Seeing Through Deepfakes: A Human-Inspired Framework for Multi-Face Detection

Supplementary Material

1. Comparisons of the cross-compression detection performance

In practice, the compression factor of a video that we need to detect may be unknown. To assess our model's performance across different compression levels, we conduct experiments involving cross-compression. Specifically, DF-Platter provides videos compressed at levels C0, C23, and C40. We train our model on videos compressed at C23 and test it on videos compressed at C0 and C40. As shown in Table 1, our method improves frame-level detection performance in both "C23 to C40" and "C40 to C23" scenarios. Notably, SBI achieves a 1.3% improvement in FLAC and 0.5% in FLAU in the "C23 to C40" scenario. Although the lower visual quality of C40 videos typically impacts detection, our method consistently enhances per-face frame-level detection by an average of 4.8% in "C23 to C0" and 1.1% in "C23 to C40", thanks to the integration of four contextual feature modules.

2. Effects of human factors

Our method leverages human-inspired features, so it is crucial to assess detection performance when these factors are removed. Our method contains four modules, i.e., the Spatial-Temporal module (M1), Pixel-Wise module (M2), Gaze module (M3), and Body-Face module (M4). These modules are related to different human factors. For M1, the temporal component is inspired by the human observation of inconsistencies across frames. In M2, the face comparison element is based on how humans compare faces within a frame to detect abnormalities. Both M3 and M4 are entirely driven by human insights, focusing on gaze inconsistencies and age/gender mismatches.

To evaluate the impact of these human factors, during our ablation studies, we exclude the temporal component from M1, the comparison element from M2, and both M3 and M4, respectively. As shown in Table 2, removing these elements results in degraded detection performance. When using the original group inference network [4] without our custom designs, the performance significantly decreases. These findings demonstrate that human-inspired features and our specific designs enhance detection performance.

3. Effects of different baseline networks

We utilize a group inference network [4] for training, with its default backbone, VGG [3], to extract features relevant to each module. To assess the impact of different backbone networks, we experiment with VGG [3], ResNet [1], and Xception [2]. As shown in Table 3, the choice of backbone network has a slight influence on detection performance, likely because our human-inspired features are effectively extracted by all three networks. ResNet slightly outperforms VGG and Xception.

4. Effects of different fusion strategy

We employ the XOR fusion method to integrate the outputs of the four detection modules. It is crucial to assess the impact of different fusion strategies on detection performance. Specifically, we evaluated three strategies: AND operator fusion, average fusion, and XOR operator fusion.

In the AND operator fusion, a face is classified as fake only if all modules detect it as fake. In the average fusion, a face is classified as fake if the average score from the four modules indicates it is fake. Lastly, in the XOR operator fusion, a face is classified as fake if any one of the four modules detects it as fake.

The results, as shown in Table 4, demonstrate that the XOR operator fusion strategy yields the best detection performance. The AND operator fusion and average fusion strategies fall short primarily because the gaze and face-body modules capture features that the spatial-temporal and pixel-wise modules do not. Consequently, we adopt the XOR fusion strategy in our final method.

Furthermore, we evaluate false negative rate: 7.8% (FFIW), 1.0% (OpenForensics), and 4.4% (DF-Platter). The lower FNR confirms that 'XOR' fusion helps reduce missed false faces in multi-face detection, improving per-face frame-level accuracy.

5. Effects of four modules

We conduct experiments using the Spatial-Temporal module (M1), Pixel-Wise module (M2), Gaze module (M3), and Body-Face module (M4) independently. The outputs of M1 and M2 are binary labels, and the outputs of M3 and M4 are binary labels or NA. NA represents a scenario that does not extract devised gaze or body-face features. The results are presented in Table 5. While both M1 and M2 achieved acceptable performance on their own, M3 and M4 did not perform as well.

