GLOBAL RESEARCH IMMERSION PROGRAM
FOR YOUNG SCIENTISTS

Innovations in AI Security: Detection of AI-Forged Images

Xinran Hu
University of Illinois, Urbana_Champaign
xinranh6@illinois.edu

1 Objectives

Present Recent Technological Advances: Highlight the latest advancements in neural network technologies for detecting AI-generated images, emphasizing their significance in AI cybersecurity.

Introduce Three State-of-the-Art Models: Face X-ray; Frequency in Face Forgery Network; and Two-Stream Network.

Demonstrate Effectiveness and Encourage Research: Present experimental results comparing the models' accuracy and inspire further development in Deepfake detection technologies to enhance cybersecurity.

2 Background

In recent years, the technique of swapping people's faces, commonly known as Deepfakes, has raised significant public concerns. Deepfakes leverage advanced deep learning algorithms to create highly realistic fake images and videos, often indistinguishable from authentic ones to the human eye [1]. This capability has been misused for malicious purposes, such as spreading misinformation, defaming individuals, and creating deceptive media content for profit [2].

The rapid development of deep learning and AI technologies has contributed to these manipulations. There are four primary types of face manipulations: entire face synthesis, attribute manipulation, identity swap, and expression swap [3]. The detection of such fake faces has become a hot research topic, with significant implications for cybersecurity, digital forensics, and public trust. This poster aims to introduce three state-of-the-art models in the area of deepfake detection. These models utilize neural networks to identify AI-generated pictures with high precision. Each model has undergone extensive experimentation and has demonstrated remarkable outcomes in terms of accuracy, offering promising solutions to the growing challenge of deepfake detection.

3 Methods

3.1 Face X-ray

A forgery image can often be decomposed into two images from different sources: one for the face and one for the background. The Face X-ray model excels at detecting the blending area where these two images are combined, allowing it to determine whether an image is real or forged [4].

To achieve this, a mask is defined during the blending process, which helps the model identify the boundaries and transitions between the face and the background. This capability enables the Face X-ray model to detect subtle inconsistencies in the blending areas, making it highly effective in identifying Deepfakes.

