WUHAN UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

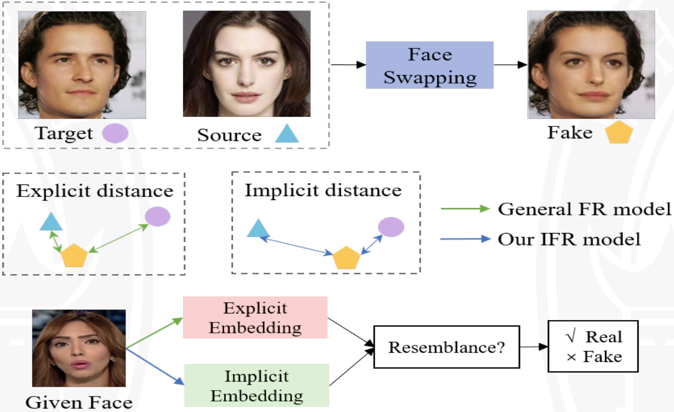
# Implicit Identity Driven Deepfake Face Swapping Detection (TUE-PM-034)

*Baojin Huang†, Zhongyuan Wang†, Jifan Yang†, Jiaxin Ai†*
*Qin Zou†, Qian Wang‡, Dengpan Ye‡*

*†NERCMS, School of Computer Science, Wuhan University*
*‡School of Cyber Science and Engineering, Wuhan University*

May 30, 2023

# Motivation

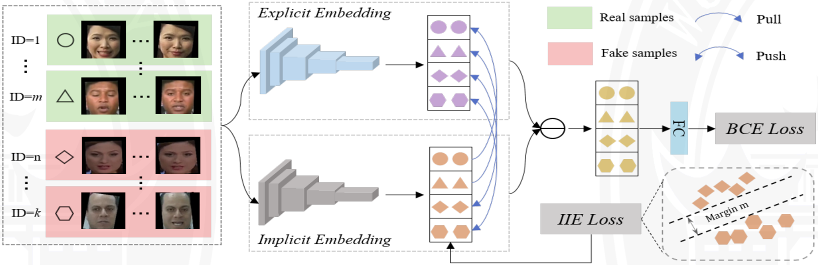



With the similarity between explicit and implicit embeddings of the given face, we can significantly distinguish it as real and fake, which facilitates forgery detection.

# Contribution

- From a completely new perspective, we propose the implicit identity driven framework for face swapping detection, which explores the implicit identity of fake faces. This enhances the deep network to distinguish fake faces with unknown manipulations.
- We specially design explicit identity contrast (EIC) loss and the implicit identity exploration (IIE) loss. EIC aims to pull real samples closer to their explicit identities and push fake samples away from their explicit identities. IIE is margin-based and guides fake faces with known target identities to have small intra-class distances and large inter-class distances.
- Extensive experiments and visualizations demonstrate the superiority of our method over the state-of-the-art approaches.

# Implicit Identity Driven Framework




The outline of our proposed implicit identity driven framework for deepfake face swapping detection. We hybridize real face samples (green boxes) and fake face samples (red boxes) as training set.

# Explicit Identity Contrast

$$\mathcal{L}_{\mathrm{eic}} = \frac{1}{N_F}\sum_{i\in F}\delta\left(F_{im}\left(x_i\right), F_{em}\left(x_i\right)\right) - \frac{1}{N_R}\sum_{i\in R}\delta\left(F_{im}\left(x_i\right), F_{em}\left(x_i\right)\right), \quad (1)$$

where $R$ and $F$ indicate the set of real and fake samples, respectively. $N_R$ and $N_F$ denote the number of real samples and fake samples, respectively. $\delta\left(\cdot,\cdot\right)$ represents the cosine similarity calculation function, which is defined as $\delta(u, v) = \frac{u}{\|u\|} \cdot \frac{v}{\|v\|}$.

# Implicit Identity Exploration

$$\mathcal{L}_{iie}^{+} = -\mathbb{E}_{x_i, y_i \sim \mathcal{K}} \left[ \log \frac{e^{s\left(\cos\left(\theta_{y_i}\right) - m\right)}}{e^{s\left(\cos\left(\theta_{y_i}\right) - m\right)} + \sum_{j \neq y_i} e^{s \cos \theta_j}} \right]. \tag{2}$$

Here, $\theta_j$ represents the angle between normalized $F_{im}(x_i)$ and the normalized proxy of j-th identity on the hypersphere. $s$ and $m$ stand for feature rescale and margin hyperparameter, respectively.

$$m_{fake} = \alpha \cdot \frac{1}{N_r} \sum_{i \in R_{mini}} \cos\left(\theta_{y_i}\right), \tag{3}$$

where $R_{mini}$ denotes the set of real samples for a mini-batch. $N_r$ represents the number of samples in $R_{mini}$. $\alpha$ is a hyperparameter to limit the maximum value of the margin, which is empirically set to 0.5. The margin $m_{real}$ for the real sample is set to a fixed value of 0.4.