The suboptimal performance of M3 is primarily due to its focus on detecting abnormal gaze patterns, specifically cases where most individuals are looking at the camera, while a few are not. This module does not account for

Method	C23 TO C0				C23 TO C40			
	FLAC	FLAU	PFAC	PFAU	FLAC	FLAU	PFAC	PFAU
Xception	86.7	88.3	76.7	77.8	86.2	88.0	72.2	72.9
SBI	96.6	98.5	90.1	91.2	**88.6**	90.4	70.1	71.3
NoiseDF	90.2	90.6	80.1	82.8	82.0	81.5	68.6	69.4
TALL	94.9	98.6	92.9	93.6	88.3	89.4	72.1	72.6
Li et al.	92.3	93.9	83.9	84.6	83.2	83.8	70.2	70.8
S-MIL	91.4	92.9	79.1	79.6	83.3	84.8	70.4	71.2
Zhou et al.	90.8	91.7	80.2	80.3	84.3	85.5	70.8	71.6
Ma et al.	95.4	96.7	89.4	90.9	87.0	87.7	70.4	71.2
FILTER	96.4	97.8	89.8	90.7	86.8	87.9	71.9	73.6
COMISC	93.5	94.4	89.4	91.3	88.1	89.4	72.0	73.4
Ours	**98.0**	**98.8**	**95.4**	**96.3**	88.2	**90.5**	**73.2**	**74.7**

Table 1. Comparisons of the cross-compression detection performance on DF-Platter.

Removing human factors	FFIW		OpenForencics		DF-Platter	
	PFAC	PFAU	PFAC	PFAU	PFAC	PFAU
M1 w/out temporal	86.3	88.2	93.0	94.9	90.0	90.5
M2 w/out comparison	89.0	90.3	94.4	95.8	91.5	92.8
Whole method w/out M3 and M4	89.7	90.2	95.3	96.8	91.9	92.7
Original Group inference network [4]	80.2	82.1	85.5	85.7	82.1	83.7
Whole method	**91.3**	**92.1**	**97.8**	**98.9**	**93.5**	**94.6**

Table 2. Ablation study - Detection performance in removing human factors.

Network	FFIW		OpenForencics		DF-Platter	
	PFAC	PFAU	PFAC	PFAU	PFAC	PFAU
VGG	90.3	91.2	97.0	98.4	93.8	**94.4**
Xception	90.2	91.0	97.2	98.6	93.2	94.3
Resnet	**90.9**	**91.4**	**97.8**	**98.8**	**93.9**	**94.4**

Table 3. Ablation study - Detection performance in different networks.

Fusion	FFIW		OpenForencics		DF-Platter	
	PFAC	PFAU	PFAC	PFAU	PFAC	PFAU
AND Operator	86.3	88.5	93.3	94.9	90.3	90.6
Average Fusion	88.0	90.2	94.4	95.8	91.2	92.6
XOR Operator	**91.3**	**92.1**	**97.8**	**98.9**	**93.5**	**94.6**

Table 4. Ablation study - Detection performance in different fusion strategies.

all possible gaze variations, limiting its effectiveness when used in isolation. Furthermore, this gaze anomaly pattern is observed in current multi-face deepfake datasets but may not generalize to all datasets.

Module	FFIW		OpenForencics		DF-Platter	
	PFAC	PFAU	PFAC	PFAU	PFAC	PFAU
M1	87.7	89.0	93.9	95.6	90.4	91.1
M2	87.6	89.8	93.3	95.9	90.4	91.5
M3	69.5	70.9	78.7	79.9	72.2	73.9
M4	62.8	63.9	67.1	68.4	68.4	69.7
M1+M2+M3+M4	**91.3**	**92.1**	**97.8**	**98.9**	**93.5**	**94.6**

Table 5. Ablation study - Detection performance in different modules.

We explored the relationship between gazes but found dataset constraints limited the cue extraction. Specifically, converging gaze detection of multi-faces requires camera-related parameters (e.g., distance from the camera), which are absent in existing multi-face deepfake datasets

Similarly, M4, which detects discrepancies in age and gender between the face and body, did not achieve strong results on its own. This is because some manipulated faces do not exhibit mismatched age or gender, making them detectable by M1 and M2 instead. Although the results in original manuscripts show M4 contributes to overall perfor-

mance, it is less effective when used independently.

