

Augmentation and Cutout in Deepfake Detection: A Comparative Study of Accuracy, Calibration, and Attention

1st Mert Kaya
Dept. of Computer Engineering
TED University
Ankara, Turkey
mertkayacs@gmail.com

2nd Venera Adanova
Dept. of Computer Engineering
TED University
Ankara, Turkey
venera.adanova@tedu.edu.tr

Abstract—Deepfake generation has advanced faster than the tools meant to contain it. Proactive defenses such as invisible watermarking have already been broken, which leaves forensic detection as the practical line of defense. Detection alone is not enough, however, when a verdict must be trusted in high-stakes settings, where a result that cannot be explained is of little use. This work studies how preprocessing choices shape both the accuracy and the explainability of convolutional deepfake detectors. We build a parametric pipeline around an EfficientNet-B4 backbone and compare nine preprocessing configurations that vary data augmentation and cutout strategy on a paired subset of the FaceForensics++ dataset. Each configuration is evaluated with four complementary metrics that separate ranking ability from probability calibration, and the decisions of every model are examined region by region with Grad-CAM aligned to facial landmarks. Combining augmentation with a black-filled cutout produces the strongest overall detection, reaching an area under the curve of 0.8971. The four metrics do not agree on a single best configuration, and the per-region analysis exposes a vulnerability in the nose region that aggregate accuracy hides. These findings show that preprocessing is a design decision affecting both how well a detector ranks samples and how trustworthy and interpretable its decisions are.

Keywords—deepfake detection, explainable artificial intelligence, Grad-CAM, data augmentation, convolutional neural networks, model calibration, FaceForensics++

I. INTRODUCTION

Synthetic faces produced by modern generative models have become so realistic that people can no longer reliably tell them apart from photographs of real individuals; in controlled studies such faces are even judged as more trustworthy than genuine ones [1]. The techniques that create this content have moved quickly, from the generative adversarial networks behind early face swaps to diffusion models that synthesize photorealistic imagery from a text prompt [2]. As the quality of manipulation rises, so does its potential for fraud, disinformation, and the erosion of trust in recorded media [3], [4]. These harms are concrete: in 2024 a single deepfake video call defrauded the engineering firm Arup of about USD 25 million [5], and generative-AI-enabled fraud losses in the United States are projected to reach USD 40 billion by 2027 [6].

One response has been to mark synthetic media at the

moment of creation. Provenance schemes embed an invisible watermark that downstream tools can later read to confirm an image was machine generated, and recent systems such as SynthID aim to do this at internet scale [7]. Such defenses are proactive, but they are not unbreakable. A universal attack, UnMarker, strips the strongest watermarks, including semantic ones, while leaving the image visually intact [8]. If the marks meant to certify provenance can be removed, then reactive detection of manipulated faces remains a practical necessity rather than a fallback.

Detection is rarely the end of the story. When a judgment about whether a face is real or fake informs a consequential decision, the people who rely on it need to understand why the model decided as it did. A classifier that outputs only a probability is hard to scrutinize or challenge [9]. This raises two requirements that are usually treated separately: a detector must be accurate, and its decisions must be interpretable and well calibrated, so that a reported confidence can be trusted [10].

Most work on convolutional deepfake detection treats the training recipe as fixed and adds explainability, when present, as a visualization placed on top of an already trained model. The preprocessing choices that shape the detector, in particular the data augmentation policy and the use of cutout regularization, are tuned for accuracy and then left unexamined from the standpoint of what the model learns to look at. Whether those same choices also change where a detector attends, and how trustworthy its probabilities are, has received little attention.

This paper studies that question directly. Building on Seferbekov’s detection pipeline [11], we construct a parametric framework around an EfficientNet-B4 backbone [12] and vary two preprocessing axes, augmentation and cutout fill strategy, across nine configurations. Each configuration is evaluated on a paired subset of FaceForensics++ [13] with four metrics that separate ranking ability from probability calibration, and the decisions of every trained model are examined region by region using Grad-CAM [14] registered to 68 facial landmarks [15]. The contributions are as follows.

