IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 March 2026, accepted 20 March 2026, date of publication 3 April 2026, date of current version 13 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3680327

RESEARCH ARTICLE

Enhancing Deepfake Detection: Leveraging StyleGAN3 for Robust AI-Generated Forgery Identification

MOHAMMAD HAFEZI, ESSA Q. SHAHRA, SHADI BASURRA, ADEL ANEIBA, AND JACK DEVEY

Faculty of Computing, Engineering and Built Environment, Birmingham City University, B4 7RQ Birmingham, U.K.

Corresponding author: Essa Q. Shahra (Essa.Shahra@bcu.ac.uk)

ABSTRACT The rapid advancement of generative models has significantly increased the realism of AI-generated Deepfake content, posing serious challenges to digital media integrity and forensic analysis. A key difficulty in Deepfake detection lies in achieving robust generalization when confronted with synthetic images generated by previously unseen models that exhibit reduced visual artifacts. This study investigates the effectiveness of augmenting training data with StyleGAN3-generated images to enhance the generalization capability of Deepfake detection systems. Unlike earlier generative models, StyleGAN3 mitigates common artifacts such as texture sticking and aliasing, producing highly realistic synthetic faces that better represent modern forgery characteristics. We train a convolutional neural network (ResNet-18) under two controlled conditions: using a standard Deepfake dataset and using a dataset augmented with StyleGAN3-generated images. Experimental results demonstrate that the proposed augmentation strategy yields a 20.5% absolute improvement in test accuracy, along with a substantial increase in true positive rate and a significant reduction in false negatives. These findings indicate that exposure to more realistic synthetic samples enables the model to learn deeper and more transferable representations of manipulated content. However, the improvement in fake detection performance is accompanied by a moderate rise in false positives, highlighting an important trade-off that must be considered in practical deployment. Overall, this work demonstrates that incorporating artifact-reduced synthetic images during training can improve the robustness of Deepfake detection models. The study contributes to ongoing efforts in digital media forensics by emphasizing the importance of realistic data augmentation strategies for strengthening detection systems against evolving generative techniques.

INDEX TERMS Deepfake, AI generation, deep learning, StyleGAN3.

I. INTRODUCTION

Deepfakes, which are highly realistic media generated using deep neural networks, have gained widespread attention due to their ability to convincingly manipulate human imagery [1]. They allow malicious actors to manipulate and control an individual's identity, which can lead to serious consequences such as misinformation, defamation, and other harmful uses of a person's likeness, raising critical concerns regarding privacy and trust in digital media [2], [3]. The increasing realism of Deepfakes makes it increasingly challenging to distinguish them from authentic media, raising concerns in digital forensics and security [4], [5], [6], in particular images generated by Generative Adversarial Networks (GANs). Among GAN models, the state-of-the-art StyleGAN3 is designed to enhance image realism by overcoming texture inconsistencies present in earlier models. StyleGAN3's ability to produce high-quality, consistent images with fewer artefacts has expanded its potential for both creating Deepfakes and testing the robustness

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Kashif Bashir.

of detection models [7], [8], [9], [10], [11], [12], [13]. As a result, StyleGAN3 produces highly authentic-looking images with smoother transitions, posing a particularly difficult challenge for face forgery detection models. Another challenge arises from the limited generalisability and robustness of existing detection methods. Generalization refers to the ability to detect fake images generated by models not encountered during training, while robustness refers to maintaining detection accuracy under image perturbations such as compression or adversarial attacks [11], [14]. Many detection models are trained on specific datasets containing images generated by known GAN models. Although such models may achieve high accuracy on these datasets, their performance often drops significantly when tested on images generated by unseen GAN models or subjected to post-processing techniques such as compression, blurring, or resizing [6], [15], [16]. This limitation poses a significant obstacle to real-world utilisation of these models, where they need to handle diverse forgery techniques and unknown GAN architectures [14], [15], [16]. Researchers are continuously developing various detection algorithms and content authentication frameworks focusing on identifying inconsistencies in facial expressions, blending artefacts, and physiological signals [8], [11]. Other methods explore frequency domain analysis to detect high-frequency patterns unique to GAN-generated content [17]. One of the primary solutions is to improve the generalization capabilities of detection models, ensuring they can effectively identify Deepfakes generated by unseen models or subjected to various post-processing techniques [14], [15], [18]. Transfer learning and adversarial training are key strategies to achieve this [14]. Another vital approach is to develop robust evaluation frameworks that reflect real-world scenarios, which include standardising evaluation benchmarks and expanding datasets [6], [19]. The fight against AI-generated forgeries is an ongoing challenge that requires a multi-faceted approach. By continuously adapting detection techniques, exploring novel cues, and fostering collaboration between researchers, developers, and policymakers, it is possible to mitigate the risks posed by Deepfakes and preserve the integrity of digital information.

One of the major challenges in Deepfake detection is the rapid evolution of generative models, which continuously reduce visual artifacts and introduce distribution shifts between training data and real-world forgeries. Detection models often learn generator-specific cues or low-level artifacts present in earlier GAN architectures. When confronted with images produced by newer models—such as StyleGAN3, which mitigates issues like texture sticking and aliasing—these detectors frequently fail because the learned artifacts no longer exist. This mismatch between training distributions and real-world synthetic content leads to poor generalization and reduced robustness. Addressing this challenge requires training strategies that expose detectors to more realistic and artifact-free synthetic images, enabling them to learn deeper, more intrinsic characteristics of manipulated content. This study advances the field by investigating whether training detection models on StyleGAN3-generated images enhances their ability to identify manipulated content. Unlike previous generative models, StyleGAN3 mitigates issues such as "texture sticking," producing highly realistic images with fewer detectable artifacts. By leveraging these images for training, we hypothesize that detection models can move beyond reliance on superficial artifacts characteristic of earlier GAN architectures and instead learn to identify deeper, more fundamental features of synthetic images. This work not only introduces a novel augmentation strategy but also establishes a foundation for future research in strengthening Deepfake detection through advanced synthetic data techniques. In this work, we particularly frame Deepfake detection within the context of digital forensics and media authentication, where the cost of misclassifying forged content can be severe. In forensic investigations, journalism, and online content verification, detection systems must operate under real-world conditions involving unseen manipulation techniques and highly realistic synthetic media. StyleGAN3 represents a new level of visual realism that challenges traditional detection approaches. Therefore, improving detector robustness against such advanced generative models is not only a technical goal but a practical necessity for safeguarding trust in digital evidence and public media.

