

**JOURNAL OF INFORMATION
SYSTEM AND TECHNOLOGY
MANAGEMENT
(JISTM)**www.gaexcellence.com/jistm**IDENTITY-REFERENCED TWO-STAGE DEEPPFAKE
DETECTION WITH CONSISTENCY-GATED FUSION**Zhen Yin^{1*}, Robiatul A'dawiah Jamaluddin²¹Kuala Lumpur University of Science and Technology, Malaysia 243925358@s.klust.edu.my <https://orcid.org/0009-0001-0433-0141>²Kuala Lumpur University of Science and Technology, Malaysia robiatul@klust.edu.my <https://orcid.org/0009-0006-2617-0069>

*Corresponding Author

Article Info:**Article history:**

Received date: 18.08.2026

Revised date: 30.08.2026

Accepted date: 27.09.2026

Published date: 30.09.2026

To cite this document:

Zhen, Y., & Jamaluddin, R. A. (2026). Identity-Referenced Two-Stage Deepfake Detection with Consistency-Gated Fusion. *Journal of Information System and Technology Management*, 11 (44), 904-916.

Abstract:

Deepfake detection methods that rely on manipulation-specific artifacts tend to overfit the forgery types observed during training and generalize poorly to unseen manipulations and datasets. In this paper, we propose an identity-referenced two-stage framework that detects face forgeries by comparing a test face against a reference face of the claimed identity, rather than by memorizing generator-specific artifacts. Unlike earlier identity-referenced detectors, which assume that the original identity that was replaced is known at test time and fine-tune the identity encoder, our framework relies only on the claimed identity, which is available in deployment (for example from an ID photo), and keeps the identity encoder frozen. In the first stage, the frozen MobileFaceNet encoder, pre-trained with ArcFace for face recognition, produces an identity-consistency score between the reference and the test face. In the second stage, a trained EfficientNet-B0 encoder extracts forgery artifact. The two scores are then combined through a consistency-gated fusion, which combines the two scores equally when they are consistent and progressively suppresses the forgery score when they strongly disagree. Experiments on FaceForensics++ and Celeb-DF show that the identity-consistency signal generalizes across manipulation types and datasets: on the most challenging cross-manipulation case (FaceSwap), the identity reference attains 77.8% video-level AUC while the forgery branch collapses to 43.4%, and the fusion attains 67.7%, outperforming the classical baselines by 7 to 20 percentage points on FaceSwap. In cross-dataset evaluation on Celeb-DF, the fusion reaches 91.3% video-level AUC, competitive with strong recent detectors while using only 6.0M parameters. These results demonstrate that identity consistency is a robust prior for identity-swapping manipulations, and that using it directly as a standalone detection score, combined with an independent artifact branch through a validation-defined gate, is an effective and

lightweight route to cross-manipulation and cross-dataset generalization.

DOI: 10.35631/JISTM.1144053 **Keyword:**

Deepfake Detection; Identity Reference; Face Manipulation Detection; Cross-Manipulation Generalization; Cross-Dataset Generalization; Consistency-Gated Fusion



© The authors (2026). This is an Open Access article distributed under the terms of the Creative Commons Attribution (CC BY NC) (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact jistm@gaexcellence.com.

Introduction

The rapid advancement of face forgery techniques, such as face swapping and face reenactment, has made it increasingly difficult to distinguish manipulated videos from authentic ones. Deepfake detection has therefore become an essential task for preventing the misuse of synthetic media in identity fraud, disinformation, and social engineering. However, a well-known limitation of most detectors is their poor generalization: a model trained on one forgery type or dataset frequently degrades sharply when applied to an unseen manipulation technique or a different dataset.

The root cause of this weakness is that many detectors rely on forgery-specific artifacts, such as boundary seams, blending traces, or frequency anomalies, that are characteristic of a particular generation method. Once the generation method changes, these artifacts no longer hold, and the detector collapses. For example, on the FaceForensics++ benchmark, classical detectors such as Xception and CNN-GRU achieve only about 50% AUC on the FaceSwap manipulation when trained on the other three manipulation types.