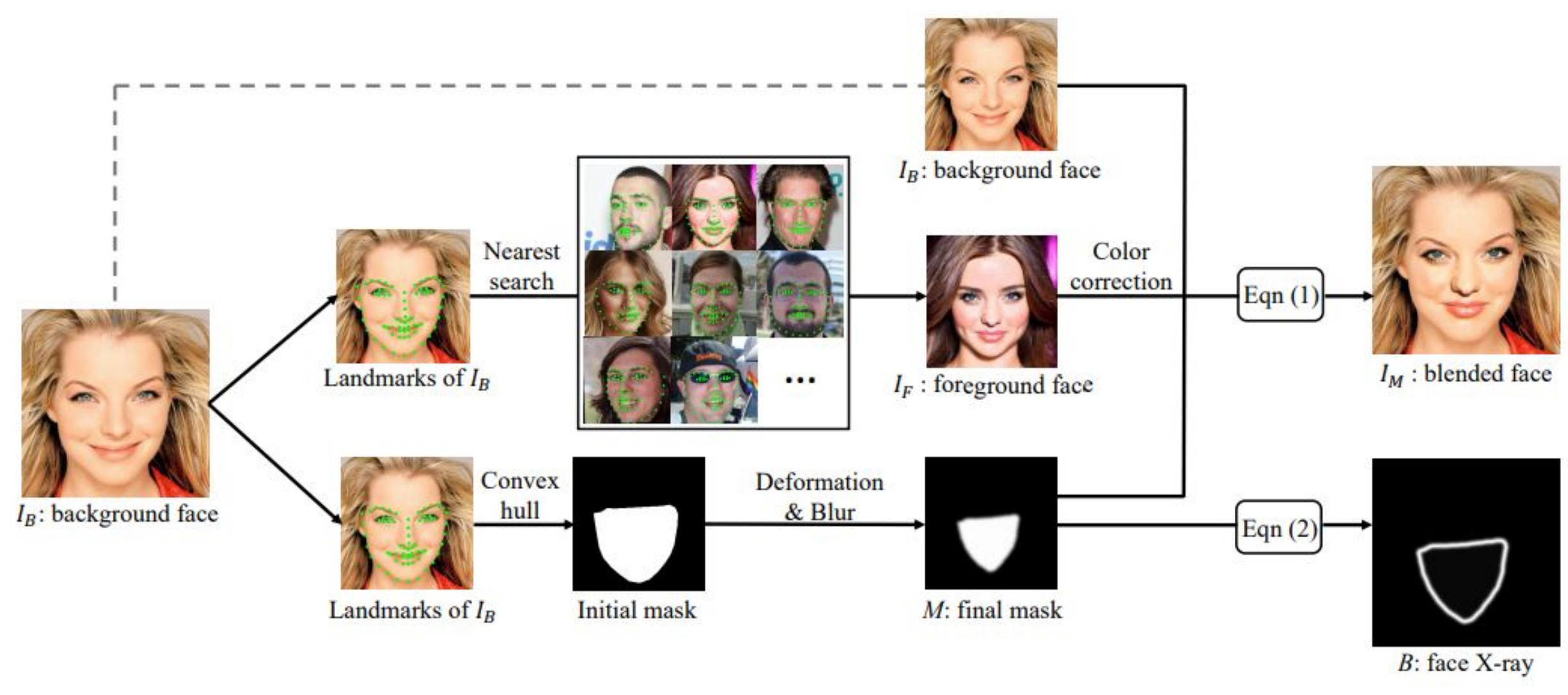

Figure 1. The process of training a sample using Face X-ray.

3.2 Frequency in Face Forgery Network

The Frequency in Face Forgery Network (F3-Net) leverages the frequency domain to detect subtle forgery features often overlooked in the spatial domain. Former studies primarily focused on the spatial domain of images, but F3-Net utilizes two complementary approaches: frequency-aware decomposed (FAD) image components and local frequency statistics (LFS) [5].

The image is first decomposed into frequency components through FAD, which are then filtered and transformed back to the spatial domain. These transformed components are examined by a convolutional neural network to detect inconsistencies. Simultaneously, LFS fully exploits the frequency domain by analyzing local frequency distributions to identify abnormal patterns indicative of forgery. The two branches are connected by a module called MixBlock, which facilitates interaction between the spatially transformed frequency components and the local frequency statistics. This integration ensures a comprehensive analysis, improving the model's ability to detect subtle forgeries and enhancing overall detection accuracy. By combining spatial and frequency domain analysis, F3-Net demonstrates significant advancements in identifying AI-generated forgeries.