The key contributions of this study can be summarized as follows:

1) StyleGAN3-based Data Augmentation for Deepfake Detection: We introduce a novel augmentation strategy that incorporates highly realistic StyleGAN3-generated images into the training process, exposing detection models to more advanced and artifact-reduced synthetic content.
2) Comprehensive Generalization Evaluation: We systematically compare model performance when trained with and without StyleGAN3 augmentation under identical architectures and evaluation conditions, enabling a controlled analysis of its impact on generalization to unseen manipulations.
3) Robustness and Trade-off Analysis: We provide an in-depth evaluation of performance metrics and confusion matrices, demonstrating that StyleGAN3 augmentation significantly improves fake detection capability (reducing false negatives) while revealing an important trade-off in increased false positives.

The structure of the paper is organized as follows: Section II presents the literature review, Section III provides a detailed explanation of the methodology, Section IV elaborates on the experimental results and discussion, and finally, Section V concludes the study.

II. LITERATURE REVIEW

Recent research has highlighted the growing challenge of generalization in Deepfake detection. The Face Deepfake Detection Challenge [11] provides a comprehensive

benchmark study evaluating the robustness of existing detection methods. The study revealed a significant gap between in-dataset performance and real-world generalization, particularly in ''in-the-wild'' scenarios. Most successful approaches employed deep learning architectures, with EfficientNet-based models demonstrating strong performance. Extensive data augmentation was widely used to enhance generalization. Real images were taken from the CelebA [16] and FFHQ ([20] datasets, while Deepfake images were generated using architectures like StyleGAN [20], StyleGAN2 [7], StarGAN [21], and others. The dataset included both manipulated and unmanipulated samples with various post-processing operations. The top-performing model achieved 93.61% accuracy, highlighting the effectiveness of deep convolutional architectures. However, the study emphasized the need for more robust detectors capable of generalizing across unseen manipulations and real-world conditions. Another study, [22] addresses the challenge of overfitting in Deepfake detection models, particularly when generalising to unseen forgeries. The research introduces a multi-task disentanglement framework, with modified Xception architecture as the backbone, that separates forgery- irrelevant, method-specific, and common forgery features. By focusing only on the common features, the model improves its generalization capabilities. The framework is tested on benchmark datasets like FaceForensics++, DFDC, and CelebDF, achieving performance with AUC scores up to 0.945. Future work suggests further exploring disentanglement techniques to enhance detection robustness across different forgery types.

Yan et al. [23] investigated the generalization gap caused by distribution shifts between training and testing data. They proposed Latent Space Data Augmentation (LSDA) using EfficientNet-B4 as the backbone. LSDA introduced both within-domain and cross-domain augmentations in the latent space to expand forgery diversity and improve decision boundaries. Evaluated on FaceForensics++, DFDC, and CelebDF, the approach achieved AUC improvements up to 0.867. Future work suggested exploring broader latent space transformations to further strengthen robustness. Yin et al. [24] focused on detecting highly realistic face forgeries by introducing a Dynamic Fine-Grained Difference Capture (DFDC) module combined with a Multi-Scale Spatio-Temporal Aggregation (MSA) module. Their architecture emphasized inter-frame inconsistencies while suppressing natural facial movements. The method achieved state-of-the-art performance on FaceForensics++, Celeb-DF, and DFDC datasets, demonstrating strong improvements in both accuracy and AUC. The authors proposed further enhancements in intra-frame and inter-frame feature extraction.

The generalisability and pre-processing in Deepfake detection models are stated as a critical gap in Wodajo and Atnafu's [25] study. They proposed a method that combines CNN and ViT architectures, termed as Convolutional Vision Transformer (CViT), to enhance local and global feature extraction. The model was trained on the DFDC dataset, achieving 91.5% accuracy. Expanding the model by incorporating more diverse datasets to improve its accuracy and robustness is considered future work. Several other works have explored generalization from alternative perspectives. Yao et al. [26] proposed a game-theoretical framework, using EfficientNet, ResNet, and Xception architectures, to tackle the generalisation problem by diagnosing multi-order interactions among visual concepts. The models were trained using a custom dataset and evaluated across FaceForensics++, Celeb-DF (v1) and Celeb-DF (v2). The implementations acquired the highest AUC of 99.93% on the in-dataset evaluation. The paper suggests that further exploration of interactions and multi-order representations could inspire more effective methods for improving the generalisation of Deepfake detectors.

Beyond Deepfake detection, broader advances in computer vision have also focused on improving feature representation, generalization, and robustness. Transformer-based architectures, such as collaborative compensative transformer networks for salient object detection and recent efficient segmentation frameworks, demonstrate how combining global contextual modeling with fine-grained local features can significantly enhance visual understanding tasks [27]. These developments highlight the importance of learning robust multi-scale and context-aware representations, which are equally critical in Deepfake detection, where subtle inconsistencies must be distinguished from natural facial variations. Inspired by these trends, our work similarly emphasizes improving the model's exposure to diverse and highly realistic synthetic data so that it learns deeper, more transferable visual features rather than dataset-specific artifacts. According to S and Thillaiarasu [28], the increasing threat posed by Deepfake needs more accurate and precise detection models to counteract. They used three CNN-based models - ResNext, Xception, and an ensemble of both – and DFDC dataset for their experiments. The ensemble model achieved superior results compared to individual models, with an accuracy of 93%. The paper suggests expanding the ensemble model by incorporating additional facial features like skin tone and eye-blinking patterns. The authors also propose exploring other deep learning architectures to enhance detection accuracy further. Authors in [29] identifies the challenge of distinguishing real from fake facial images, particularly with novel datasets and complex synthetic faces. The authors employed Error Level Analysis (ELA) to pre-process images by analysing compression discrepancies. Several deep learning models, including VGG-16, VGG-19, ResNet-50, and InceptionV3 were fine-tuned for the task. They employed the Real and Fake Face Detection dataset published by Yonsei University [30]. The experiment faced a huge gap between validation and test metrics, achieving an accuracy of 91.97% for training and 64.49% for testing. The authors aim to develop a new CNN architecture capable of achieving better accuracy across multiple datasets by

TABLE 1. Recent CNN-based related work review.