In this paper, we address cross-manipulation and cross-dataset generalization from the perspective of identity consistency. The key observation is that, for identity-swapping manipulations, a forged face usually no longer matches the identity it claims to be. Identity is therefore a strong prior for identity-swapping manipulations: it does not depend on any specific generator and remains informative even when low-level artifacts are absent. We propose an identity-referenced two-stage detection framework that compares a test face against a reference face of the claimed identity. Unlike earlier identity-referenced detection (Shen et al., 2022), which assumes that the identity that was manipulated away is known at test time and fine-tunes the identity encoder on the target manipulation, our framework requires only the claimed

identity and keeps the identity encoder frozen. A frozen MobileFaceNet encoder thereby provides a stable identity-consistency score, while a trained EfficientNet-B0 encoder captures forgery artifacts. A consistency-gated fusion combines the two scores so that the fusion falls back to the identity baseline when the forgery branch becomes unreliable.

Our main contributions are summarized as follows. (1) We formulate deepfake detection as an identity-consistency verification problem against the claimed identity. This is more than a change of notation: earlier identity-referenced detection (Shen et al., 2022) assumes that the manipulated-away identity is known at test time, whereas the claimed identity is the identity that the content is presented as, so the required reference is available in deployment (for example from an ID photo) and is never taken from the forged video itself. (2) We decouple the identity-consistency score and the forgery-artifact score into two independent branches and fuse them with a consistency gate whose threshold is taken once from the validation split. Because the gate depends only on the two scores and this validation-defined threshold, it needs no calibration on the target test set, and the same fusion rule can be applied unchanged across manipulation types and datasets; when the two branches conflict, the gate suppresses the forgery branch and falls back to the identity baseline. (3) We provide quantitative evidence on FaceForensics++ and Celeb-DF that the identity branch stays stable where the artifact branch collapses on the hardest cross-manipulation case (FaceSwap) the fusion attains 67.7% video-level AUC, 7 to 20 percentage points above the classical baselines, and 91.3% cross-dataset AUC on Celeb-DF, with only 6.0M parameters.

Literature Review

Deepfake detection has been studied from several perspectives. Artifact-based methods detect low-level inconsistencies left by generators (Wang et al., 2020). Xception and related CNN detectors learn forgery traces directly from face images, while frequency-based methods such as F3Net (Qian et al., 2020) mine frequency-domain clues. However, these methods tend to overfit the forgery types seen during training, as the artifacts are generator specific.

Identity-based detection is a more recent direction. Shen et al. (2022) proposed identity-referenced deepfake detection with contrastive learning, using MobileFaceNet as the identity backbone and a contrastive loss to pull apart real and fake embeddings of the same identity. They reported near-perfect cross-manipulation performance on FaceForensics++, but that result was obtained under a different and easier setting than the one considered here: the reference was the original (target) identity that the manipulation replaced, and the identity encoder was fine-tuned on the target manipulation. Such a setting assumes that the manipulated-away identity is known at test time, which is rarely the case in deployment, where only the claimed identity, that is, the identity that the content is presented as, is available. Our method therefore operates under a strictly harder setting: it uses the claimed identity as the reference and keeps the identity encoder frozen. Compared with Shen et al., our method also differs in three further respects: it treats identity consistency as a standalone detection score rather than as a learned representation, it adds an independent forgery-artifact branch, and it fuses the two scores through a validation-defined consistency gate. The original contribution of this work is thus the combination of these complementary signals, namely a frozen identity reference used directly as a detection score, an independently trained artifact branch, and a gate that regulates them, rather than the use of an identity reference as such, which has been explored before.

Contrastive learning has been widely used to learn representations that are robust to domain shift. Paired-consistency approaches, such as PCL+I2G (Zhao et al., 2021), and self-supervised methods, such as RealForensics (Haliassos et al., 2022) and Self-Blended Images (Shiohara & Yamasaki, 2022), improve cross-dataset generalization by exploiting consistency or by augmenting training data with synthetic blends. These methods show that removing identity and background confounders is critical for generalization. Our work shares this motivation but uses an explicit identity reference rather than an implicit consistency constraint.

