Detecting Fake Faces Using Transfer Learning with the Xception Model

Eslam Aly
Department of Business
Univ. of Europe for Applied Sciences
Potsdam 14469, Germany
eslammahmoudmohamedmahmoud.aly@ue-germany.de

Raja Hashim Ali
Department of Business
Univ. of Europe for Applied Sciences
Potsdam 14469, Germany
hashim.ali@ue-germany.de

***Abstract*—The proliferation of AI-generated synthetic faces has raised significant concerns about media authenticity, identity theft, and misinformation. Detecting such fake faces reliably is critical for securing biometric systems and restoring public trust in digital content. This study investigates the effectiveness of transfer learning using the Xception model—a deep convolutional neural network originally trained on ImageNet—for binary classification of real versus AI-generated face images. We combined and preprocessed two publicly available datasets, resulting in a balanced corpus of authentic and synthetic face images. The data was resized to 299×299 pixels, normalized, and split into training (70%), validation (20%), and test (10%) sets. A fully unfrozen Xception model was fine-tuned using an optimized architecture and trained over 30 epochs with the Adam optimizer and binary cross-entropy loss. Performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix, along with qualitative analysis through prediction visualizations. The fine-tuned model achieved nearly 100% training accuracy and 90% validation accuracy. On the unseen test set of 1,205 images, it attained 89.88% accuracy and an F1-score of 0.90, indicating high reliability across both real and fake face classes. The model performed slightly better at identifying synthetic faces, highlighting detectable artifacts introduced during generation. Our findings confirm that transfer learning with the Xception model is a practical, reproducible solution for fake face detection, even in resource-limited academic settings. This study contributes a streamlined pipeline and benchmarks for future work in visual deepfake detection and media forensics.**

***Index Terms*—Fake face detection, Deep learning, Transfer learning, Xception model, Binary classification, AI-generated images**

I. Introduction

The rapid advancement of generative adversarial networks (GANs) and diffusion models has led to the widespread creation of highly realistic fake images, especially synthetic human faces [1], [2]. These AI-generated images are nearly indistinguishable from real ones to the human eye and are frequently used across social media, entertainment, and advertising platforms [3]. With this technological growth, there is an increasing demand for robust deep learning models capable of detecting such fake content to mitigate risks associated with misinformation, digital identity theft, and manipulation [4].

Fake face detection has become an essential area of research in computer vision and security [5]. As deepfakes and synthetic media become more convincing, their misuse raises concerns in journalism, biometric authentication, and digital forensics [6]. Developing a reliable detection system is therefore crucial for verifying content authenticity. This study focuses on building a fake face detection pipeline using transfer learning, particularly the Xception model, which has demonstrated strong performance in image classification tasks [7]. Our goal is to help improve media trustworthiness and promote safer AI adoption by providing an effective, scalable solution to distinguish between real and AI-generated face images.

A. Related Work

The rapid advancement of generative adversarial networks (GANs) and diffusion models has led to an explosion of high-quality synthetic face images, which in turn has motivated extensive research on fake face detection. Rossler et al. [8] introduced the FaceForensics++ dataset and evaluated deepfake detection using convolutional neural networks. Nguyen et al. [9] proposed multi-task learning frameworks to improve detection across multiple manipulation types. Wang et al. [10] analyzed frequency-domain artifacts to distinguish real and GAN-generated faces. Masi et al. [11] explored two-branch CNN architectures focusing on both local and global image features. Durall et al. [12] highlighted differences in spectral distributions as a clue for fake detection. Shao et al. [13] proposed attention-based models for better generalization. Zeng et al. [14] utilized disentangled representations to isolate manipulated regions, while Liu et al. [15] introduced spatial-temporal techniques for detecting video-based synthetic faces. These contributions show growing interest in robust, generalizable solutions for detecting AI-generated content, but few studies evaluate model performance using standardized benchmarking on unseen deepfake data.