- A controlled comparison of nine augmentation and

cutout configurations for convolutional deepfake detection, evaluated across AUC, F1-score, Brier score, and logarithmic loss on identity-preserving real-fake pairs.

- Evidence that these four metrics disagree on the best configuration, so that no single accuracy number characterizes a detector: the configuration with the strongest ranking ability is not the best calibrated.
- A per-region Grad-CAM analysis showing that augmentation and cutout choices reshape where a detector attends, and that the highest-scoring configuration hides a pronounced failure in the nose region that accuracy metrics do not reveal.

II. RELATED WORK

A. Deepfake Generation and Detection

Early face manipulations relied on autoencoders and generative adversarial networks to swap or reenact identities, and surveys of this period catalogue a fast-growing family of methods alongside the detectors built to counter them [3], [16]. Generation has since shifted toward diffusion models that produce photorealistic faces from text or reference images [2], [17], widening the gap between what can be synthesized and what can be reliably detected [4].

B. Detection Approaches

Detectors fall into a few broad groups. Frame-level methods treat detection as image classification with a convolutional backbone, where efficient architectures such as EfficientNet offer a strong accuracy-to-cost ratio [12]. Temporal methods add a recurrent stage to model inconsistencies across frames [18]. A third group targets the manipulation process rather than a specific generator: Face X-ray detects the blending boundary left when a synthetic face is composited into a real frame, which helps it generalize to unseen manipulations [19]. A related line improves generalization through synthetic-artifact augmentation, such as frequency-enhanced self-blended images [20]. The winning solution of the DeepFake Detection Challenge combined MTCNN face detection with heavy augmentation and an ensemble of EfficientNet backbones [11], [21], and it serves as the methodological basis for our pipeline.

Generalization across datasets remains the central difficulty. Recent work shows that careful training is needed for a detector to transfer beyond its training data [22], and detectors trained on academic datasets degrade sharply on in-the-wild deepfakes [23]. Benchmarks such as FaceForensics++ [13] and Celeb-DF [24] are used to measure this.

C. Artifacts, Explainability, and Calibration

A separate line of work asks what detectors actually key on. Visual-artifact studies show that manipulations leave traces in specific facial regions [25], and gradient-based attribution methods such as Grad-CAM expose where a network attends when it makes a decision [14]. Such tools underpin explainable artificial intelligence for vision [9], yet in deepfake detection they are typically applied after training

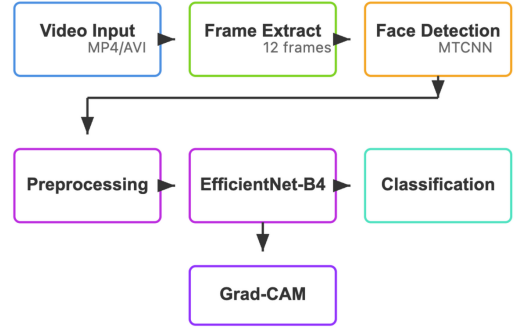


Fig. 1: Detection pipeline. A video is reduced to 12 frames, faces are detected with MTCNN and preprocessed, an EfficientNet-B4 backbone scores each frame, and Grad-CAM provides region-level explanations of the trained model.

to illustrate individual examples rather than to compare how training choices change a model’s attention; only recently has explanation quality for deepfake detectors been evaluated quantitatively [26]. Accuracy is also not the same as trustworthiness: a model can rank samples well while being poorly calibrated, reporting confidences that do not match its true correctness [10], and calibration is increasingly treated as a reliability requirement for deepfake detectors [27]. Our study connects these threads by treating augmentation and cutout as variables whose effect on calibration and region-level attention we measure directly.

III. EXPERIMENTS

A. Pipeline Overview

Fig. 1 summarizes the detection pipeline, which adapts Seferbekov’s DeepFake Detection Challenge solution into a flexible framework so that individual preprocessing factors can be varied without altering the rest of the system [11]. Each input video is reduced to a short frame sequence, faces are detected and aligned, optional augmentation and cutout are applied, and an EfficientNet-B4 backbone produces a frame-level score that is averaged into a video-level prediction. Grad-CAM is attached to the trained model to recover region-level explanations.