Cross-manipulation and cross-dataset evaluation have become the standard protocols for measuring generalization. Following prior work, we adopt the leave-one-out protocol on FaceForensics++ and the cross-dataset protocol (train on FaceForensics++, test on Celeb-DF). Under these protocols, we show that the identity-consistency branch remains stable where the artifact-based forgery branch collapses.

Table 1 positions the present work against the closest related methods. The dimensions that separate them are the reference that is available at test time, whether the identity encoder is trained or kept frozen, and the role that the identity signal plays in the final decision. Prior identity-based detectors use the identity signal either as an auxiliary training constraint or as a learned representation inside a single branch. The present framework instead keeps the identity encoder frozen, uses its output directly as a standalone detection score, and fuses that score with an independently trained forgery-artifact branch through a consistency gate. This combination, rather than the use of an identity reference as such, is the contribution of this work.

Table 1. Positioning Of The Proposed Method Relative To The Closest Related Work.

Method	Reference face at test time	Identity encoder	Role of the identity signal
PCL+I2G (Zhao et al., 2021)	None (self-consistency)	Not applicable	Auxiliary training constraint only; not used as a detection score
Shen et al. (2022)	Original (target) identity	MobileFaceNet, fine-tuned	Single branch; the identity representation is the detector
RealForensics (Haliassos et al., 2022)	None	Cross-modal self-supervision	Auxiliary training signal only
Self-Blended Images (Shiohara & Yamasaki, 2022)	None	Self-blended augmentation	Not used explicitly
Ours	Claimed identity (for example from an ID photo)	MobileFaceNet, frozen	Standalone detection score, fused with an independent artifact branch by a consistency gate

Methodology

Overview

The proposed framework receives a test face x and a reference face r of the claimed identity. It produces two complementary scores: an identity-consistency score and a forgery score, which are combined by a consistency-gated fusion into a final decision. The two stages are trained separately and are coupled only at inference time, through the gate. Figure 1 illustrates the architecture.

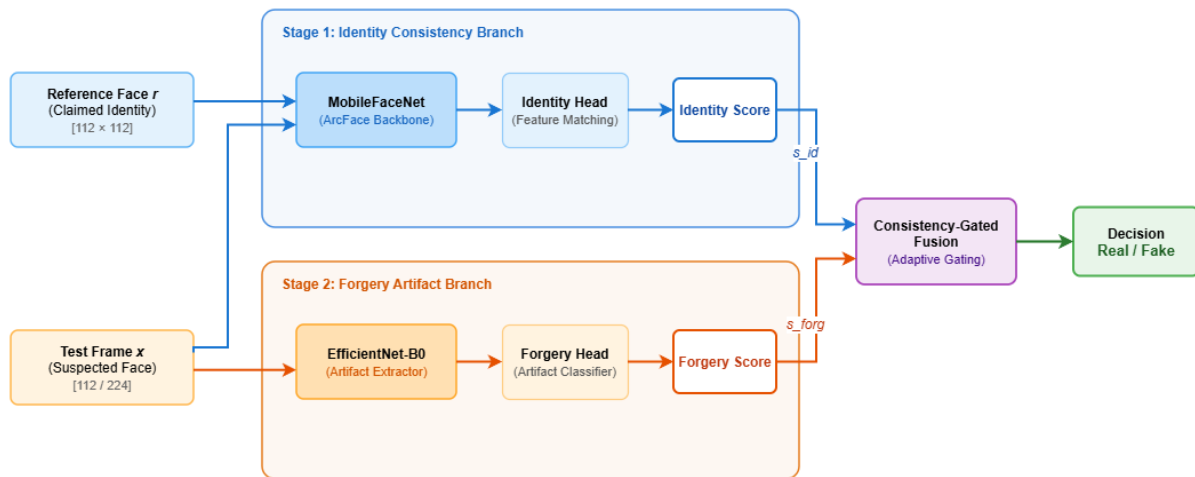


Figure 1. The Proposed Identity-Referenced Two-Stage Architecture.