# Implicit Identity Exploration

During the implicit identity embedding network forward propagation, we calculate the distance between sample $x_i$ and unknown identities in the lookup table by cosine similarity, denoted as $V^T F_{im}(x_i)$. During backward, we update the $y_i^*$-th column in the lookup table by $v_{y_i^*} \leftarrow \beta v_{y_i^*} + (1-\beta) F_{im}(x_i)$, where $\beta \in [0, 1]$. Moreover, we define the probability that sample $x_i$ is classified as $y_i^*$ by the Softmax function and maximize the expected log-likelihood

$$\mathcal{L}_{iie}^{-} = -\mathbb{E}_{x_i, y_i^* \sim \mathcal{U}} \left[ \log \frac{e^{\left(v_{y_i^*}^T F_{im}(x_i)/\tau\right)}}{\sum_{j=1}^{Q} e^{\left(v_j^T F_{im}(x_i)/\tau\right)}} \right]. \tag{4}$$

The higher temperature $\tau$ leads to softer probability distribution.

# Ablation Study

| Model | $\mathcal{L}_{eic}$ | $\mathcal{L}_{iie}$ | Celeb-DF ACC (%) | Celeb-DF AUC (%) | DFDC ACC (%) | DFDC AUC (%) |
|---|---|---|---|---|---|---|
| A | | | 70.34 | 74.09 | 69.85 | 72.65 |
| B | ✓ | | 77.76 | 82.24 | 76.39 | 78.80 |
| C | | ✓ | 76.40 | 81.46 | 74.95 | 77.22 |
| D | ✓ | ✓ | **79.16** | **83.80** | **79.37** | **81.23** |

**Table 1:** Effectiveness of the proposed constraints in our method on the Celeb-DF and DFDC datasets. Specifically, $\mathcal{L}_{eic}$ and $\mathcal{L}_{iie}$ denote the EIC loss and IIE loss, respectively.

The best performance is achieved when combining all the proposed constraints with 79.16%, 83.80% ACC and 79.37%, 81.23% AUC on Celeb-DF and DFDC, respectively.

# Cross-dataset Evaluation

| Method | FF++ | | Celeb-DF | | DFD | | DFDC | |
|---|---|---|---|---|---|---|---|---|
| | AUC (%) | EER (%) | AUC (%) | EER (%) | AUC (%) | EER (%) | AUC (%) | EER (%) |
| Xception [1] | 99.09 | 3.77 | 65.27 | 38.77 | 87.86 | 21.04 | 69.90 | 35.41 |
| EN-b4 [2] | 99.22 | 3.36 | 68.52 | 35.61 | 87.37 | 21.99 | 70.12 | 34.54 |
| Face X-ray [3] | 87.40 | - | 74.20 | - | 85.60 | - | 70.00 | - |
| MLDG [4] | 98.99 | 3.46 | 74.56 | 30.81 | 88.14 | 21.34 | 71.86 | 34.44 |
| F3-Net [5] | 98.10 | 3.58 | 71.21 | 34.03 | 86.10 | 26.17 | 72.88 | 33.38 |
| MAT(EN-b4) [6] | 99.27 | 3.35 | 76.65 | 32.83 | 87.58 | 21.73 | 67.34 | 38.31 |
| GFF [7] | 98.36 | 3.85 | 75.31 | 32.48 | 85.51 | 25.64 | 71.58 | 34.77 |
| LTW [8] | 99.17 | 3.32 | 77.14 | 29.34 | 88.56 | 20.57 | 74.58 | 33.81 |
| Local-relation [9] | **99.46** | 3.01 | 78.26 | 29.67 | 89.24 | 20.32 | 76.53 | 32.41 |
| DCL [10] | 99.30 | 3.26 | 82.30 | 26.53 | 91.66 | 16.63 | 76.71 | 31.97 |
| UIA-ViT [11] | 99.33 | - | 82.41 | - | **94.68** | - | 75.80 | - |
| Ours | 99.32 | **2.99** | **83.80** | **24.85** | 93.92 | **14.01** | **81.23** | **26.80** |

**Table 2:** Cross-database evaluation from FF++(C23) to Celeb-DF, DFD, and DFDC in terms of AUC and EER. The FF++ belongs to the intra-testing results while others represent to the unseen dataset testing.