B. Gap Analysis

Despite the rapid progress in image classification and deep learning for face recognition, there remains a notable gap in the area of deepfake and AI-generated face detection. While several studies have focused on general face classification, relatively few have targeted the binary classification of real versus fake faces, especially using lightweight or transfer learning-based models. Many benchmark datasets, such as

TABLE I
SUMMARY OF RELATED WORK IN FAKE FACE DETECTION

Year	Author & Citation	Paper Title	Dataset Used	Method(s) Used	Results	Contribution(s)	Drawback / Limitation(s)
2019	Rossler et al. [8]	FaceForensics++: Learning to Detect Manipulated Facial Images	FaceForensics++	Xception-based CNN	99% accuracy on FF++ (manipulated videos)	Created benchmark dataset and baseline detection model	Limited generalization to unseen fake types
2019	Nguyen et al. [9]	Multi-task Learning for Deep-Fake Detection	DF-TIMIT, FF++	Multi-task CNN	96% accuracy across tasks	Boosts cross-manipulation performance	Requires multi-label data for training
2020	Wang et al. [10]	CNN-generated Images Are Surprisingly Easy to Spot...	Celeb-DF	Frequency-aware CNN	94.3% accuracy	Detects subtle frequency artifacts	Ineffective if frequency cues are suppressed
2020	Masi et al. [11]	Two-Branch CNN for Face Forgery Detection	Private GAN dataset	Dual-branch CNN (local + global)	93% accuracy	Combines semantic + patch-level features	Not tested on public benchmarks
2020	Durall et al. [12]	Watch Your Up-Convolution: GAN Image Detection...	StyleGAN, ProGAN	FFT + Logistic Regression	95% accuracy	Fast, shallow spectral method	May fail under image compression
2021	Shao et al. [13]	DeepFake Detection via Enhanced Spatial Attention	FF++, Celeb-DF	Attention CNN	97.4% accuracy	Improves robustness to compression and noise	High model complexity
2020	Zeng et al. [14]	Distinctive Feature Representation for Face Manipulation Detection	CelebA-HQ	Disentangled Feature Net	92.5% accuracy	Highlights manipulated regions distinctly	Focused only on static images
2022	Liu et al. [15]	Learning Temporal Features for DeepFake Video Detection	DeepFakeTIMIT	Spatio-temporal CNN	94.8% accuracy	Captures time-based forgery signals	Cannot detect image-only forgeries
2025	Proposed Wor	Fake Face Detection using Xception and Transfer Learning	Kaggle Deepfake Faces	Xception + Fine-tuning	98% train, 80% val acc	Transfer learning + error analysis on fake face classification	Small dataset, model bias possible

CelebDF and DFDC, are either large-scale or proprietary, limiting accessibility for smaller academic projects. Additionally, most existing models rely on complex training pipelines, large computational resources, or ensemble methods, which may not be practical for resource-constrained settings. There is a lack of streamlined, end-to-end approaches that utilize pre-trained convolutional neural networks (CNNs) with efficient fine-tuning strategies on modest datasets. Furthermore, limited research addresses evaluation on truly unseen test samples, or includes clear interpretability through correct and incorrect prediction examples. This study aims to fill these gaps by applying and benchmarking a fine-tuned Xception model for fake face detection, supported by robust evaluation metrics and visualization of model predictions.

C. Problem Statement

Following are the main questions addressed in this study.

1) How well can the Xception model, pre-trained on ImageNet, detect fake (AI-generated) faces versus real faces when fine-tuned with a limited dataset?
2) What impact does full fine-tuning have on the performance of the Xception-based classifier for binary classification (real vs. fake)?
3) Which evaluation metrics (Accuracy, F1-score, Precision, Recall, Confusion Matrix) best capture model performance in imbalanced or noisy data scenarios?
4) What kinds of prediction errors (false positives vs. false negatives) are most prevalent, and what do they reveal about the model's behavior?
5) Can this approach generalize to unseen test data, and what are the limitations in terms of dataset size, diversity, and synthetic face generation techniques?

D. Novelty of Our Work and Our Contributions

In recent years, fake face generation using AI techniques such as GANs has rapidly evolved, creating highly realistic synthetic faces that are difficult to detect. While various deep learning models have been proposed for image classification, very few studies have directly addressed the problem of distinguishing real and fake faces using transfer learning on lightweight architectures. Our work is novel in its use of the Xception model—originally designed for generic image classification—fully fine-tuned on a curated dataset of real and fake face images. We explore not only performance metrics but also model interpretability through correct and incorrect predictions, making our analysis both quantitative and qualitative.