Reference	Train/Test Dataset	Technique(s)	Performance	Future Work
[11]	FDDC + DFDC + OpenForensics + FF++ + CelebA + FFHQ + Custom	EfficientNet-B3	Acc. 93.61%	Generalisation and Robustness Enhancement
		EfficientNet-B5	Acc. 90.05%	
		Not mentioned	Acc. 75.38%	
		EfficientNet-B0 + EfficientNetV2-M + Vision Transformers	Acc. 72.62%	
		Discrete Cosine Transformer + custom CNN	Acc. 63.97%	
		Not mentioned	Acc. 40.61%	
		Not mentioned	Acc. 36.85%	
[23]	FF++_c23 + CDF-v1 + CDF-v2 DFD + DFDC + DFDCP	EfficientNet-B4 + LSDA	Ave. AUC 0.826	Generalisation and Robustness Enhancement
[24]	FF++ + Celeb-DF + DFDC	CNN + custom (DFDC + MSA)	Ave. Acc. 91.26%	Performance Improvement
[22]	FF++_c23 + CDF-v1 + CDF-v2 DFD + DFDC + DFDCP	Xception	Ave. AUC 0.763	Robustness Enhancement
[28]	DFDC + Custom	ResNext	Acc. 80%	Detection Accuracy Enhancement
		Xception	Acc. 78%	
		ResNext + Xception	Acc. 93%	
[25]	DFDC	CNNs and + Vision Transformers	Acc. 91.5%	Accuracy and Robustness Improvement
[31]	Real and Fake Face Detection	ShuffleNet + k-NN	Acc. 88.2%	Overcome Current Manipulation
		Alex Net + K-NN	Acc. 86.8%	
[29]	Real and Fake Face Detection	ResNet-50	Acc. 53.64%	Accuracy Improvement
		InceptionV3	Acc. 57.25%	
		VGG-19	Acc. 60.63%	
		VGG-16	Acc. 64.49%	
[26]	FF++ + Celeb-DF (v1) + Celeb-DF (v2)	ResNet-18	Ave. F-AUC 0.7976	Generalisation Improvement
		ResNet-34	Ave. F-AUC 0.8403	

exploring various feature extraction approaches. Another paper that has the same approach of combining ELA and CNNs (AlexNet and ShuffleNet) as the previous [31]. The deep features extracted from these models are classified using Support Vector Machine (SVM) and K-Nearest Neighbours (KNN). They used the same Yonsei University dataset of Real and Fake Face Detection [30]. ShuffleNet combined with KNN achieved the highest accuracy of 88.2%, while AlexNet combined with KNN reached 86.8%. The paper suggests further exploration into deepfake detection methods to overcome current manipulation techniques.

Recent advances (2023–2025) increasingly focus on transformer-based detection frameworks, cross-dataset robustness, explainability, and real-world deployment challenges. These works emphasize that modern Deepfake detection must move beyond artifact-based learning toward more generalized representation learning. Our approach aligns with this direction by improving training data realism rather than architectural complexity.

Table. 1 summary shows that most recent works are looking for improvements in the generalisation and robustness of their proposed models or techniques. Also, it indicates many published works have relied on the train dataset performance and this is the bottleneck when it comes to the real-world dataset tests as they might struggle even to get close to their train metrics performances.

A. RESEARCH GAP

Despite significant progress, a common limitation persists: many detection models perform well on training datasets but struggle when confronted with unseen generative models or post-processing operations. While prior research has explored architectural enhancements, multi-task learning, and latent-space augmentation, the potential of leveraging highly realistic synthetic data from StyleGAN3 as a training augmentation strategy remains largely unexplored. StyleGAN3's artifact-reduced image generation provides an opportunity to expose detection models to more challenging and realistic forgeries during training. This study addresses this gap by incorporating StyleGAN3-generated images into the training dataset and systematically evaluating the impact on model generalization and robustness. Rather than focusing on architectural novelty, we investigate whether improved training data realism can help models learn more transferable forgery features.

III. METHODOLOGY

This study follows a systematic experimental design to evaluate the impact of StyleGAN3-based data augmentation on Deepfake detection performance. The methodology includes dataset construction, synthetic data generation, model training using a controlled baseline architecture, and evaluation using standard performance metrics. An overview of the experimental pipeline is illustrated in Fig. 1.

A. DATASET SELECTION AND PREPARATION

In this study, our aim is to use a base model with different training datasets, one without and one with StyleGAN3 images to measure the effect of StyleGAN3 augmentation on the detection performance of the model. Both mentioned datasets were constructed and utilized, all rooted in "The Face Deepfake Detection Challenge" [11] which from now on we name it FDDC. The first dataset, the Original Dataset, sets the baseline metrics. The second dataset, the StyleGAN3 Augmented Dataset, aims to investigate the impact of StyleGAN3 augmentation on the same model performance metrics.

1) ORIGINAL DATASET

The Original Dataset consists of 10,000 real images and 5,000 fake images. The real images are evenly drawn from the FFHQ and CelebA datasets to provide diversity in facial appearances. The fake images include 1,000 samples each generated by StarGAN, GDWCT, AttGAN, StyleGAN, and StyleGAN2, covering a range of GAN-based synthesis methods. This dataset serves as the baseline for evaluating the impact of augmentation. A schematic representation of this composition is shown in Fig. 2.

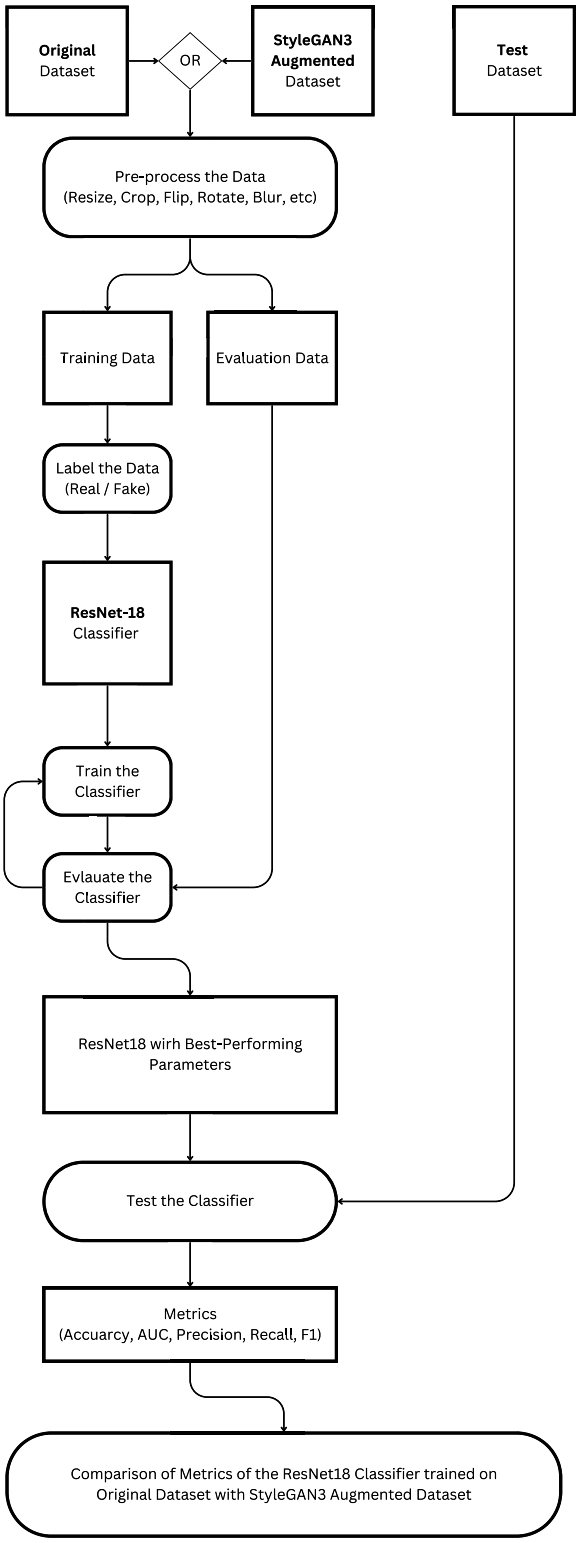

FIGURE 1. Workflow for deepfake detection using ResNet-18 and StyleGAN3 augmentation.