B. Preprocessing

From each video, 12 frames are sampled at equal intervals from 32 candidates to form a consistent temporal sequence. Faces are located with a multitask cascaded convolutional network (MTCNN) [28] and cropped to a fixed 224 by 224 resolution. For the region-level analysis, 68 facial landmarks are extracted with dlib [15] and used to associate image areas with the eyes, eyebrows, nose, mouth, and jaw. Frames are normalized with ImageNet statistics before being passed to the backbone.

C. Augmentation and Cutout

Data augmentation uses the Albumentations library [29] and applies flipping, scaling, rotation, noise, blur, and brightness changes to duplicated copies of the training frames. Cutout regularization [30] follows the FaceCrop strategy of the



Fig. 2: The three cutout fill variants applied to a manipulated frame: random RGB (left), black (center), and white (right).

TABLE I: The Nine Preprocessing Configurations

#	Configuration	Augmentation	Cutout fill
1	Baseline	none	none
2	Augmentation only	standard	none
3	Augmentation only	intense	none
4	Cutout only	none	random
5	Cutout only	none	black
6	Cutout only	none	white
7	Augmentation + cutout	standard	random
8	Augmentation + cutout	standard	black
9	Augmentation + cutout	standard	white

baseline [11]: the structural similarity index (SSIM) [31] between a real frame and its manipulated counterpart locates regions of high similarity, and polygonal cutouts are placed there on the synthetic frames. The masked area is filled in one of three ways, with random RGB values, with black, or with white (Fig. 2).

A complementary star-shaped cutout of radius 8 to 16 pixels is applied to real frames with probability 0.5, discouraging the model from overfitting to pristine facial detail. This asymmetric design, FaceCrop on fakes and star cutout on reals, exposes the detector to detection-resistant artifacts without erasing genuine cues.

D. Model and Training

The backbone is an EfficientNet-B4 [12] initialized with ImageNet weights and applied independently to each of the 12 frames in a time-distributed configuration. Frame features are reduced by global average pooling and passed to a compact head consisting of dropout with rate 0.55 and a single sigmoid unit with L2 kernel regularization of 0.001. Video-level predictions are obtained by averaging the frame-wise probabilities, and the model is trained with binary cross-entropy on this averaged output to match the inference rule. Optimization uses Adam with a cosine learning-rate schedule, gradient clipping at 1.0, a batch size of 4, and early stopping on validation loss with a patience of 5 epochs. Every configuration is trained three times, and we report the mean and standard deviation across runs.

E. Experimental Configurations

Nine configurations span two axes, the presence and intensity of augmentation and the cutout fill strategy, as listed in Table I. A baseline uses neither augmentation nor cutout. Two configurations apply augmentation alone at standard and increased intensity. Three apply cutout alone with random, black, or white fill. The final three combine augmentation with each fill type. All other architectural and training choices are held at the baseline values, so any difference in outcome can be attributed to the two factors under study.

F. Dataset

Experiments use a subset of FaceForensics++ [13] consisting of 1,000 real and 1,000 manipulated videos drawn from a pool of roughly 5,000. The manipulated set is balanced across four families, FaceSwap, Face2Face, FaceShifter, and DeepFakes, with 250 pairs each. Real and fake videos are kept as identity-preserving pairs, and the data is divided 70/15/15 into training, validation, and test partitions on a pair basis, so an identity never appears across splits. Larger datasets such as Celeb-DF [24] and the DeepFake Detection Challenge set [21] were left to future work because of their size and cost.

G. Evaluation Metrics

Four complementary metrics separate ranking ability from probability quality. The area under the receiver operating characteristic curve (AUC) and the F1-score measure how well a model separates the two classes, threshold-independently and at the operating threshold; the Brier score and logarithmic loss measure how well its reported probabilities are calibrated [10]. Reporting both families matters because a model can rank samples well while being poorly calibrated.

