly blend the perturbed face rendering I_{render} with the target image I_t to produce the final composite image:

$$I_{\mathrm{blend}} = M_{\mathrm{final}} \odot I_{\mathrm{render}} + \left(1 - M_{\mathrm{final}}\right) \odot I_t \tag{7}$$

IV. EXPERIMENTS

A. Experimental Setting

Datasets. The primary dataset used in our experiments is FaceForensics++ (FF++) [10], which contains 1 000 publicly available original videos, each lasting around ten seconds. FF++ provides four different deepfake generation methods: FaceSwap, Deepfakes, Face2Face, and NeuralTextures. FaceSwap and Deepfakes perform face replacement, whereas Face2Face and NeuralTextures target facial expression manipulation. Each method produces 1 000 forged videos, for a total of 4 000 fakes. Moreover, FF++ includes three compression levels: C0 (lossless, original), C23 (high quality), and C40 (low quality). With its large volume, high fidelity, and diverse manipulation techniques, FF++ has become a standard benchmark for deepfake video detection. We train and evaluate our model on the official FF++ train/validation/test splits. Because we employ a forgery-free training regime, only real video frames are used during training; in testing we evaluate on both the real and forged videos. To assess cross-dataset generalization, we evaluate on three additional benchmarks: Celeb-DF [26], DFDCP, and DFDC [1] .

Preprocessing. We extract frames from each FF++ real video (8 frames per video). Faces are detected with RetinaFace, and the largest face per frame is cropped and landmarked. Landmark and bounding-box data are saved for later cropping and mask generation. Next, each cropped face is passed through the 3D reconstruction network to predict 257-dimensional parameters and to compute its UV texture map via mesh projection and sampling; all parameters and textures are stored for subsequent perturbation and rendering.

Training. Leveraging both SBI's statistical perturbations and RBI's reconstruction-based perturbations, we propose a mixed-training strategy: the training set comprises 70% SBIs and 30% RBIs. SBI and RBI share the same spatial and frequency perturbation ranges (e.g. $\Delta\mathrm{RGB} \sim U(-20, 20)$). For each training example, we randomly choose between the SBI or RBI pipeline to generate it, ensuring the model learns both 2D and 3D forgery cues. Validation uses real frames from the FF++ validation split; we save the model weights at the highest validation AUC for final testing. We employ a pre-trained EfficientNet-B4 backbone, with input size 380×380, optimized using Sharpness-Aware Minimization. The initial learning rate is 1×10^{-2} with momentum 0.9. Standard data augmentations (horizontal flips, random rotations, scaling, and compression) are applied. All experiments run on an NVIDIA RTX 3090 GPU.

B. Evaluation

1) In-Dataset Evaluation: To evaluate the impact of the synthetic data proposed in this chapter on model detection performance, several representative algorithms in the field of deepfake detection are selected for comparison. First, in-dataset testing is conducted on the FF++ dataset to compare performance across four types of forgery methods. The evaluation metric is AUC (%), and the results are shown in Table VI.

Experimental results show that when the detection model is trained on a dataset combining SBI and RBI methods, it achieves the best performance on Deepfake, Face2Face, and FaceSwap. In particular, for the Face2Face method, which is based on 3D facial reconstruction

2) Cross-Dataset Evaluation: To evaluate our method, we conducted cross-dataset evaluation on different datasets. The results are shown in Table IV-B2. It can be observed that the mixed training approach using both SBI and RBI achieves the best performance on the CDF and DFDC datasets, and performs close to the best on the DFDCP dataset.

This demonstrates that combining SBI and RBI during training can effectively enhance the generalization ability of

Method	AUC (%)				
	DeepFake	Face2Face	FaceSwap	NeuralTextures	Avg
Face X-Ray [12]	-	-	-	-	87.35
SBI [14]	<u>97.62</u>	84.96	93.92	**80.58**	<u>89.28</u>
RBI	87.28	<u>95.18</u>	**98.97**	58.54	84.99
SBI+RBI	**98.81**	**95.75**	<u>98.83</u>	<u>78.06</u>	**92.85**

TABLE II
THE IN-DATASET EVALUATION PERFORMANCE ON THE HIGH-QUALITY VERSION (C23) OF THE FF++ DATASET.BOLDANDUNDERLINEDVALUESCORRESPONDTOTHEBESTANDTHESECOND-BESTVALUE, RESPECTIVELY.

deepfake detection models. In particular, the strong performance on more challenging datasets like CDF and DFDC highlights the potential of our method in cross-domain forgery detection tasks.