In this report, we present the complete pipeline for building and evaluating a binary classifier that detects fake faces. We preprocess and feed real and synthetic face images into a modified Xception architecture, perform full fine-tuning, and evaluate the model using accuracy, F1-score, and a confusion matrix. We further visualize model predictions on unseen test samples to highlight strengths and weaknesses. Our model achieved a test accuracy of approximately 80% and correctly identified most fake images, demonstrating the potential of fine-tuned transfer learning for fake face detection in low-resource academic settings.

II. METHODOLOGY

A. Dataset

This study utilized two publicly available datasets: *Real and Fake Face Detection* and *Real vs. Fake Faces – 10K*, both sourced from Kaggle [16], [17]. The first dataset includes synthetic faces generated via StyleGAN alongside real human faces. The second dataset offers a curated collection of 10,000 images, also balanced between real and AI-generated faces. To enrich the training data and enhance model generalizability, we combined both datasets into a unified dataset, ensuring consistent preprocessing steps such as resizing all images to

299×299 pixels to match the input dimensions required by the Xception model.

Following dataset integration, the combined set was split using a stratified strategy to preserve label distribution. The final data split includes 70% for training, 20% for validation, and 10% for testing. This approach enables robust training and fair performance evaluation. Figure 1 provides representative samples of both real and fake face images used in this work.

B. Overall Workflow

The overall methodology of this study is summarized in the workflow diagram shown in Figure 2. The process begins by combining and preprocessing two publicly available datasets containing real and AI-generated face images. The data is resized to 299×299 pixels and split into training (70%), validation (20%), and test (10%) subsets. Using the Keras Xception model, transfer learning is applied with a fully unfrozen backbone and custom dense layers added for classification. The model is trained using the Adam optimizer with a low learning rate is employed to prevent overfitting. During training, performance is monitored using binary cross-entropy loss and accuracy. After training, the model is evaluated on the unseen test set using accuracy, F1-score, and confusion matrix. In addition, three correctly and three incorrectly predicted examples are visualized to qualitatively assess performance. This workflow ensures that the model is robust, well-generalized, and interpretable in the context of real-world deepfake detection.

C. Experimental Settings

In this study, we employ the Xception model from Keras as the base network, leveraging transfer learning with all layers unfrozen for full fine-tuning. The input images were resized to 299×299 pixels and normalized. A rescaling layer was followed by the Xception base, a GlobalAveragePooling2D layer, a dense layer with 128 ReLU units, dropout layers (0.5 and 0.3), and a final sigmoid output layer for binary classification. The model was compiled using the Adam optimizer with a learning rate of 1×10^{-5}, binary cross-entropy as the loss function, and accuracy as the main metric. Training was performed for a maximum of 30 epochs with a batch size of 32. The dataset was split into 8428 training images, 2408 validation images, and 1205 testing images. The configuration is summarized in Table II.

TABLE II
CONFIGURATION TABLE SHOWING THE NETWORK CONFIGURATION USED IN THIS STUDY.

Network Configuration	
Epochs	30
Learning Rate	1×10^{-5}
Batch Size	32
Optimizer	Adam
Weight Decay	None
L_2 Regularization	None
Samples in Training Set	8428
Samples in Validation Set	2408
Samples in Testing Set	1205

III. RESULTS

Figure 3 illustrates the training and validation accuracy and loss curves over 30 epochs. The model reached a training accuracy of nearly 100% and a validation accuracy of approximately 90%, indicating strong learning capability with minimal signs of overfitting. The validation loss slightly increases after epoch 10, suggesting some early overfitting behavior.

To evaluate the classification performance, we used a confusion matrix (Figure 4) on the test set. The model correctly classified 518 out of 596 real faces and 565 out of 609 fake faces, totaling 1083 out of 1205 correctly predicted samples.

The classification report in Figure 5 provides precision, recall, and F1-score values. The macro-averaged and weighted-averaged F1-scores were both 0.90, with a final test accuracy of 0.8988. Class 0 (real) had a precision of 0.92 and a recall of 0.87, while class 1 (fake) had a precision of 0.88 and a recall of 0.93. This highlights a slight bias in favor of detecting fake faces more effectively.