Given a test face and a reference face of the claimed identity, the frozen MobileFaceNet encoder produces an identity-consistency score, while the trained EfficientNet-B0 encoder produces a forgery score. The two scores are then combined by a consistency gate: when the branches agree, the fusion reduces to their plain average, and when they disagree beyond the validation-defined threshold τ , the forgery score is progressively suppressed and the decision falls back to the identity baseline.

Reference Identity Construction

The claimed identity of a manipulated video is read from its filename: in a FaceForensics++ video such as FaceSwap_000_003, the recording presents itself as subject 000 (its original subject), so 000 is the claimed identity, whereas 003 is the source face swapped in. The reference image r is a single real face frame randomly sampled from the real video of the claimed identity 000. The reference is never taken from the manipulated test video itself; it comes from the original (real) video, so the identity branch compares the test face against an independent real face of the claimed identity rather than against any content of the fake video. This mirrors the deployment assumption, where a reference photo of the claimed identity is available from an external source (for example an ID photo), and the original real frame serves as its proxy. The same rule applies to training, validation, test, and cross-dataset (Celeb-DF) references. Because the face crops contain only the face region, the comparison is driven by identity content rather than by background or scene cues shared with the source video.

Identity Consistency Stage

The identity stage is built on MobileFaceNet (Chen et al., 2018), a lightweight face-recognition network pre-trained with ArcFace (Deng et al., 2019) on large-scale face data. The encoder is frozen during training so that its identity-invariant representations are preserved. Given the reference r and the test face x , the encoder produces two embeddings, from which a two-layer identity head ($1537 \rightarrow 256 \rightarrow 2$, with batch normalization, ReLU, and dropout) maps the concatenation of the two embeddings, their cosine similarity, and their element-wise difference to an identity-consistency score; the head is trained with cross-entropy on real (label 0) and fake (label 1) pairs, while the encoder remains frozen. A high identity score indicates that the test face is unlikely to belong to the claimed identity (a likely fake), while a low score indicates a likely match. The identity and forgery heads share the same label convention (0 = real, 1 = fake), so both scores are forgery probabilities aligned in direction before fusion.

Forgery Artifact Stage

The forgery stage uses EfficientNet-B0 (Tan & Le, 2019) as the backbone. In contrast to the identity stage, the forgery encoder is trained end-to-end to capture low-level artifacts such as blending boundaries, texture inconsistencies, and lighting mismatches. A forgery head maps the backbone features to a forgery score. This branch is strong in-domain but, as our experiments show, degrades substantially on unseen manipulation types because the artifacts it learns are generator specific.

Consistency-Gated Fusion

A naive fusion, such as feature concatenation or a learned linear combination, learns to rely on the forgery branch during training and therefore inherits its failure on unseen distributions. To avoid this, we fuse the two scores with a consistency gate, which is defined by Eq. (1) and Eq. (2):

$$s_{\text{fused}} = s_{\text{id}} + (s_{\text{forg}} - s_{\text{id}}) \cdot g \quad (1)$$

$$g = 0.5 \cdot \exp(-k \cdot \max(0, |s_{\text{forg}} - s_{\text{id}}| - \tau)) \quad (2)$$

Here Eq. (1) interpolates the two scores and Eq. (2) sets the interpolation weight. In Eq. (2), $k = 3$ is a positive constant controlling the gate sharpness, and the threshold τ is set to the 95th percentile (p95) of the branch disagreement $|s_{\text{forg}} - s_{\text{id}}|$ on the validation set. When the two branches agree within τ , the gate is $g = 0.5$ and the fusion reduces exactly to the plain average; as the disagreement grows beyond τ , g decays toward 0 and the fusion falls back to the identity baseline. Because the gate depends only on the two scores and a validation-set threshold, it requires no target-test calibration, so the same fusion rule can be applied consistently across unseen test distributions.

In Eq. (1) the weight is $w = g$ in $[0, 0.5]$, so the gate acts as a single, global fusion rule: it averages the two branches when they agree and falls back to the identity branch when they conflict. The plain average is the special case $w = 1/2$ obtained whenever the disagreement is within τ . This adaptive suppression is decisive on FaceSwap, where the forgery branch collapses (43.4%): the gate keeps the fused score at 67.7%, whereas the plain average is dragged down to 61.0% and the learnable concat to 50.4%. Because g is monotonic in the

disagreement, a larger k suppresses the forgery branch more aggressively once the disagreement exceeds τ , and the fused score interpolates smoothly between the identity baseline and the plain average.