Method	AUC (%)		
	CDF	DFDCP	DFDC
F3-Net [6]	71.21	-	-
Two-branch [4]	73.41	-	-
DAM [18]	75.30	72.80	-
LipForensics [19]	72.49	67.17	-
Face X-Ray [12]	-	80.92	-
PCL+I2G [13]	90.03	74.37	67.52
SBI [14]	<u>90.90</u>	**86.88**	<u>76.50</u>
PFake [15]	90.17	85.01	-
RBI	87.33	62.43	64.45
SBI+RBI	**94.27**	<u>85.29</u>	**77.50**

TABLE III
CROSS-DATASETE VALUATION ON CDF,DFDC,DFDCP.OUR METHOD ACHIEVES THE BEST PERFORMANCE ON THE CDF AND DFDC DATASETS, AND THE SECOND-BEST PERFORMANCE ON THE DFDCP DATASET.

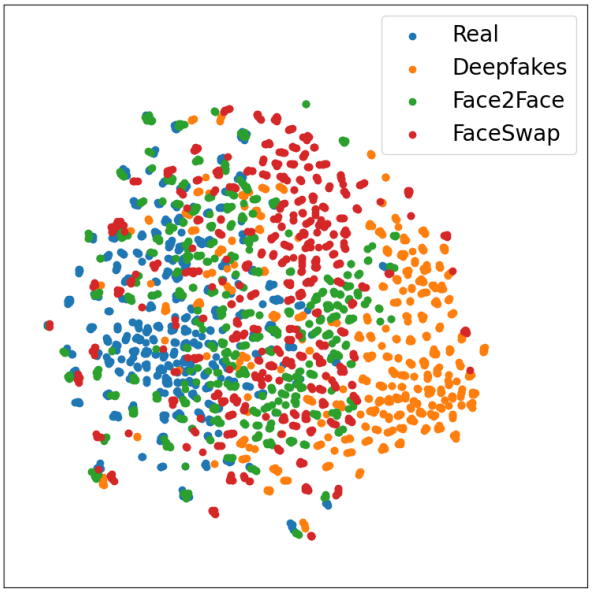

Fig. 3. SBI only

C. Ablation Studies

1) SBI/RBI Mixing Ratio: We vary the SBI:RBI ratio and evaluate detection on FF++ (Table IV-C1). When RBI comprises only 10%, the model underfits reconstruction cues; at 30% RBI, performance peaks, showing complementary benefits of 2D and 3D perturbations. At 100% RBI, performance drops, indicating the necessity of general SBI disturbances.

2) Perturbation Component Analysis: We remove shape, expression, pose, and texture perturbations individually and evaluate on 3D-based forgeries (Table V). Removing expression perturbation impacts Face2Face most (AUC drops from 95.18 to 91.23), highlighting its importance. Shape and pose removal also degrade performance, while texture perturbation has the largest overall effect, confirming the strength of statistical inconsistencies.

3) Backbone Comparison: Although we default to EfficientNet-B4 as the primary backbone, our mixed-data strategy is architecture-agnostic. As shown in Table VI, the proposed method achieves consistently high performance across various backbone networks, including ResNet and ConvNeXt. This demonstrates the generality and adaptability of our approach, making it compatible with a wide range of model architectures commonly used in deepfake detection tasks.

D. Feature Space Visualization

Figure 4 presents the t-SNE [22] visualization of the final-layer feature embeddings for different training strategies. When trained solely on SBI data, the model exhibits poor feature separation between real images and 3D-reconstructed forgeries, indicating limited ability to capture reconstruction-specific cues. In contrast, the model trained with a mixture of SBI and RBI data demonstrates a much clearer separation between real and fake samples in the embedding space. This result confirms the effectiveness of the proposed RBI augmentation, which introduces distinctive artifacts associated with 3D reconstruction, thereby enhancing the model's capability to generalize to more challenging forgery types.

Proportion	AUC (%)						
	Deepfake	Face2Face	FaceSwap	NeuralTextures	CDF	DFDCP	DFDC
0%	97.62	84.96	93.92	**80.58**	90.90	86.88	76.50
10%	98.72	90.36	97.09	79.03	94.19	**87.88**	76.42
30%	**98.81**	**95.75**	98.83	78.06	94.27	85.29	**77.5**
50%	98.72	95.01	98.63	78.46	**94.33**	85.48	75.69
100%	87.28	95.18	**98.97**	58.54	87.33	62.43	64.45

TABLE IV
EXPERIMENTS WITH DIFFERENT MIXING RATIOS OF SBI AND RBI WERE CONDUCTED, WHERE THE PROPORTION OF RBI WAS SET TO 0%, 10%, 30%, 50%, AND 100%, RESPECTIVELY.