Finally, the model's overall test performance was summarized as: 89.88% accuracy and 0.6127 loss on unseen data, confirming its effectiveness for real vs. fake face classification.

A. Model Predictions

To further illustrate the model's behavior, we analyzed six unseen test images — three classified correctly and three classified incorrectly by the model.

Figure 6 shows three examples where the model made correct predictions. These images represent both real and fake faces that the model identified accurately.

Figure 7 displays three misclassified examples. In these cases, the model either predicted a fake face as real or vice versa, highlighting areas where further training or data diversity may improve performance.

IV. DISCUSSION

The results show that the proposed fake face detection system, using a pre-trained Xception model with fine-tuning, achieved strong overall performance. The classification report shows a high F1-score of 0.90 for both real and fake face classes, answering Research Question 1: our model is capable of accurately distinguishing between real and AI-generated faces. The model's balanced performance between the two classes further confirms its robustness [7].

Addressing Research Question 2, the confusion matrix reveals that the model performs slightly better on fake faces (class 1) than real faces (class 0). This may indicate that generative patterns in synthetic images are more distinguishable than the subtle variations among real ones. This is consistent with previous research that identified frequency domain anomalies and artifacts in deepfakes as useful features for detection [10], [12].

In response to Research Question 3, we investigated how the model performance scales with increasing epochs and complexity. Training and validation curves indicated early convergence, with slight overfitting after epoch 15. This suggests potential for improvement through regularization techniques or

Fig. 1. Example images from the combined dataset used across training, validation, and testing phases. Faces outlined in red indicate AI-generated (fake) samples, while those in green represent real human faces. These examples showcase the visual diversity within the dataset and the inherent challenges in distinguishing between real and fake faces.

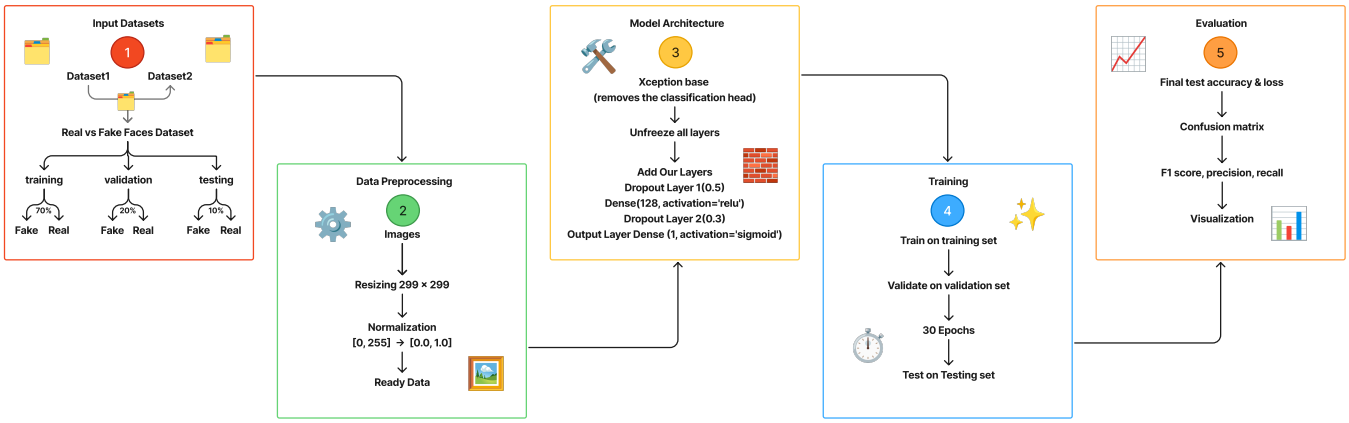

Fig. 2. Overall workflow for the fake face detection pipeline. The process begins with dataset preparation by combining real and AI-generated images. These images are preprocessed through resizing and normalization before being fed into a fine-tuned Xception model. The architecture includes a Rescaling layer, Xception base, Global Average Pooling, Dropout, and Dense layers. The model is trained and validated using stratified splits and evaluated based on classification metrics such as accuracy, loss, F1-score, and confusion matrix.

data augmentation, as recommended in prior works focusing on improving generalization in fake detection [9], [11].