2) StyleGAN3 AUGMENTED DATASET

To investigate the effect of training data realism, we augmented the fake subset by incorporating 5,000 additional StyleGAN3-generated images. This augmentation includes

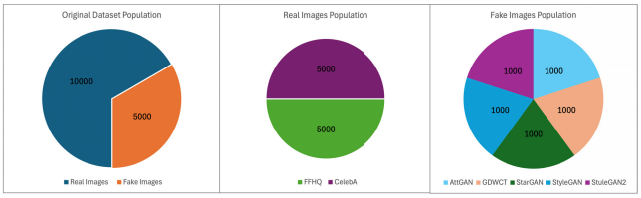

FIGURE 2. The original dataset includes 10000 real (picked from FFHQ and CelebA) and 5000 fake images (Picked from StarGAN, GDWCT, AttGAN, StyleGAN, and StyleGAN2.

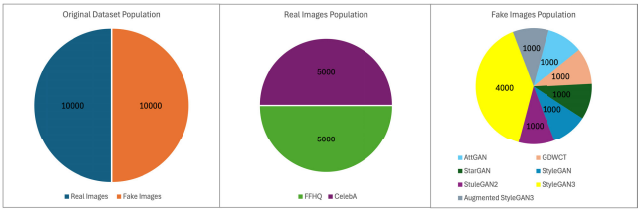

FIGURE 3. StyleGAN3 augmented dataset has 5000 StyleGAN3 images added to the original dataset.

4,000 original StyleGAN3-generated images and 1,000 further augmented versions of these images. The real image subset remains unchanged to ensure that performance differences arise solely from variations in synthetic training data. As shown in Fig. 3, this strategy enhances the diversity and realism of fake samples while maintaining a consistent real-image distribution. The goal of this augmentation is to expose the model to more sophisticated and artifact-reduced synthetic images, encouraging it to learn deeper and more transferable forgery features rather than relying on superficial GAN-specific artifacts.

3) TEST DATASET

The test dataset contains 7,000 images derived from the FDDC dataset and remains fixed across all experiments to ensure fair comparison. It includes a balanced distribution of manipulated images, with 1,000 images from each of seven source categories. To simulate real-world conditions, all test images are subjected to randomized post-processing operations including rotation, scaling, Gaussian noise, mirroring, and JPEG compression. This design ensures that the evaluation reflects generalization to unseen and perturbed samples. The structure of the test dataset is illustrated in Fig. 4.

- Rotation: a random integer is generated to determine the degree of rotation, as defined by the distribution presented in Equation (1), where R represents the angle of rotation in degrees.

$$R = 45 + 45U, \quad U \sim \text{Uniform}(0, 6) \tag{1}$$

- Scaling: a random integer number determined whether to reduce the image by 50% or magnify it by 100%.
- Gaussian Noise: A randomly selected integer which determines the size of the kernel (K) to be applied over the image, where $K \in \{3^2, 9^2, 12^2\}$.

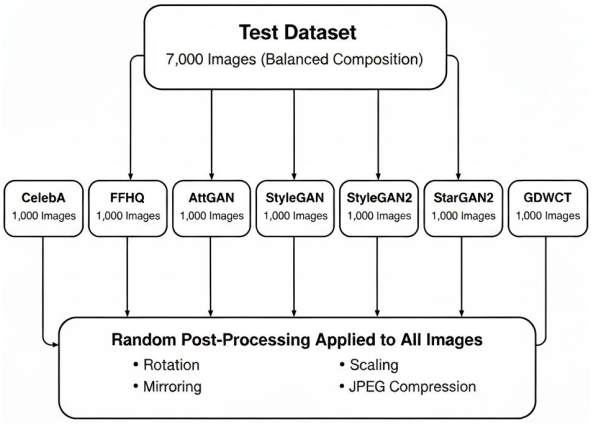

FIGURE 4. The test set includes 7000 images, 1000 from each subcategory.

- Mirror: a random integer number that represents mirroring horizontally, vertically or both.
- JPEG Compression: a random integer number generated in the range [50, 99] determined the quality factor parameter.

This study focuses on image-level Deepfake detection to specifically evaluate the influence of synthetic image augmentation on spatial feature learning. While video Deepfakes represent an important real-world scenario, video detection involves additional temporal modeling challenges beyond the scope of this work.

B. STYLEGAN3 IMAGE AUGMENTATION

A key component of this study is the generation of high-quality synthetic images using StyleGAN3. StyleGAN3 represents an advanced GAN architecture capable of producing highly realistic and artifact-reduced facial images. By training detection models on these images, we aim to reduce over-reliance on artifacts present in earlier generative models. We generated images using both StyleGAN3-r and StyleGAN3-t configurations to increase diversity. Images were produced at resolutions of 256×256 and 1024×1024 pixels and then resized to match the input requirements of the detection model. This approach ensures exposure to both coarse structural features and fine-grained details.

To further enhance diversity and improve generalization, we applied additional augmentation to 1,000 StyleGAN3-generated images using the following transformations:

- Random Rotation: Rotating images by a random angle introduces variability in the orientation of the faces. This is essential because, in real-world applications, faces can appear at different angles due to camera positioning or head movement. Incorporating this augmentation ensures that the model does not become overly sensitive to the original orientation of the images.
- Random Flip (Horizontal and Vertical): The random horizontal and vertical flipping of images simulates natural variations that occur in photographs, such as mirror images or reflections. While these transformations are simple, they are effective in increasing the dataset's variability, helping the model to avoid overfitting to specific image orientations. This approach is particularly useful in scenarios where the directionality of a face might change, ensuring that the model remains invariant to such changes.
- Random Gaussian Blur: Gaussian blur is applied to simulate the blurring that can occur in images due to motion, camera shake, or focus issues. By introducing varying degrees of blur, the augmentation process creates images that test the model's ability to detect forgeries even when the details are partially obscured. This augmentation is critical for ensuring that the model can perform well on low-quality or degraded images, which are common in real-world applications.
- Random Scaling: Scaling the images by a random factor allows for the simulation of different distances from the camera. This augmentation mimics the effect of zooming in or out, which is a common occurrence in photographs, aiming to improve robustness.
- Random Crop: Cropping the images to a random size helps the model learn to focus on different parts of the face. This augmentation is particularly useful for encouraging the model to recognize forgeries based on partial information, which is often the case when faces are partially obscured or only parts of the face are visible. By training on cropped images, the model becomes more adept at identifying key facial features that are indicative of Deepfakes. It also is a key strategy not to let the model to memorize patterns in the faces and consequent early overfitting of the model.
- Random Recompression: Recompressing images with varying quality levels introduces artifacts that are common in lossy image formats like JPEG. These artifacts can affect the visual quality of the images and potentially mask or reveal forgery cues. By incorporating recompression as an augmentation, we simulate the real-world scenario where images undergo multiple compression stages before being analysed.