H. Grad-CAM Region Analysis

To study what each model attends to, Grad-CAM [14] is applied per frame and the maps are summarized at the video level. Activations are registered to the 68 landmarks and averaged within eight facial regions: the left and right eyes, the left and right eyebrows, the nose, the inner and outer mouth, and the jaw. For each region we record its mean activation, computed separately for the four prediction categories: true positive, true negative, false positive, and false negative. Comparing these region scores across categories links a model’s attention to its specific successes and failures.

IV. RESULTS

A. Quantitative Performance

Table II reports all nine configurations across the four metrics. Combining augmentation with a black-filled cutout gives the strongest aggregate result, with an AUC of 0.8971, an F1-score of 0.8429, and the best Brier score of 0.1242. This is about a 3% AUC improvement over the baseline (0.8678) and about an 8% gain in F1-score (0.7780). Every configuration that combines augmentation with cutout improves on the baseline Brier score, by between 4.5% and 18.5%, indicating better-calibrated probabilities once both factors are present.

B. Metrics Disagree on the Best Configuration

No single configuration wins on every metric. While the black-filled combination leads on AUC, F1-score, and Brier score, the lowest logarithmic loss belongs to the random-filled combination (0.4656 against 0.4710), so the best-ranked model is not the best calibrated. Augmentation alone is the clearest case: it ranks last on AUC (0.8610) and worst

TABLE II: Detection performance of the nine preprocessing configurations on the FaceForensics++ subset, reported as mean $\pm$ standard deviation over three runs. Arrows indicate whether higher ($\uparrow$) or lower ($\downarrow$) is better. The best value in each column is shown in bold.

Configuration	AUC $\uparrow$	F1-score $\uparrow$	Brier $\downarrow$	LogLoss $\downarrow$
Baseline	0.8678 $\pm$ 0.0044	0.7780 $\pm$ 0.0065	0.1524 $\pm$ 0.0013	0.4827 $\pm$ 0.0069
Augmentation only (standard)	0.8610 $\pm$ 0.0051	0.7955 $\pm$ 0.0084	0.1572 $\pm$ 0.0010	0.5719 $\pm$ 0.0084
Augmentation only (intense)	0.8718 $\pm$ 0.0043	0.7932 $\pm$ 0.0077	0.1431 $\pm$ 0.0008	0.5288 $\pm$ 0.0081
Cutout only, random fill	0.8711 $\pm$ 0.0051	0.7769 $\pm$ 0.0069	0.1524 $\pm$ 0.0012	0.5241 $\pm$ 0.0078
Cutout only, black fill	0.8666 $\pm$ 0.0043	0.7912 $\pm$ 0.0077	0.1537 $\pm$ 0.0009	0.4989 $\pm$ 0.0074
Cutout only, white fill	0.8639 $\pm$ 0.0047	0.7704 $\pm$ 0.0074	0.1462 $\pm$ 0.0011	0.5244 $\pm$ 0.0076
Random fill + augmentation	0.8837 $\pm$ 0.0072	0.7950 $\pm$ 0.0098	0.1450 $\pm$ 0.0011	0.4656 $\pm$ 0.0067
Black fill + augmentation	0.8971 $\pm$ 0.0064	0.8429 $\pm$ 0.0101	0.1242 $\pm$ 0.0009	0.4710 $\pm$ 0.0063
White fill + augmentation	0.8734 $\pm$ 0.0061	0.7951 $\pm$ 0.0095	0.1455 $\pm$ 0.0012	0.4761 $\pm$ 0.0071

on logarithmic loss (0.5719), yet reaches the second-best F1-score (0.7955). Augmentation without cutout therefore improves the precision-recall balance at the chosen threshold while degrading both the probability ranking and the calibration. Because the threshold-independent metric (AUC), the threshold-dependent metric (F1-score), and the calibration metrics (Brier score and logarithmic loss) order the configurations differently, reporting any one of them in isolation gives an incomplete picture of a detector.

C. Correct and Incorrect Decisions Rely on Different Regions

Grad-CAM links each decision to the facial regions that drove it. Fig. 3 reports the mean per-region activation of all nine configurations across the four prediction categories; the description below focuses on the black-filled combination, the strongest model, with the baseline for contrast.