Experiments

Datasets and Implementation Details

We evaluate on FaceForensics++ (FF++, C23 light compression) (Rössler et al., 2019) and Celeb-DF v2 (Y. Li et al., 2020). FF++ contains 1000 original videos and their manipulated versions generated by four methods: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Celeb-DF contains 590 real and 5639 fake videos of 59 celebrities. Faces are cropped and resized to 112x112 for the identity encoder and 224x224 for the forgery encoder. For each test face, the reference image r is a single real face frame randomly sampled from the real video of the claimed identity (extracted from the manipulated video's filename); the reference and the test face are cropped and aligned with the same face detector, and each identity contributes up to five reference - test pairs, with the reference frame sampled independently in each pair. The FF++ identities are split into training and test sets at an 80/20 ratio with no identity overlap between them, and 20% of the training identities are held out for validation (64% train / 16% validation / 20% test). Following prior work, we report the video-level AUC by averaging frame-level predictions within each video. The identity encoder (MobileFaceNet) is frozen, and the forgery encoder (EfficientNet-B0) is trained for 10 epochs with the Adam optimizer. The model has 6.0M parameters in total (1.2M frozen MobileFaceNet, 4.3M trained EfficientNet-B0, and 0.4M trainable heads). All experiments use three random seeds (42, 123, 2024), which also fixes the random reference sampling for reproducibility, reporting the mean over seeds.

In-domain and Cross-Manipulation Results

Table 2 reports the in-domain performance (training and testing on the same four manipulation types with disjoint identities) and the cross-manipulation generalization under the leave-one-manipulation-out protocol (training on three types and testing on the held-out type). In-domain, the forgery branch achieves 96.5% AUC, and the fusion improves the identity baseline of 82.8% to 95.6%. The most striking result is on FaceSwap (FS): the forgery branch collapses to 43.4% AUC, close to random, whereas the identity branch remains at 77.8% AUC. This confirms that identity consistency is robust to identity-swapping manipulations, whereas forgery artifacts are generator specific. The fusion tracks the identity baseline on FS, reaching 67.7%. Conversely, on NeuralTextures (NT), which preserves identity and alters only texture, the identity branch is uninformative (54.8%, near chance) and its fluctuations slightly perturb the fusion below the forgery-only score (70.7% to 64.1%). Table 2 also serves as a two-way ablation: removing the identity stage reduces the method to the forgery-only detector, which collapses on FaceSwap (43.4%); removing the forgery stage reduces it to the identity-only detector, which is stable across manipulations (77.8% on FaceSwap) but weaker in-domain (82.8%). The full framework inherits the forgery branch's in-domain strength and the identity branch's cross-manipulation stability.

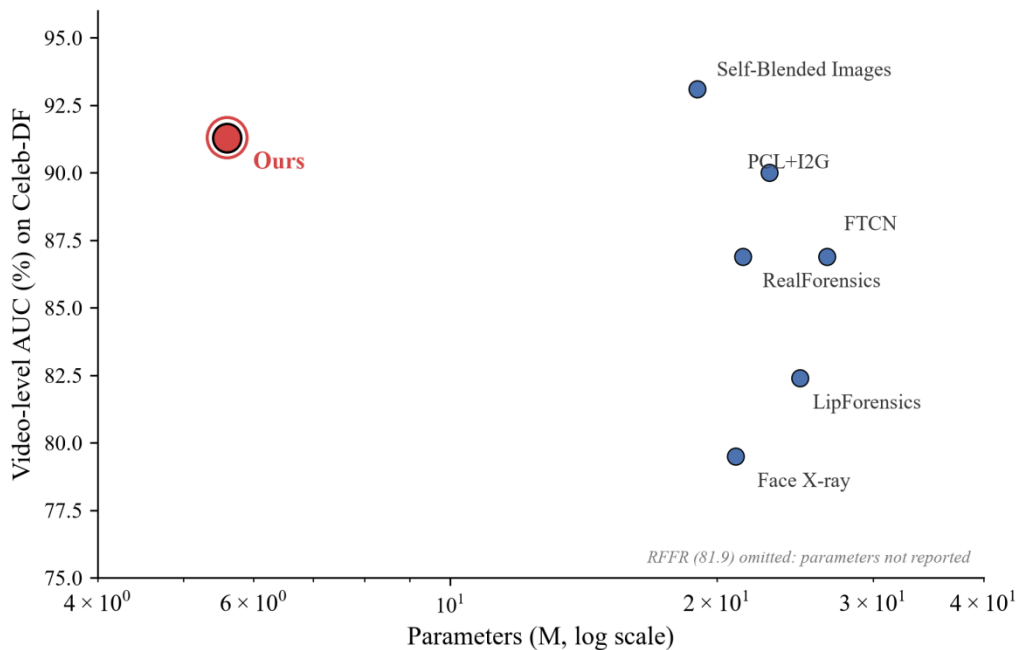
Table 2. In-Domain And Cross-Manipulation Results (Video-Level AUC %).