Fig. 5 aims to give an insight into the 5000 StyleGAN3 images added to the Original Dataset.

C. MODEL ARCHITECTURE AND TRAINING

While more advanced architectures such as EfficientNet and Vision Transformers have shown strong performance in recent Deepfake detection studies, we intentionally selected ResNet-18 as a controlled baseline model. This allows us to isolate the impact of StyleGAN3 augmentation without architectural complexity influencing the results. Our goal is not to claim state-of-the-art detection performance, but to evaluate whether exposure to StyleGAN3-generated images improves generalization in a consistent and model-agnostic manner. Future work will extend this augmentation strategy to more advanced architectures. Given computational constraints, we opted for ResNet-18, balancing efficiency

FIGURE 5. This illustration shows a random grid image of 5000 StyleGAN3 images added to the original dataset to make the StyleGAN3 augmented dataset.

and performance. The ResNet-18 model was initialized with ImageNet pre-trained weights and fine-tuned for the binary classification task. Training was conducted using the Adam optimizer with a fixed learning rate of 0.0001 and a batch size of 64. The model was trained for 20 epochs using CrossEntropyLoss as the objective function. These settings were kept identical for both datasets to ensure a fair comparison of the impact of StyleGAN3 augmentation. We utilized PyTorch to facilitate rapid implementation and prototyping, ensuring a streamlined and flexible development process. We employ 75% of the whole training data for the training set to ensure the model can generalize across different types of images available in our datasets. The validation set, which is 25% of the rest of the data, was used to fine-tune the model during training. It included a mix of images not seen during training, allowing for an unbiased evaluation of the model's performance. The Test Dataset was reserved for the final evaluation of the model. It contained a mix of 7000 real and fake images, including images generated by GANs not included in the training set.

D. EVALUATION PERFORMANCE METRICS

The models are evaluated using several key metrics:

- Accuracy: This metric indicates the percentage of correctly classified instances. It is reported for both validation and test datasets.
- Loss: The training and validation loss values are tracked to assess model performance over epochs. Lower loss values indicate better model performance. Equation (2) represents the formula for calculating binary cross entropy loss, as utilised for the pipeline in this work.

$$L_{CE} = -\sum_{i=1}^{N} y_i \cdot \log \hat{y}_i \tag{2}$$

where N represents the number of images in the batch, y_i is an image's ground truth label and $\hat{y}_i$ is the Softmax probability for the i^{th} class.

- AUC (Area Under the Curve): This metric evaluates the model's ability to distinguish between classes, derived from the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. Equation (3) presents expressions for the calculation of TPR and FPR.

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN} \tag{3}$$

Hence, the AUC is computed as the integral of TPR with respect to FPR over the closed interval of $[0, 1]$, as expressed in Equation (4).

$$\text{AUC} = \int_0^1 TPR \, d(FPR) \tag{4}$$

The AUC value ranges from 0 to 1 with a value of 1 representing a perfect classifier and a value of 0.5 suggesting no better performance than random guessing.

- Precision, Recall, and F1 Score: These metrics provide insights into the model's performance concerning false positives and false negatives. Precision measures the accuracy of positive predictions, recall measures the ability to find all positive instances, and the F1 score is the harmonic mean of precision and recall, providing a balance between the two.

E. IMPLEMENTATION DETAILS AND REPRODUCIBILITY

To ensure experimental reproducibility, we provide additional implementation details of our training pipeline. All experiments were conducted using the PyTorch deep learning framework. Input images were resized to 224×224 pixels to match the ResNet-18 architecture requirements. During training, images were normalized using ImageNet mean and standard deviation values. The training-validation split was performed using a fixed 75%/25% ratio with random shuffling, while the independent test set remained unchanged across all experiments. A fixed random seed was used to ensure consistent data splitting and model initialization. Data augmentation operations (rotation, flipping, scaling, cropping, Gaussian blur, and JPEG recompression) were applied online during training. For StyleGAN3 image generation, both StyleGAN3-r and StyleGAN3-t configurations were used at resolutions of 256×256 and 1024×1024, with final training inputs resized to 224×224. Training was performed on a workstation equipped with an NVIDIA GPU, and each

TABLE 2. This table describes quantitative results of implementing the same model with both the original dataset and StyleGAN3 augmented dataset.

Training Dataset	Val. Loss	Val. Acc.	Val. AUC	Test Loss	Test Acc.	Test AUC	Test Pre.	Test Rec.	Test F1 Sco.
Original	0.024	0.9928	0.9929	2.5100	0.6889	0.7647	0.9617	0.5880	0.7298
StyleGAN3 Augmented	0.026	0.9906	0.9906	0.8642	0.8298	0.8438	0.9425	0.8112	0.8719

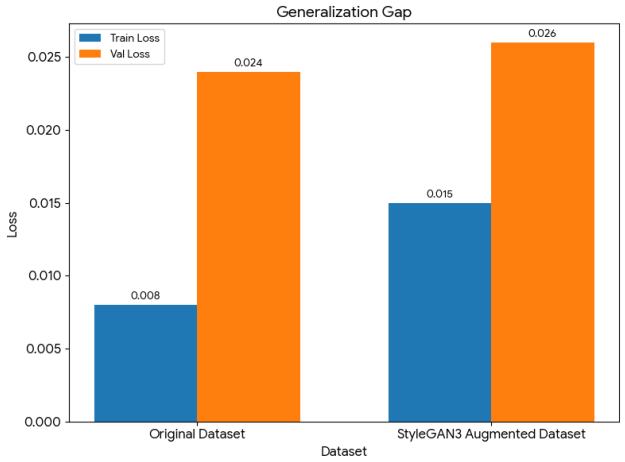

FIGURE 6. Training and validation loss comparison for the original and StyleGAN3 augmented datasets.