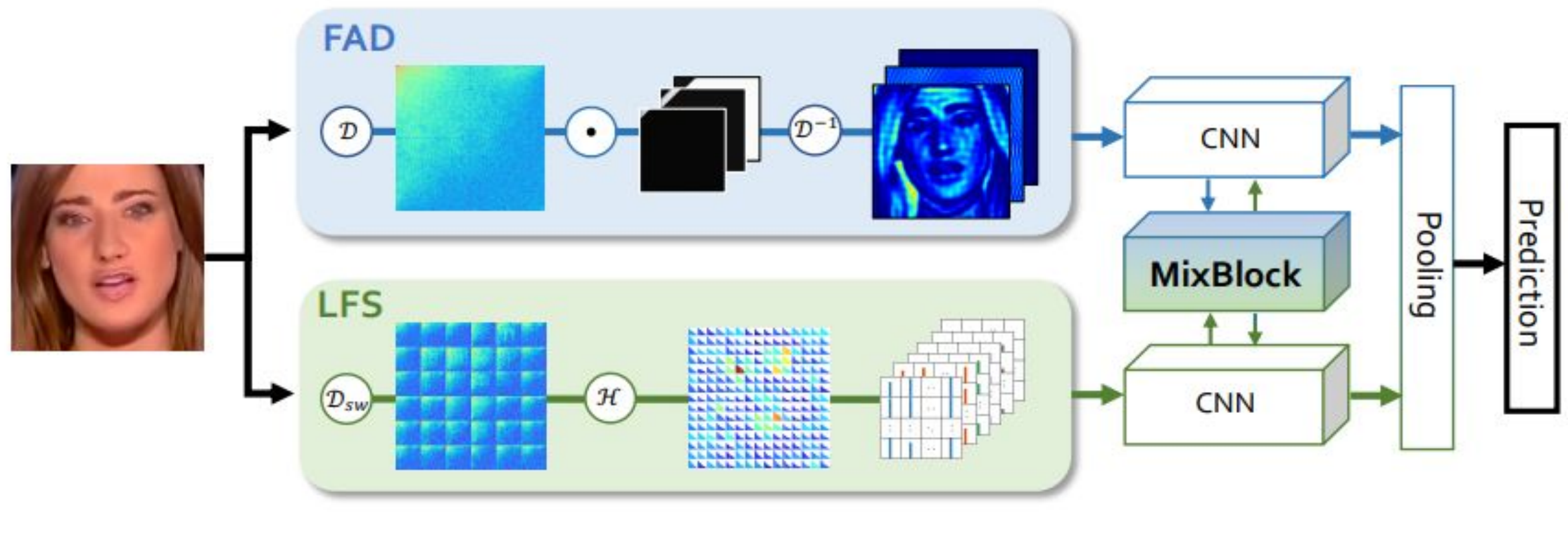

Figure 2: Overview of F3-net.

3.3 Locate and Verify

This model employs a two-stream network combined with three innovative modules and strategic approaches to identify potential forgery regions with high precision [6]. The model takes two input streams: the Spatial Rich Model (SRM) and supplemental high-frequency component. To effectively combine these two streams, the Cross-modality Consistency Enhancement (CMCE) module is employed, ensuring consistency between the spatial and frequency domains while maintaining their characteristics. The Local Forgery Guided Attention (LFGA) module directs the model's focus towards manipulated regions, improving detection accuracy by prioritizing areas more likely to contain forgeries. Additionally, the Multi-scale Patch Feature Fusion (MPFF) module aids in detecting forgeries at different scales, particularly at the shadow level, allowing the model to capture finer detail.

Furthermore, the model utilizes a Semi-supervised Patch Similarity Learning (SSPSL) strategy to estimate location annotations. By focusing on sensitive facial patches, such as nose, eyes, and mouth, the model can approximate the distribution of manipulated regions and improve its detection accuracy.