6. Single-face detection comparisons

Method	FF++
Xception	96.3
SBI	99.6
TALL	**99.9**
Zhou et al.	99.5
Ma et al.	95.6
MoNFAP	98.6
Ours	99.5

Table 6. Single-face detection comparisons on FF++ datasets.

Our method can be adapted to single-face scenarios. Specifically, we modify M1 to extract only spatial-temporal features and M2 to focus solely on single-face pixel features, removing multi-face dependencies. M3 is excluded as it is not applicable, while M4 remains unchanged. We evaluate this adaptation on the FF++ [2] dataset. Since FILTER and COMICS rely on multi-face learning and cannot operate on single-face scenarios directly, their results are not included here. Although the performance on FF++ becomes saturated, the results in Table 6 demonstrate competitive performance in single-face detection.

7. Computational cost

Method	FLOPs
Xception	$8.1 * 10^9$
SBI	$8.4 * 10^9$
NoiseDF	$4.7 * 10^9$
Li et al.	$3.8 * 10^9$
S-MIL al.	$3.1 * 10^{10}$
Zhou et al.	$5.6 * 10^{10}$
Ma et al.	$3.9 * 10^9$
FILTER	$6.2 * 10^9$
COMISC	$4.3 * 10^9$
M1	$8.7 * 10^9$
M2	$1.7 * 10^{10}$
M3	$1.8 * 10^9$
M4	$2.7 * 10^9$
Ours	$3.0 * 10^{10}$

Table 7. Computational cost (FLOPs) of the proposed method.

To assess the computational cost in terms of FLOPs (Floating Point Operations), we calculate the cost for each major component of the model. The total FLOPs for the model are obtained by summing the FLOPs for each layer, including convolutional, fully connected, and other operations.

We utilize Python libraries, such as fvcore, to compute the FLOPs based on the model architecture and input size. The results in Table 7 demonstrate that the computational cost of our method is acceptable. Notably, the FLOPs of Zhou et al. [5] are higher than those of other methods, which can be attributed to the use of temporal modules in approaches, as these temporal modules typically increase the FLOPs. Overall, the computational cost of HICOM remains manageable for practical applications.

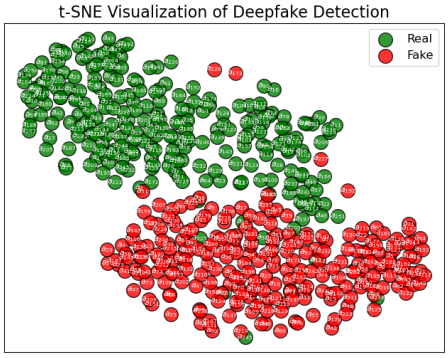

Figure 1. Visualization of M1.

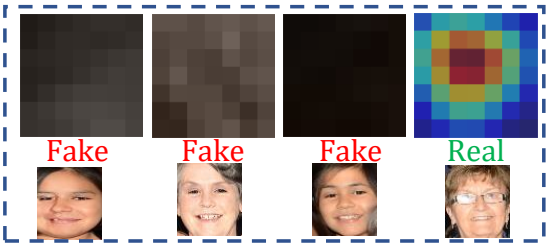

Figure 2. Visualization of M2.

8. Visualization

We visualize the M1 and M2 and show results in Fig. 1 and Fig. 2. Figures illustrate that the M1 and M2 can detect real and fake faces effectively. Since M3 and M4 focus on gaze, age, and body, we provide the accuracy of gaze/age/gender prediction accuracies of 99.2%,98.6% and 99.7% without visualizations.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforen-

sics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 1, 3

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[4] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *ICCV*, pages 7476–7485, 2021. 1, 2

[5] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *CVPR*, pages 5778–5788, 2021. 3