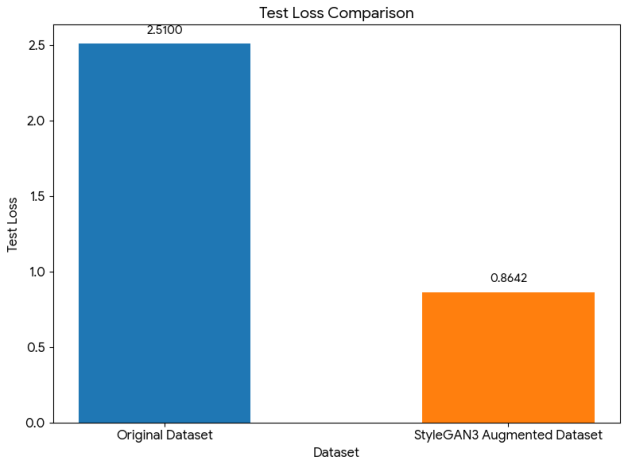

FIGURE 7. Test loss comparison for models trained on the original and StyleGAN3 augmented datasets.

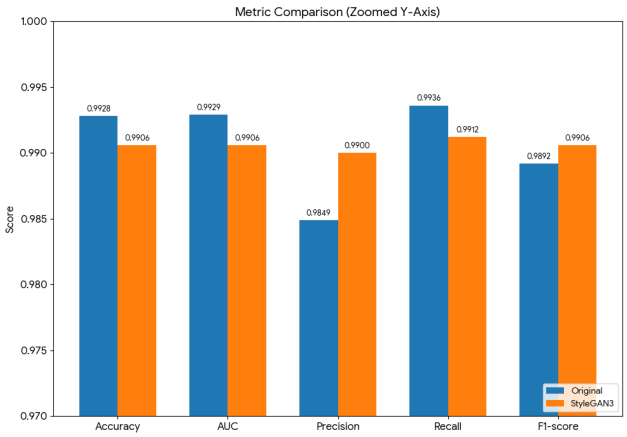

FIGURE 8. Validation performance comparison across accuracy, AUC, precision, recall, and F1-score for models trained on the original dataset and the StyleGAN3 augmented dataset.

experiment was repeated using identical training settings to ensure fair comparison between the Original Dataset and the StyleGAN3 Augmented Dataset.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Our experimental design can be interpreted as a controlled ablation study, where the only difference between the two training settings is the inclusion of StyleGAN3-generated images. This isolation enables direct assessment of the augmentation effect.

A. PERFORMANCE RESULTS

To evaluate the impact of StyleGAN3-based augmentation, we trained the ResNet-18 model under two conditions: using the Original Dataset and using the StyleGAN3 Augmented Dataset. The quantitative results are summarized in Table 2.

The generalization behavior of both training configurations is illustrated across Figs. 6–10. The training–validation loss comparison (Fig. 7) shows that the Original Dataset exhibits a larger gap between training loss (0.008) and validation loss (0.024), suggesting stronger overfitting, whereas the StyleGAN3 Augmented Dataset presents a smaller discrepancy (0.015 vs. 0.026), indicating improved regularization. This observation is further reinforced by the test loss comparison (Fig. 7), where the Original Dataset produces a substantially higher test loss (2.5100) compared to the StyleGAN3 Augmented Dataset (0.8642), demonstrating superior robustness to unseen data when StyleGAN3 augmentation is applied. While validation performance remains similarly high for both models across all metrics (Fig. 8), confirming comparable in-distribution learning, clear differences emerge in the test results (Fig.9). The StyleGAN3-augmented model achieves notable improvements in accuracy, AUC, recall, and F1-score, with a particularly significant increase in recall, indicating a substantial reduction in missed fake detections. Finally, the validation-to-test accuracy comparison (Fig. 10) highlights the pronounced generalization gap of the Original Dataset (99.28% to 68.89%) compared to the smaller performance drop observed with StyleGAN3 augmentation (99.06% to 82.98%). Collectively, these results demonstrate that incorporating highly realistic synthetic images during training enhances generalization and reduces overfitting, leading to more robust Deepfake detection on unseen manipulations.

The model trained with the StyleGAN3 Augmented Dataset exhibited a significant improvement in its ability to detect fake images (Fig. 11). Specifically, it recorded a

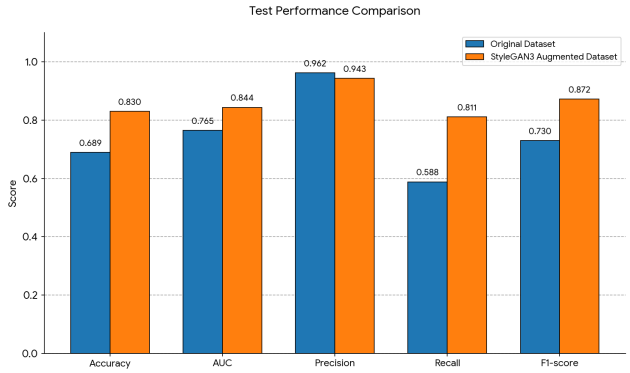

FIGURE 9. Test performance comparison across accuracy, AUC, precision, recall, and F1-score for models trained on the original dataset and the StyleGAN3 augmented dataset.

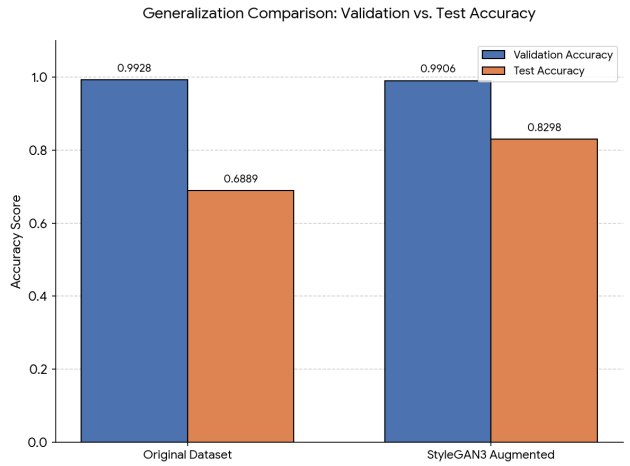

FIGURE 10. Comparison of validation and test accuracy for models trained on the original dataset and the StyleGAN3 augmented dataset.