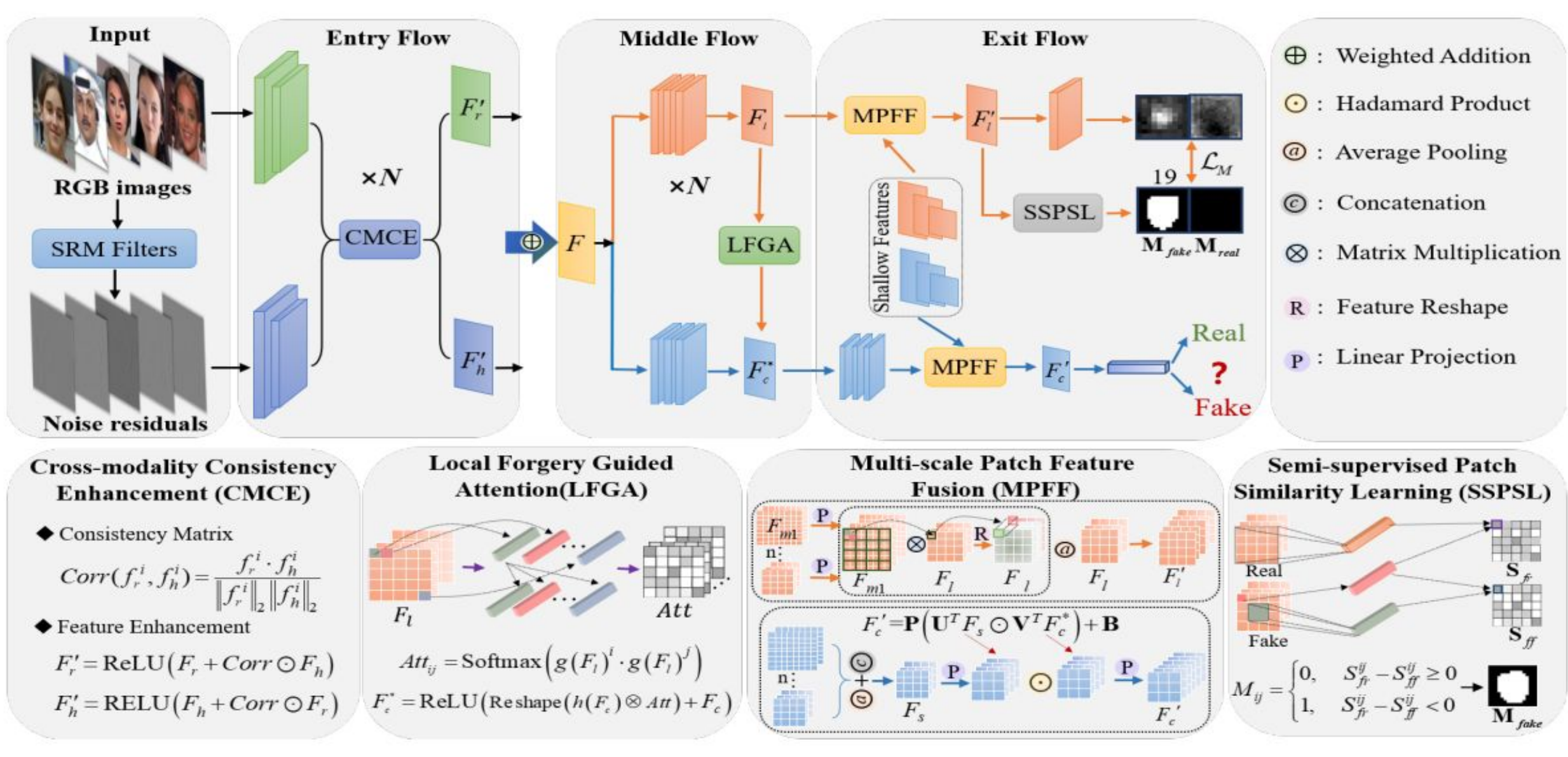
Input
RGB images
SRM Filters
Noise residuals
Entry Flow
×N
CMCE
F_r'
F_h'
Middle Flow
F
×N
F_l
LFGA
F_c^*
Exit Flow
MPFF
F_l'
Shallow Features
SSPSL
19
$\mathcal{L}_M$
$\mathbf{M}_{fake}$ $\mathbf{M}_{real}$
MPFF
F_c'
Real
?
Fake
⊕ : Weighted Addition
⊙ : Hadamard Product
@ : Average Pooling
© : Concatenation
⊗ : Matrix Multiplication
R : Feature Reshape
P : Linear Projection
Cross-modality Consistency Enhancement (CMCE)
◆ Consistency Matrix
$Corr(f_r^i, f_h^i) = \frac{f_r^i \cdot f_h^i}{\|f_r^i\|_2 \|f_h^i\|_2}$
◆ Feature Enhancement
$F_r' = \mathrm{ReLU}(F_r + Corr \odot F_h)$
$F_h' = \mathrm{RELU}(F_h + Corr \odot F_r)$
Local Forgery Guided Attention(LFGA)
F_l
Att
$Att_{ij} = \mathrm{Softmax}(g(F_l)^i \cdot g(F_l)^j)$
$F_c^* = \mathrm{ReLU}(\mathrm{Reshape}(h(F_c) \otimes Att) + F_c)$
Multi-scale Patch Feature Fusion (MPFF)
F_{m1}
n:
P
R
F_l
F_l'
$F_c' = \mathbf{P}(\mathbf{U}^T F_s \odot \mathbf{V}^T F_c^*) + \mathbf{B}$
F_s
F_c'
Semi-supervised Patch Similarity Learning (SSPSL)
Real
Fake
$\mathbf{S}_{fr}$
$\mathbf{S}_{ff}$
$M_{ij} = \begin{cases} 0, & S_{fr}^{ij} - S_{ff}^{ij} \geq 0 \\ 1, & S_{fr}^{ij} - S_{ff}^{ij} < 0 \end{cases}$
$\mathbf{M}_{fake}$