In true-positive cases, where the model correctly flags manipulated faces, the nose and the left eye stand out, both close to 73%, and the left eyebrow is also active. Black fill and augmentation give the model the ability to localize these discriminative cues, which the baseline lacks: it remains weak across every region (28 to 48%) and relies on much weaker evidence to distinguish real from fake. Several augmented configurations rely instead on the mouth, with inner and outer mouth activation around 62 to 65%, a region that carries manipulation artifacts of its own.

In true-negative cases, where the model accepts genuine faces as real, the left eye again leads, above 70% against about 35% for the baseline, and augmentation also raises the eyebrow regions that the baseline keeps low. A genuine face has no manipulation trace to detect, so the model confirms it from the eye and eyebrow, while the baseline depends on fewer regions. The black-filled cutout without augmentation produces the highest nose activation in this case (74%) and keeps the mouth high as well, where the augmented models are more selective.

In false-negative cases, where the model misses manipulated faces, the nose carries the highest activation, rising from 49% at the baseline to 78% for the black-filled combination, with the inner mouth close behind at 75%, while the eyebrows show the lowest values. The heavy black and white fill pulls attention to the nose and mouth, so the model reads a subtle manipulation there as genuine and the fake goes undetected; simpler configurations keep a better balance and miss fewer fakes.

In false-positive cases, where the model wrongly flags genuine faces, the nose and the inner mouth show the highest activation. The configurations without augmentation are the worst, with nose activation up to 77%: they read ordinary variation in the nose and mouth as manipulation, which raises false alarms on real faces, whereas the black-filled combination is more conservative and stays lower at about 69%. The same black fill that is restrained here reaches 78% in the false-negative case, so its lower false-alarm rate comes with more missed manipulations.

D. A Hidden Trade-off in the Nose Region

One pattern cuts across all four categories: the nose carries high activation whether the model is right or wrong, so the region that anchors correct detection also dominates the errors. A configuration can therefore improve its overall score while growing more dependent on a single region, a trade-off the headline metrics in Table II do not reveal. This reliance is understandable, since the nose is a structural anchor whose subtle blending artifacts are reliably exploited by detectors [25], and stable mid-facial regions reveal manipulation traces more consistently than dynamic areas such as the eyes or mouth [19]. We also observe that the black-filled cutout outperforms random fill on this FaceForensics++ subset, which runs counter to the larger-scale DeepFake Detection Challenge experience [11]. We attribute this to the smaller scale of our paired subset rather than to a property of the fill, and we treat it as a conjecture to be tested at full scale.

V. CONCLUSION

We studied how two preprocessing choices, data augmentation and cutout fill strategy, affect both the accuracy and the explainability of a convolutional deepfake detector. Across nine configurations on a paired FaceForensics++ subset, combining augmentation with a black-filled cutout produced the strongest overall detection, but no single metric captured the full picture: the best-ranked configuration was not the best calibrated, and augmentation without cutout improved threshold performance while harming calibration. A region-level Grad-CAM analysis showed that correct and incorrect decisions rely on different facial regions, and that the configuration with the best aggregate score exposes the nose as a persistent weak point in its error cases. Preprocessing is therefore a design decision in its own right: it trades discrimination against calibration and changes where a detector looks, with direct consequences for how far its decisions can