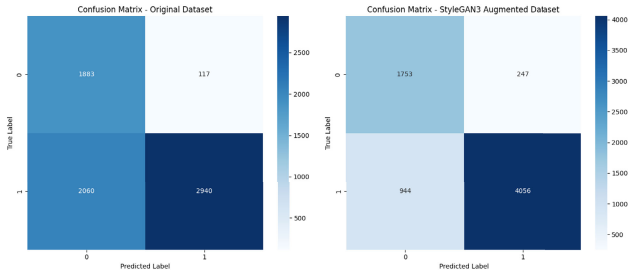

FIGURE 11. Confusion matrices show that the model trained with StyleGAN3 is better at minimising false negatives.

notable increase in true positives (4056 compared to 2940) and a reduction in false negatives (944 compared to 2060), indicating a marked improvement in the model's capacity to identify manipulated content. Nevertheless, this advancement in the classification of fake images was accompanied by a slight decline in the classification of real images, as evidenced by a decrease in true negatives (1753 compared to 1883) and an increase in false positives (247 compared to 117). These results underscore the inherent trade-offs in performance when employing StyleGAN3 augmentations to reinforce model robustness.

B. GENERALIZATION EVALUATION

The capacity of a model to generalize is paramount, particularly when evaluating its performance on real-world images that differ from those in the training set. As we discussed before the Test Dataset is a separate dataset including images that the model has not seen before. Despite both datasets achieving high validation accuracy, the testing phase revealed notable discrepancies. The Original Dataset exhibited weaker generalization, as indicated by a test accuracy of 68.89% and an AUC of 76.47%. In contrast, the StyleGAN3 Augmented Dataset demonstrated superior performance, achieving a test accuracy of 82.98% and an AUC of 84.38%, a significant 20.5% accuracy improvement (Fig. 9). These results affirm that the augmented dataset significantly enhances the model's ability to adapt to diverse and previously unseen data distributions.

C. ROBUSTNESS ANALYSIS

The robustness of the model against adversarial and manipulated data was evaluated through an analysis of the confusion matrices for both datasets. The StyleGAN3 Augmented Dataset markedly improved the model's robustness in detecting fake images, as evidenced by a reduction in the false negative rate from 2060 to 944, effectively halving the number of missed fake images. However, the slight increase in false positives (247 compared to 117) suggests that the model has become more conservative, occasionally misclassifying genuine images as fake. This trade-off indicates that while the augmented dataset enhances the model's capability in detecting forgeries, it may slightly compromise its precision regarding real images. The increase in false positives could be due to the high realisticity of StyleGAN3-generated images, which made the model unable to distinguish real images from fakes in some cases. The observed increase in false positives suggests that the model becomes more sensitive to subtle visual inconsistencies after exposure to highly realistic synthetic images. Potential mitigation strategies include threshold calibration, incorporating additional real-world negative samples, ensemble-based verification, and integrating frequency-domain or physiological cues to complement spatial detection.

D. COMPARATIVE ANALYSIS

A thorough analysis of the performance metrics reveals the substantial advantages of employing StyleGAN3 augmentation in the training of robust Deepfake detection models. The results indicate marked improvements in key performance indicators, particularly in recall and F1 score, when comparing models trained with StyleGAN3-augmented data to those trained on traditional datasets. Specifically, the recall metric for the model utilizing the StyleGAN3 Augmented Dataset reached an impressive 81.12%, in stark

contrast to the 58.80% recall achieved by the model trained on the Original Dataset. Recall, which measures the model's ability to correctly identify positive instances (in this case, manipulated images), is crucial in the context of deepfake detection. A higher recall indicates that the model is more effective at capturing a greater proportion of actual positive cases, thereby reducing the likelihood of false negatives.

Furthermore, the F1 score, which serves as a harmonic mean of precision and recall, also demonstrated significant enhancement, with the StyleGAN3-augmented model achieving an F1 score of 87.19% compared to 72.98% for the Original Dataset model. The F1 score is particularly important as it provides a balanced measure of a model's accuracy, especially in scenarios where there is an uneven distribution of classes. The improvement in the F1 score suggests that the StyleGAN3 augmentation not only increases the model's ability to detect manipulated images but also maintains a favourable precision level, thereby ensuring that the instances it identifies as positive are indeed accurate. These results strongly imply that the integration of StyleGAN3-augmented data allows the model to more effectively learn and capture the complex and nuanced features characteristic of manipulated images. By exposing the model to StyleGAN3-generated images, it becomes better equipped to recognize the subtle artifacts and variations that may be present in real-world Deepfakes.

In analyzing the training losses associated with the Original Dataset and the StyleGAN3 Augmented Dataset, we observed values of 0.008 and 0.015, respectively. Based on these figures, we might anticipate corresponding outcomes in the validation and testing phases. However, as illustrated in Fig. 6 despite the Original Dataset demonstrating a training loss that is nearly twice as favourable as that of the StyleGAN3 Augmented Dataset, both datasets exhibited comparable validation losses of 0.24 and 0.26. This unexpected similarity in validation loss suggests that the model's performance on unseen data does not align with the training loss metrics. More strikingly, during the testing phase, the Original Dataset yielded a significantly higher loss of 2.51, in stark contrast to the StyleGAN3 Augmented Dataset, which achieved a loss of only 0.86. These findings are indicative of an early overfitting issue, discussed in the research gap (Sec II-A), associated with the Original Dataset. The results from the StyleGAN3 Augmented Dataset suggest that the incorporation of synthetic images has mitigated the overfitting problem to a considerable extent. By providing a more diverse and challenging set of training examples, the StyleGAN3 Augmented Dataset has enabled the model to develop a more robust understanding of the features associated with both real and manipulated images. This enhancement in generalization capability is reflected in the significantly lower testing loss, indicating that the model is better equipped to handle variations in data that it has not encountered during training.

E. INTERPRETATION AND IMPLICATIONS

The experimental findings highlight the potential of StyleGAN3 augmentation to reinforce model robustness in the domain of face forgery detection. The enhanced detection of fake images, as evidenced by the increased true positive rate and reduced false negatives, underscores the model's effectiveness in identifying manipulated content — a critical requirement for forgery detection systems. However, the trade-off observed in the classification of real images, characterized by an increase in false positives, necessitates careful consideration in deployment scenarios where misclassifying genuine images could have severe repercussions. Future research could explore strategies to mitigate this trade-off, such as fine-tuning classification thresholds, incorporating additional real-world data during the training process, or using more complex augmentation and haperparameters during training to further decrease the overfitting issue. These results provide valuable insights into the application of generative data augmentations, such as StyleGAN3, for enhancing deep learning models in challenging domains, paving the way for the development of more reliable and robust forgery detection systems.

V. CONCLUSION

This study investigated the impact of incorporating StyleGAN3-generated images into the training process of Deepfake detection models. By exposing the detector to highly realistic and artifact-reduced synthetic faces, we aimed to improve generalization to modern forgery techniques that challenge traditional detection approaches. Experimental results demonstrated a clear performance gain, with a 20.5% absolute improvement in test accuracy and a substantial reduction in false negatives when compared to training on a conventional dataset alone. These findings suggest that realistic synthetic data can help detection models learn more transferable and robust feature representations, rather than relying on generator-specific artifacts. At the same time, the observed increase in false positives highlights an important trade-off between sensitivity to manipulated content and preservation of genuine image classification. This underscores the need for balanced detection strategies and threshold calibration in practical deployment scenarios, particularly in forensic and media verification contexts where both missed detections and false alarms carry significant consequences.