Figure 3: Overview of the two-stream network.

4 Results

The Face X-ray model demonstrates high accuracy and strong generalizability across various deepfake scenarios, significantly outperforming previous models, especially in detecting blended images.

The F3-net model achieves impressive results with an accuracy rate of more than 95% in the FaceForensics++ low-quality dataset. An ablation study on each component revealed that all components contributed to a significant increase in accuracy. The model specializes in low-quality tasks, making it particularly effective in real-world applications where image quality may vary.

The two-stream model shows significant improvements. The model's Area Under Curve (AUC) improved from 0.797 to 0.835 on the Deepfake Detection Challenge preview, and from 0.811 to 0.847 on the CelebDF_v1 dataset at the video level.

5 Discussion

The Face X-ray model's high accuracy and generalizability highlight its robustness in detecting blended Deepfake image. However, it struggles with fully synthetic images, pointing to an area for future improvement. F3-Net's dual-branch approach, leveraging frequency domain analysis, effectively uncovers forgery features missed in the spatial domain, demonstrating high precision and recall rates. The two-stream network shows significant improvements in AUC on major datasets. It improves focus on manipulated regions and underscore the importance of multi-scale and multi-domain approaches in advancing deepfake detection.

6 Conclusion

The advancements presented by the Face X-ray, F3-Net, and two-stream network mark significant strides in the field of Deepfake detection. The Face X-ray model excels in identifying blended images, though it requires further development to handle fully synthetic images. The F3-Net model proves effective in low-quality scenarios and demonstrates the power of leveraging the frequency domain, achieving high precision and recall through its dual-branch approach. The two-stream network showcases substantial improvements in detection accuracy and robustness by integrating spatial and high-frequency components with innovative modules like CMCE, LFGA, and MPFF, complemented by the SSPSL strategy. Collectively, these models enhance our capabilities in identifying AI-generated forgeries, underscoring the importance of continuous research and multi-faceted approaches in AI cybersecurity. Future work should focus on addressing current limitations and expanding the models' applicability to a broader range of Deepfake scenarios.

7 References

[1] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv:1812.08685 [cs]*, Dec. 2018, Available: https://arxiv.org/abs/1812.08685
[2] R. Cellan-Jones, "Deepfake videos 'double in nine months,'" *BBC News*, Oct. 07, 2019. Available: https://www.bbc.com/news/technology-49961089
[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, Dec. 2020, doi: https://doi.org/10.1016/j.inffus.2020.06.014.
[4] L. Li *et al.*, "Face X-ray for More General Face Forgery Detection," Apr. 2020.
[5] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues," Oct. 2020.
[6] C. Shuai *et al.*, "Locate and Verify: A Two-Stream Network for Improved Deepfake Detection," *Proceedings of the 31st ACM International Conference on Multimedia*, Oct. 2023, doi: https://doi.org/10.1145/3581783.3612386.