Shape	Expression	Pose	Texture	FaceSwap	Face2Face	Avg
✓	✓	✓	✓	98.97	**95.18**	**97.08**
-	✓	✓	✓	98.37	91.23	94.80
✓	-	✓	✓	**99.12**	93.94	96.53
✓	✓	-	✓	98.72	91.34	95.03
✓	✓	✓	-	95.43	83.27	89.35

TABLE V
ABLATION STUDY ON PERTURBATION METHODS (AUC). THE CONTRIBUTION OF DIFFERENT PERTURBATION FACTORS TO THE MODEL'S DETECTION PERFORMANCE WAS EVALUATED BY INDIVIDUALLY REMOVING SHAPE, EXPRESSION, POSE, AND TEXTURE PERTURBATIONS. THE RESULTS SHOW THAT THESE COMBINED PERTURBATIONS HAVE A SIGNIFICANT IMPACT ON THE PERFORMANCE OF FORGERY DETECTION.

backbone	AUC (%)				
	DeepFake	Face2Face	FaceSwap	NeuralTextures	Avg
ResNet-50	97.83	95.88	97.67	77.70	92.27
Convnext	98.01	93.49	97.27	79.02	91.95
EfficientNet-b4	98.81	95.75	98.83	78.06	92.85

TABLE VI
COMPARATIVE EXPERIMENTS WITH DIFFERENT BACKBONE NETWORKS

V. DISCUSSION

This paper proposes an innovative improvement to video forgery detection methods that require no forged samples for training. By introducing a 3D Morphable Model (3DMM) to construct a parameter perturbation strategy, the approach effectively overcomes the limitations of traditional self-blending image methods, which struggle to simulate forgery traces caused by 3D facial reconstruction. Experimental results demonstrate that the method significantly enhances the detection accuracy for reconstruction-based forgeries such as Face2Face and FaceSwap, by simulating geometric distortion, expression mismatches, and texture anomalies of the source face. Notably, the hybrid training strategy combining SBI and RBI further improves the model's generalization performance on cross-domain datasets such as Celeb-DF, highlighting the potential of self-synthesized data in addressing unknown forgery techniques.

ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation of China under Grants U2436208, 62302467 and 62402459, Fundamental Research Funds for the Central Universities (ID: CUC22GZ034), Public Computing Cloud.

REFERENCES

[1] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer, "The DeepFake Detection Challenge Dataset," CoRR, vol. abs/2006.07397, 2020.

[2] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv:1312.6114, 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," arXiv:1406.2661, 2014.

[4] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch Recurrent Network for Isolating Deepfakes in Videos," ArXiv, vol. abs/2008.03412, 2020.

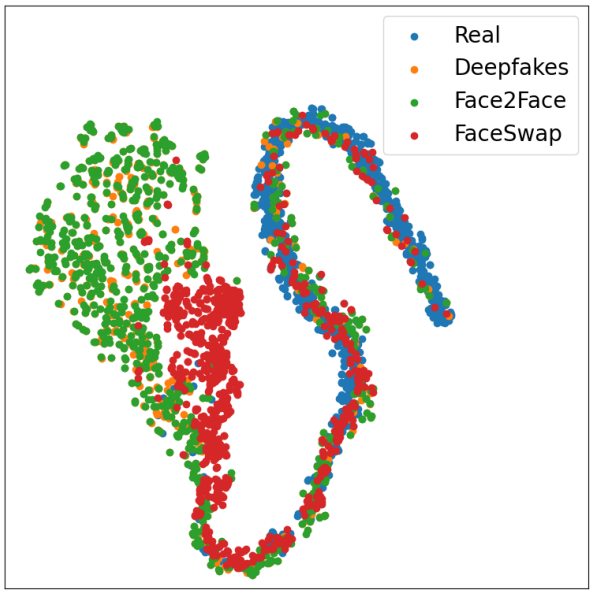

Fig. 4. SBI only

[5] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6454-6463.

[6] J. Wei, S. Wang, and Q. Huang, "F^3Net: fusion, feedback and focus for salient object detection," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, 2020, pp. 12321–12328.

[7] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), 2018, pp. 1–6.

[8] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features," in 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1-7.

[9] D. Cozzolino, A. Rossler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15108–15117.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11.

[11] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[12] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for More General Face Forgery Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[13] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning Self-Consistency for Deepfake Detection," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15003-15013.

[14] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18699-18708.

[15] J. Guan, H. Zhou, M. Gong, Y. Zhao, E. Ding, and J. Wang, "Detecting Deepfake by Creating Spatio-Temporal Regularity Disruption," ArXiv, vol. abs/2207.10402, 2022.

[16] W. Guan, W. Wang, J. Dong, and B. Peng, "Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 5345-5356, 2024.

[17] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional Deepfake Detection," arXiv:2103.02406, 2021.

[18] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face Forensics in the Wild," arXiv:2103.16076, 2021.

[19] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection," arXiv:2012.07657, 2021.

[20] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009, pp. 296–301.

[21] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 285-295.

[22] L. van der Maaten and G. E. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

[23] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," arXiv:1910.01717, 2020.

[24] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," arXiv:1610.02357, 2017.

[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv:1905.11946, 2020.

[26] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," arXiv:1909.12962, 2020.