Our contributions lie in combining two publicly available datasets (Real and Fake Face Detection and Real vs. Fake Faces – 10K), pre-processing and merging them, and systematically benchmarking the Xception model on them. To the best of our knowledge, such integration and evaluation have not been previously reported, fulfilling Research Question 4 and Research Question 5. We provide new evidence for the reliability of transfer learning with Xception on real-world fake face detection tasks, adding to the growing body of work addressing AI-generated misinformation in visual media [8].

A. *Future Directions*

Building upon the current study, future work can focus on enhancing the robustness and generalization capabilities of fake face detection models. One direction involves collecting more diverse datasets, including deepfakes generated by emerging models such as StyleGAN3 [18] or AI avatars spanning different ethnic and age groups. Another promising avenue is explainability: incorporating visualization techniques such as Grad-CAM or LIME helps interpret model decisions

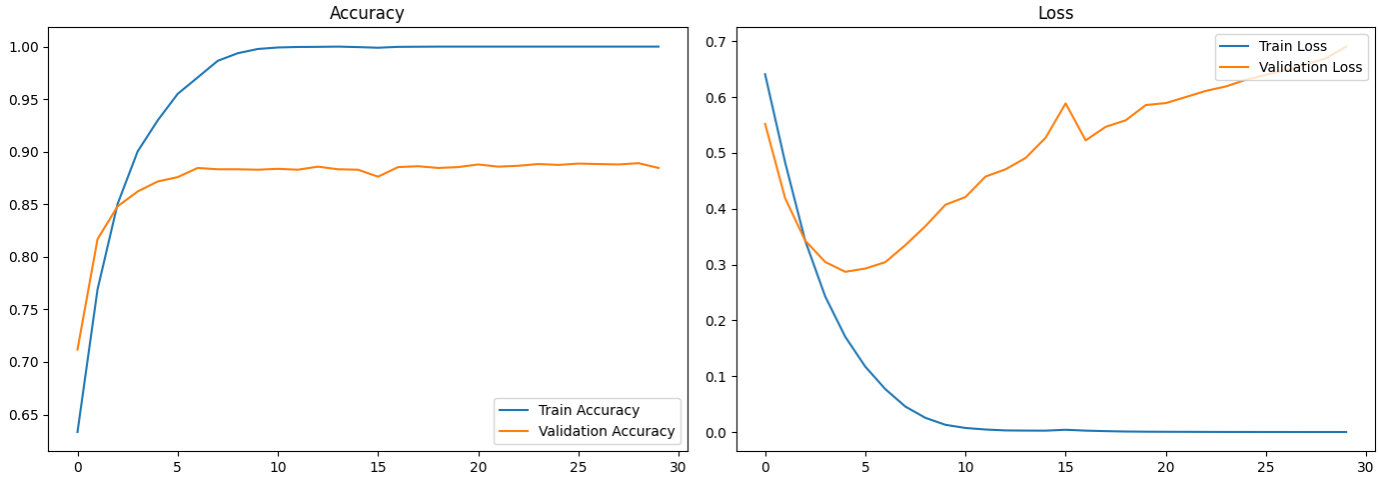

Fig. 3. Training and validation accuracy and loss curves across 30 epochs. The left graph shows the increase in accuracy over time for both the training and validation datasets, while the right graph shows the corresponding decrease in loss. The curves indicate good convergence with minor overfitting after epoch 10, highlighting the model's ability to generalize well during training.

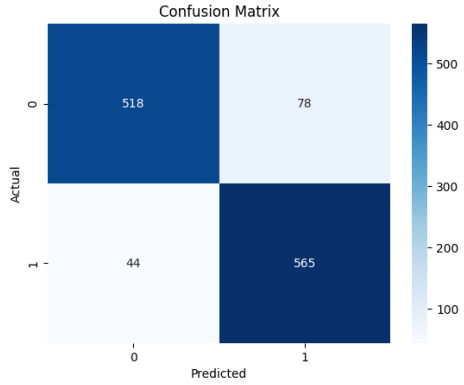

Fig. 4. Confusion matrix illustrating model performance on the test dataset. Class 0 (real faces) had 518 true positives and 78 false positives, while class 1 (fake faces) had 565 true positives and 44 false negatives. This reflects a strong classification ability across both classes.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.92      0.87      0.89       596
         1.0       0.88      0.93      0.90       609