# Cross-manipulation Evaluation

| Train | Method | DF | FS | FST | Mean |
|---|---|---|---|---|---|
| DF | EN-b4 | 99.97 | 46.24 | 51.26 | 65.82 |
| | MAT | 99.92 | 40.61 | 45.39 | 61.97 |
| | GFF | 99.87 | 47.21 | 51.93 | 66.34 |
| | DCL | **99.98** | 61.01 | 68.45 | 76.48 |
| | Ours | 99.51 | **63.83** | **73.49** | **78.94** |
| FS | EN-b4 | 69.25 | 99.89 | 60.76 | 76.63 |
| | MAT | 64.13 | 99.67 | 57.37 | 73.72 |
| | GFF | 70.21 | 99.85 | 61.29 | 77.12 |
| | DCL | 74.80 | **99.90** | 64.86 | 79.85 |
| | Ours | **75.39** | 99.73 | **66.18** | **80.43** |
| FST | EN-b4 | 61.11 | 56.19 | 99.52 | 72.27 |
| | MAT | 58.15 | 55.03 | 99.16 | 70.78 |
| | GFF | 61.48 | 56.17 | 99.41 | 72.35 |
| | DCL | 63.98 | 58.43 | 99.49 | 73.97 |
| | Ours | **65.42** | **59.50** | **99.50** | **74.81** |

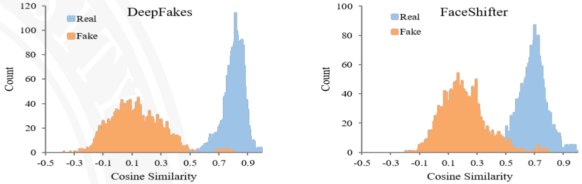
**Table 3:** Cross-manipulation evaluation in terms of AUC. Diagonal results indicate the intra-testing performance. DF, FS and FST denote the DeepFakes, FaceSwap and FaceShifter datasets, respectively.

# Multi-source manipulation Evaluation

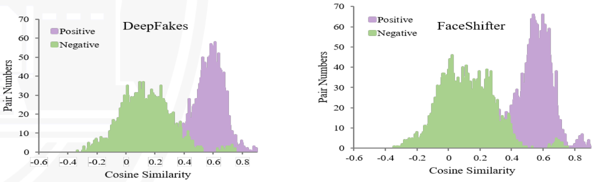
| Method | GID-DF (C23) | | GID-DF (C40) | | GID-F2F (C23) | | GID-F2F (C40) | |
|---|---|---|---|---|---|---|---|---|
| | ACC (%) | AUC (%) | ACC (%) | AUC (%) | ACC (%) | AUC (%) | ACC (%) | AUC (%) |
| EfficientNet [2] | 82.40 | 91.11 | 67.60 | 75.30 | 63.32 | 80.10 | 61.41 | 67.40 |
| Focalloss [12] | 81.33 | 90.31 | 67.47 | 74.95 | 60.80 | 79.80 | 61.00 | 67.21 |
| ForensicTransfer [13] | 72.01 | - | 68.20 | - | 64.50 | - | 55.00 | - |
| Multi-task [14] | 70.30 | - | 66.76 | - | 58.74 | - | 56.50 | - |
| MLDG [4] | 84.21 | 91.82 | 67.15 | 73.12 | 63.46 | 77.10 | 58.12 | 61.70 |
| LTW [8] | 85.60 | 92.70 | 69.15 | 75.60 | 65.60 | 80.20 | 65.70 | 72.40 |
| DCL [10] | 87.70 | 94.9 | 75.90 | 83.82 | 68.40 | 82.93 | 67.85 | **75.07** |
| Ours | **88.21** | **95.03** | **76.90** | **84.55** | **69.36** | **84.37** | **67.99** | 74.80 |

**Table 4:** Performance on multi-source manipulation evaluation. GID-DF means traning on the other three manipulated methods of FF++ and test on DeepFakes. The same for the others.

# Visualization




Cosine similarity distribution of explicit and implicit identities for real and fake samples.




Cosine similarity distribution for positive and negative samples.

If you have any questions or concerns, please do not hesitate to email:

huangbaojin@whu.edu.cn


Code will be available soon…

# References i


- A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *IEEE ICCV*, 2019, pp. 1–11.
- M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*. PMLR, 2019, pp. 6105–6114.
- 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 *IEEE CVPR*, 2020, pp. 5001–5010.
- D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI*, vol. 32, no. 1, 2018.

- J. Wei, S. Wang, and Q. Huang, "F$^3$net: fusion, feedback and focus for salient object detection," in *AAAI*, vol. 34, no. 07, 2020, pp. 12 321–12 328.
- H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *IEEE CVPR*, 2021, pp. 2185–2194.
- Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *IEEE CVPR*, 2021, pp. 16 317–16 326.
- K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu, L. Shao, and R. Ji, "Domain general face forgery detection by learning to weight," in *AAAI*, vol. 35, no. 3, 2021, pp. 2638–2646.
- S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *AAAI*, vol. 35, no. 2, 2021, pp. 1081–1088.

- K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in *AAAI*, vol. 36, no. 2, 2022, pp. 2316–2324.
- W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *ECCV*. Springer, 2022.
- T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE CVPR*, 2017, pp. 2980–2988.
- D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *arXiv preprint arXiv:1812.02510*, 2018.

# References iv

H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2019, pp. 1–8.