Scenario	Forgery	Identity	Fusion
Full (in-domain)	96.5	82.8	95.6
DF (leave-out)	88.1	86.9	92.2
F2F (leave-out)	81.3	69.4	81.8
FS (leave-out)	43.4	77.8	67.7
NT (leave-out)	70.7	54.8	64.1

Cross-Dataset Results

For cross-dataset evaluation, we train on FF++ and test on Celeb-DF v2. The identity branch achieves 89.3% video-level AUC, substantially higher than the forgery branch (79.0%), demonstrating that identity consistency transfers across datasets with different generators, subjects, and compression settings. The fusion reaches 91.3% video-level AUC, outperforming both individual branches.

Figure 2 compares our method with existing detectors in terms of accuracy and parameter count on Celeb-DF. Our method attains the second-highest AUC (91.3%) with only 6.0M parameters, roughly 3.2 to 4.4 times fewer than the baselines.

**Figure 2. Efficiency comparison on Celeb-DF (video-level AUC %).**

Comparison with Existing Methods

Table 3 compares our method with existing detectors on cross-manipulation generalization, including the classical detectors Xception (Chollet, 2017), CNN-GRU (Sabir et al., 2019), and a patch-based classifier (Chai et al., 2020), as well as the recent unsupervised inconsistency-aware detector UIA-ViT (Zhuang et al., 2022). The baseline values are taken from published results under the same leave-one-manipulation-out protocol; minor differences in

implementation or preprocessing may therefore exist. In the hardest case, FaceSwap, the classical detectors fail with 47.6% to 60.5% AUC, and our method attains 67.7%, outperforming them by 7 to 20 percentage points. Only the unsupervised UIA-ViT is higher on this manipulation (70.7%), and it requires about fourteen times more parameters (86M against 6.0M); relative to its parameter budget, our fusion therefore gives the best cross-manipulation accuracy reported in Table 3. This demonstrates the advantage of the identity reference over purely artifact-based detection on unseen identity-swapping manipulations. On NeuralTextures, a manipulation that preserves identity while altering expression and texture, our identity reference provides little signal (54.8%) and the fusion (64.1%) underperforms the artifact-based baselines (79.7% to 86.6%); on Face2Face, our fusion (81.8%) is likewise below the strongest artifact-based methods (up to 94.2%). These are expected limitations of identity-based detection for identity-preserving manipulations, and they are discussed further in the Conclusion.

Table 3. Cross-Manipulation Comparison on Faceforensics++ (Video-Level AUC %). Parameter Counts (M) Are Backbone Estimates from The Original Papers.