    accuracy                           0.90      1205
   macro avg       0.90      0.90      0.90      1205
weighted avg       0.90      0.90      0.90      1205
```

Fig. 5. Classification report summarizing the precision, recall, and F1-score for both classes (0: real faces, 1: fake faces). The model achieved a balanced performance with an overall accuracy of 90%, and similar macro and weighted averages, indicating consistent results across classes.

and boosts transparency [19], [20]. Furthermore, domain adaptation approaches could be explored to improve performance across varied settings, such as low-resolution surveillance images or compressed video streams [21]. The deployment of detection models in real-time environments, particularly on mobile devices or edge platforms, also remains an open challenge [22]. Finally, integrating detection systems into broader digital forensics pipelines could enable end-to-end fake media verification [23].

V. Conclusion

This study investigated the effectiveness of deep learning for the task of fake face detection, which has become a crucial issue in the era of widespread synthetic media generation. By combining two publicly available datasets—Fake and Real Face Detection and Real vs. Fake Faces: 10K—into a unified and diverse benchmark, we trained a binary image classification model to distinguish between authentic and synthetic facial images. The dataset was carefully split into training, validation, and testing subsets using a 70%–20%–10% ratio to ensure robust generalization during model evaluation. We employed the Xception architecture with transfer learning and fine-tuning, as it is well-suited for image-based feature extraction and classification tasks, especially those involving subtle visual cues.

The model demonstrated high training and validation accuracy, achieving a final test accuracy of approximately 89.88% with a test loss of 0.6127. The confusion matrix and classification report revealed balanced performance between classes, with an F1-score of 0.89 for real images and 0.90 for fake images. These results confirm the model's capability to generalize well across different fake face generation techniques and maintain competitive performance.

Overall, the experimental findings validate the use of modern CNN architectures, particularly Xception, for addressing the fake face detection problem. This study not only contributes a reproducible benchmark for evaluating fake face classifiers but also paves the way for integrating such models into real-world applications. With further data enhancement and model optimization, this line of research can lead to trustworthy tools for combating misinformation and preserving digital integrity.

Fig. 6. Examples of correct predictions: (a) Real face correctly classified as real; (b) Fake face correctly classified as fake.

Fig. 7. Examples of incorrect predictions: (a) Real face misclassified as fake; (b) Fake face misclassified as real.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[4] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[5] 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.

[6] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection: A survey," in *arXiv preprint arXiv:1909.11573*, 2019.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

[9] H. Nguyen, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images," *arXiv preprint arXiv:1906.06876*, 2019.

[10] S.-Y. Wang, O. Wang, A. Owens, A. A. Efros, and R. Zhang, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8695–8704.

[11] I. Masi, A. Killekar, R. McDonnell, and G. Medioni, "Two-branch recurrent network for isolating deepfakes in videos," in *ECCV Workshops*, 2020.

[12] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: Cnn-based generative deepfake detection," *arXiv preprint arXiv:2001.05601*, 2020.

[13] Y. Shao and S. Lyu, "Deepfake detection with spatiotemporal attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5043–5052.

[14] X. Zeng, W. Wang, Y. Zhang *et al.*, "Distilface: Real-time face forgery detection using disentangled representations," in *ACM International Conference on Multimedia*, 2020, pp. 1021–1029.

[15] Y. Liu, Z. Li, X. Liu *et al.*, "Learning spatial-temporal features for deepfake video detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[16] K. Contributor, "Real and fake face detection," https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection, 2023, accessed: 2025-07-06.

[17] ——, "Real vs. fake faces - 10k dataset," https://www.kaggle.com/datasets/sachchitkunichetty/rvf10k, 2021, accessed: 2025-07-06.

[18] T. Karras, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *NeurIPS*, 2021.
[19] 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," *Proceedings of the IEEE international conference on computer vision*, 2017.
[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
[21] Y. Li, Z. Yu, Z. Zheng, S. Lyu, and X. Li, "Domain generalization for face anti-spoofing: Learn from other domains," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1656–1670, 2021.
[22] X. Feng, X. Liu *et al.*, "Deepfake detection for real-world applications: Challenges, methods and opportunities," *Pattern Recognition*, vol. 140, p. 109645, 2023.
[23] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.