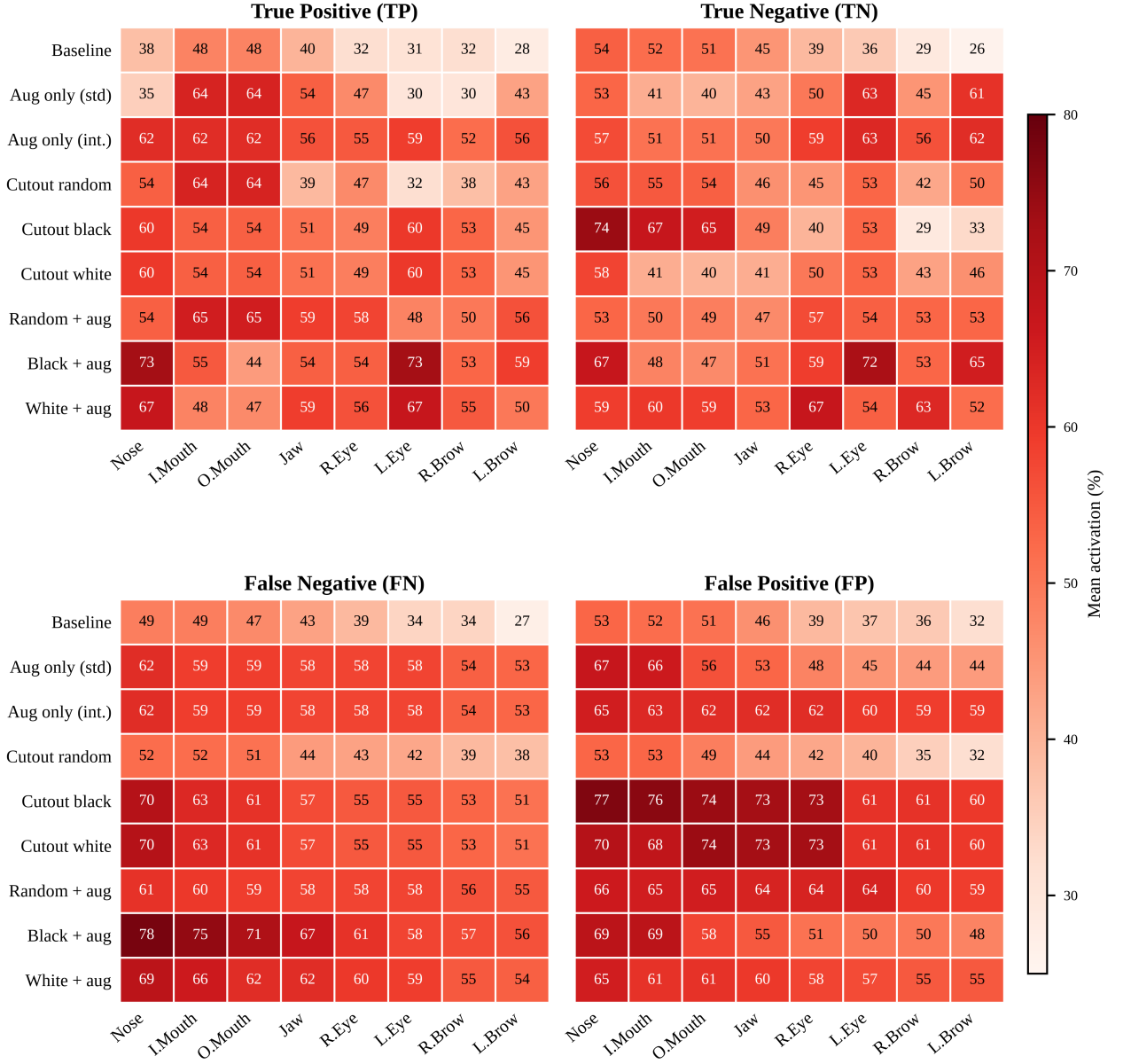


Fig. 3: Mean Grad-CAM activation (%) per facial region for the nine configurations across the four prediction categories: true positive and true negative (top row), false negative and false positive (bottom row). Rows follow the order of Table II and the color scale is shared across all panels. When correct, the black-filled combination is most active in the nose, left eye, and left eyebrow; its errors shift to the nose and mouth.

be trusted, not only how often they are correct.

These results come with clear limits. The study uses a single architecture, EfficientNet-B4, and a single dataset at reduced scale, without cross-dataset evaluation or diffusion-era manipulations, and the explainability analysis relies on Grad-CAM alone. Each of these is a natural direction for future work: evaluating the same configurations across datasets such as Celeb-DF and the DeepFake Detection Challenge set, extending to diffusion-generated faces, scaling to the full FaceForensics++ corpus, and comparing additional attribution methods and architectures. The broader point is that accuracy, calibration, and region-level explanation should be reported together, because for high-stakes use a detector must

be judged on all three at once, and a strong score on one does not offset a weak score on another.