Method	DF	F2F	FS	NT	Params (M)	Avg	Params Ours (×)
Xception	93.9	86.8	51.2	79.7	~22.9	77.9	3.8
CNN-GRU	97.6	85.8	47.6	86.6	~23.9	79.4	4.0
Patch-based	94.0	87.3	60.5	84.8	~25.6	81.7	4.3
UIA-ViT	96.7	94.2	70.7	82.8	~86	86.1	14.3
Ours	92.2	81.8	67.7	64.1	6.0	76.5	1.0

Table 4 summarizes the cross-dataset comparison on Celeb-DF. Our method attains 91.3% video-level AUC, higher than Face X-ray (L. Li et al., 2020), RFFR (Shi et al., 2025), LipForensics (Haliassos et al., 2021), RealForensics (Haliassos et al., 2022), FTCN (Zheng et al., 2021), and PCL+I2G (Zhao et al., 2021), and slightly below the self-supervised method Self-Blended Images (Shiohara & Yamasaki, 2022), which reports 93.1%. Our 91.3% is obtained with 6.0M parameters, compared with 19 to 27M for the temporal and self-supervised baselines in Table 4, confirming that the identity reference provides a lightweight route to cross-dataset generalization.

Table 4. Cross-Dataset Comparison on Celeb-DF (Video-Level AUC %). Parameter Counts (M) Are Backbone Estimates from The Original Papers; A Dash Indicates Not Reported.

Method	Celeb-DF	Params (M)	Params Ours (×)	ΔAUC (pp)
Face X-ray	79.5	~21	3.5	+11.8
RFFR	81.9	—	—	+9.4
LipForensics	82.4	24.8	4.1	+8.9
RealForensics	86.9	21.4	3.6	+4.4

FTCN	86.9	26.6	4.4	+4.4
PCL+I2G	90.0	~22.9	3.8	+1.3
Self-Blended Images (SBI)	93.1	~19	3.2	-1.8
Ours	91.3	6.0	1.0	—

Ablation Study

Table 5 ablates the fusion strategy. The learnable concat fusion attains the highest in-domain score (97.0% on Full) but collapses on FaceSwap (50.4%), because it continues to trust the forgery branch, which inverts on this unseen manipulation. The plain average is more robust (61.0% on FaceSwap) but is still dragged down by the inverted forgery branch. The threshold gate, which requires no target-test calibration, keeps the fused score at 67.7% on FaceSwap, close to the identity baseline (77.8%). This confirms that the main benefit of the consistency gate is protecting the identity signal when the forgery branch becomes unreliable, rather than uniformly outperforming plain averaging across all manipulation types. The threshold τ is set to the 95th percentile (p95) of the branch disagreement on the validation set, so that the fusion uses the plain-average behavior for approximately 95% of the validation samples and falls back only for the most extreme disagreements; the choice of $k = 3$ is not critical: sweeping k over $[0.5, 20]$ shows that the fused AUC is stable for $k \geq 2$, so we use $k = 3$. The threshold τ is determined exclusively from the validation set and is fixed before test evaluation, shared across all manipulation types and datasets; we do not tune τ separately for individual test sets, and the value of k is likewise fixed at 3 for all experiments.

Table 5. Ablation of the fusion strategy (video-level AUC %); p95 denotes the gate threshold set to the 95th percentile of the validation disagreement.

Fusion	DF	F2F	FS	NT	Full
concat	90.6	83.2	50.4	70.8	97.0
avg (0.5/0.5)	92.1	83.1	61.0	66.0	96.3
Consistency-gated (Ours)	92.2	81.8	67.7	64.1	95.6

Conclusion

This paper presents an identity-referenced two-stage framework for deepfake detection that exploits identity consistency as a robust prior for identity-swapping manipulations. A frozen MobileFaceNet encoder provides a stable identity score, and a trained EfficientNet-B0 encoder captures forgery artifacts; the two are combined through a consistency-gated fusion that requires no target-test calibration, with its threshold determined once on the validation set and fixed for every test setting. It should be emphasized that the contribution of this work is not the use of an identity reference as such, which has been explored before (Shen et al., 2022), but the way in which the identity signal is used: it is kept frozen, converted into a standalone detection score, and combined with an independent artifact branch by a gate that is defined once on the validation set and shared across all test settings, so that the identity signal protects the final decision whenever the artifact branch becomes unreliable. Experiments on FaceForensics++ and Celeb-DF showed that the identity branch remains at 77.8% AUC on the most difficult cross-manipulation case where the forgery branch collapses to 43.4%, and that

the fusion attains 91.3% cross-dataset AUC under this same validation-defined fusion rule, improving over classical detectors and approaching strong self-supervised methods while remaining comparable to the identity branch alone on the manipulations on which the artifact branch fails. A limitation is that identity-preserving manipulations, such as NeuralTextures and, to a lesser extent, Face2Face, remain challenging for identity reference, since they alter expression and texture rather than the identity itself. Future work will extend the framework to temporal and multi-modal cues, evaluate on additional benchmarks such as DFDC (Dolhansky et al., 2020), and pursue tighter face alignment to improve the identity signal further.

Acknowledgements: The authors would like to express their sincere gratitude to Kuala Lumpur University of Science and Technology (KLUST) for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.

Funding Statement: No Funding

Conflict of Interest Statement: The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethics Statement: This study relies exclusively on secondary, publicly available benchmark datasets (FaceForensics++ and Celeb-DF v2). No human participants were directly recruited, and no biological experiments were conducted, thereby requiring no additional institutional ethical approval. All data were obtained and utilized strictly in accordance with their respective research use agreements and licensing terms (FaceForensics++ under its academic license terms and Celeb-DF under its research dataset agreement). The authors confirm that this work adheres fully to recognized academic integrity and ethical publishing standards.

Author Contribution Statement: All authors contributed significantly to the development of this manuscript. Zhen Yin was responsible for the conceptualization, methodology, experiment design and execution, data collection, analysis, and drafting of the manuscript. Robiatul A'dawiah Jamaluddin contributed to the overall supervision, critical revision of the manuscript, and provided guidance on the research direction.

References

- Chai, L., Bau, D., Lim, S.-N., & Isola, P. (2020). What makes fake images detectable? Understanding properties that generalize. In *Computer Vision—ECCV 2020* (pp. 103–120). Springer. https://doi.org/10.1007/978-3-030-58574-7_7
- Chen, S., Liu, Y., Gao, X., & Han, Z. (2018). MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Biometric Recognition* (pp. 428–438). Springer. https://doi.org/10.1007/978-3-319-97909-0_46
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1800–1807). IEEE. <https://doi.org/10.1109/CVPR.2017.195>
- Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4685–4694). IEEE. <https://doi.org/10.1109/CVPR.2019.00482>
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) dataset. arXiv. <https://arxiv.org/abs/2006.07397>
- Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14930–14942). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01453>
- Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5037–5047). IEEE. <https://doi.org/10.1109/CVPR46437.2021.00500>
- Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5000–5009). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00505>
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3204–3213). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00327>
- Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision—ECCV 2020* (pp. 86–103). Springer. https://doi.org/10.1007/978-3-030-58610-2_6
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1–11). IEEE. <https://doi.org/10.1109/ICCV.2019.00009>
- Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 80–87). IEEE. <https://doi.org/10.48550/arXiv.1905.00582>
- Shen, D., Zhao, Y., & Quan, C. (2022). Identity-referenced deepfake detection with contrastive learning. In *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security* (pp. 27–32). ACM. <https://doi.org/10.1145/3531536.3532964>

- Shi, L., Zhang, J., Ji, Z., Bai, J., & Shan, S. (2025). Real face foundation representation learning for generalized deepfake detection. *Pattern Recognition*, 161, 111299. <https://doi.org/10.1016/j.patcog.2024.111299>
- Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18699–18708). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01816>
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR. <https://doi.org/10.48550/arXiv.1905.11946>
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8692–8701). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00872>
- Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15003–15013). IEEE. <https://doi.org/10.1109/ICCV48922.2021.01475>
- Zheng, Y., Bao, J., Chen, D., Zeng, M., & Wen, F. (2021). Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15024–15034). IEEE. <https://doi.org/10.1109/ICCV48922.2021.01477>
- Zhuang, W., Chu, Q., Tan, Z., Liu, Q., Yuan, H., Miao, C., Luo, Z., & Yu, N. (2022). UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *Computer Vision—ECCV 2022* (pp. 391–407). Springer. https://doi.org/10.1007/978-3-031-20065-6_23