REPRODUCIBILITY

The source code and the trained model weights for the reported configurations are made available by the authors. The release includes the preprocessing code, the nine configuration definitions, the training and evaluation scripts, and the Grad-CAM region analysis, so that both the metrics and the per-region figures can be regenerated. Two paths are supported: training each configuration from scratch, or loading the released weights to reproduce the reported scores directly. The FaceForensics++ subset used in this study is

governed by the dataset’s own license; access to parts of the data should be arranged with the dataset providers, while the exact subset split and preprocessing configuration can be requested from the authors.

USE OF AI TOOLS

The authors disclose that the AI assistant Claude (Anthropic) was used for grammar and language editing of the manuscript. It was not used for the research design, experiments, data analysis, or results, which are entirely the authors’ own. The authors reviewed and approved the final text and take full responsibility for its content.

REFERENCES

- [1] S. J. Nightingale and H. Farid, “AI-synthesized faces are indistinguishable from real faces and more trustworthy,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, p. e2120481119, 2022.
- [2] H. Chen, Q. Xiang, J. Hu, M. Ye, C. Yu, H. Cheng, and L. Zhang, “Comprehensive exploration of diffusion models in image generation: A survey,” *Artificial Intelligence Review*, vol. 58, 2025.
- [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, 2020.
- [4] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.
- [5] Fortune, “British engineering firm arup revealed as victim of a \$25 million deepfake video-call scam,” <https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo/>, 2024, accessed: 2026-06-26.
- [6] Deloitte Center for Financial Services, “Deepfake banking fraud risk on the rise,” <https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html>, 2024, accessed: 2026-06-26.
- [7] S. Goyal *et al.*, “SynthID-Image: Image watermarking at internet scale,” arXiv:2510.09263, 2025.
- [8] A. Kassis and U. Hengartner, “UnMarker: A universal attack on defensive image watermarking,” in *Proc. IEEE Symposium on Security and Privacy (S&P)*, 2025.
- [9] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.
- [10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.
- [11] S. Seferbekov, “Deepfake detection challenge (DFDC) winning solution,” https://github.com/selimsef/dfdc_deepfake_challenge, 2020, accessed: 2026-06-25.
- [12] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.
- [13] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.
- [14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.
- [15] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [16] L. Verdoliva, “Media forensics and deepfakes: An overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
- [17] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea *et al.*, “Deepfake media generation and detection in the generative AI era: A survey and outlook,” arXiv:2411.19537, 2024.
- [18] D. Guera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.
- [19] 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 *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5000–5009.
- [20] A. A. Hasanaath, H. Luqman, R. Katib, and S. Anwar, “FSBI: Deepfake detection with frequency enhanced self-blended images,” *Image and Vision Computing*, vol. 154, p. 105418, 2025.
- [21] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The DeepFake detection challenge (DFDC) dataset,” arXiv:2006.07397, 2020.
- [22] A. Yermakov, J. Cech, J. Matas, and M. Fritz, “Deepfake detection that generalizes across benchmarks,” in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2026.
- [23] N. A. Chandra, H. Lee, R. Murtfeldt, L. Qiu *et al.*, “Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024,” arXiv:2503.02857, 2025.
- [24] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for DeepFake forensics,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3204–3213.
- [25] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *Proc. IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019, pp. 83–92.
- [26] K. Tsigos, E. Apostolidis, S. Baxevas, S. Papadopoulos, and V. Mezaris, “Towards quantitative evaluation of explainable AI methods for deepfake detection,” in *Proc. 3rd ACM Int. Workshop on Multimedia AI against Disinformation (MAD)*, 2024.
- [27] X. Jin, W. Guan, W. Wang, and J. Dong, “Towards reliable deepfake detection from uncertainty calibration perspective,” *Visual Intelligence*, vol. 3, 2025.
- [28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [29] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, p. 125, 2020.
- [30] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” arXiv:1708.04552, 2017.
- [31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.