The primary contribution of this work lies in demonstrating that improvements in training data realism—rather than architectural complexity alone—can meaningfully enhance Deepfake detection robustness. While the present study focuses on image-based analysis using a controlled baseline model, the findings provide a foundation for extending StyleGAN3-based augmentation to more advanced architectures, cross-dataset evaluations, and video-based detection frameworks. Future research should further explore integration with transformer-based models, adversarial robustness

testing, and hybrid detection approaches that combine spatial, frequency, and temporal cues to address the evolving landscape of AI-generated media.

REFERENCES

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 7:1–7:41.

[2] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[3] C. Liu, "Deep image forgery: An investigation on forensic and anti-forensic techniques," Univ. Technol., Sydney, NSW, Australia, 2023.

[4] L. A. Passos, D. Jodas, K. A. P. Costa, L. A. Souza Júnior, D. Rodrigues, J. Del Ser, D. Camacho, and J. P. Papa, "A review of deep learning-based approaches for deepfake content detection," *Expert Syst.*, vol. 41, no. 8, p. 13570, Aug. 2024, doi: 10.1111/exsy.13570.

[5] M. Joslin and S. Hao, "Attributing and detecting fake images generated by known GANs," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2020, pp. 8–14.

[6] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," 2019, *arXiv:1909.06122*.

[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[8] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, and D. Tao, "Deepfake generation and detection: A benchmark and survey," *ACM Comput. Surv.*, 2024.

[9] H. Goyal, M. S. Wajid, M. A. Wajid, A. M. Ud Din Khanday, M. Neshat, and A. Gandomi, "State-of-the-art AI-based learning approaches for deepfake generation and detection, analyzing opportunities, threading through pros, cons, and future prospects," 2025, *arXiv:2501.01029*.

[10] A. R. Revanda, C. Fatichah, and N. Suciati, "Utilization of generative adversarial networks in face image synthesis for augmentation of face recognition training data," in *Proc. Int. Conf. Comput. Eng., Netw., Intell. Multimedia (CENIM)*, Nov. 2020, pp. 396–401.

[11] L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L. M. Q. Bui, M. Fontani, D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, N. Messina, G. Amato, G. Perelli, S. Concas, C. Cuccu, G. L. Marcialis, and S. Battiato, "The face deepfake detection challenge," *J. Imag.*, vol. 8, no. 10, p. 263, 2022.

[12] Y. Alaluf, O. Patashnik, Z. Wu, A. Zamir, E. Shechtman, D. Lischinski, and D. Cohen-Or, "Third time's the charm? Image and video editing with StyleGAN3," in *Comput. Vision—ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., Cham, Switzerland: Springer, 2022, pp. 204–220.

[13] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 852–863.

[14] J. Sabel and F. Johansson, "On the robustness and generalizability of face synthesis detection methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 962–971.

[15] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Comput. Vis. Image Understand.*, vol. 204, Mar. 2021, Art. no. 103170.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[17] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6454–6463.

[18] G. J. Aniano Porcile, J. Gindi, S. Mundra, J. R. Verbus, and H. Farid, "Finding AI-generated faces in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 4297–4305.

[19] X. Wang, H. Guo, S. Hu, M. Chang, and S. Lyu, "GAN-generated faces detection: A survey and new perspectives," in *Proc. ECAI*, 2022, pp. 2533–2542.

[20] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[21] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[22] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 22412–22423.

[23] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 8984–8994.

[24] Q. Yin, W. Lu, B. Li, and J. Huang, "Dynamic difference learning with spatio-temporal correlation for deepfake video detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4046–4058, 2023.

[25] D. W. Deressa, P. Lambert, G. Van Wallendael, S. Atnafu, and H. Mareen, "Improved deepfake video detection using convolutional vision transformer," in *Proc. IEEE Gaming, Entertainment, Media Conf. (GEM)*, Jun. 2024, pp. 1–6.

[26] K. Yao, J. Wang, B. Diao, and C. Li, "Towards understanding the generalization of deepfake detectors from a game-theoretical view," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2031–2041.

[27] J. Chen, H. Zhang, M. Gong, and Z. Gao, "Collaborative compensative transformer network for salient object detection," *Pattern Recognit.*, vol. 154, Oct. 2024, Art. no. 110600.

[28] A. G. S and N. Thillaiarasu, "Investigation of comparison on modified CNN techniques to classify fake face in deepfake videos," in *Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2022, pp. 702–707.

[29] N. Nida, A. Irtaza, and N. Ilyas, "Forged face detection using ELA and deep learning techniques," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 271–275.

[30] S. S. Pandi, M. Monesh, and B. Lingesh, "A novel approach to detect face fraud detection using artificial intelligence," in *Proc. 2nd Int. Conf. Emerg. Trends Inf. Technol. Eng. (ICETITE)*, 2024, pp. 1–6.

[31] R. Rafique, M. Nawaz, H. Kibriya, and M. Masood, "DeepFake detection using error level analysis and deep learning," in *Proc. 4th Int. Conf. Comput. Inf. Sci. (ICCIS)*, Nov. 2021, pp. 1–4.

MOHAMMAD HAFEZI received the B.Sc. degree in electrical engineering (control and instrumentation) from Montazeri University, Mashhad, Iran, in 2010, the M.Sc. degree in electrical engineering (automation and instrumentation in oil) from Shiraz University, Shiraz, Iran, in 2017, and the M.Sc. degree in big data analytics from Birmingham City University, Birmingham, U.K., in 2024. He is currently an Electrical Engineer with Berry Global, U.K., and also a Freelance Technical Data Insight Consultant. He has over 12 years of experience in the oil and gas industry, with a focus on leveraging overlooked operational data to improve process performance and maintenance strategies, bridging the eras of traditional engineering and data-driven innovation. He is passionate about combining classical engineering with modern data science to enhance decision-making and operational efficiency. His research and professional interests include hazardous area electrical design, predictive maintenance, industrial automation, and the integration of data analytics in engineering systems.

ESSA Q. SHAHRA received the bachelor's degree in computer science from Taiz University, Yemen, the master's degree from the King Fahd University of Petroleum and Minerals, Saudi Arabia, and the Ph.D. degree from Birmingham City University (BCU), U.K., with a research focus on artificial intelligence (AI) and the Internet of Things (IoT). During his Ph.D. studies, he was an Early-Stage Researcher under the prestigious Horizon 2020 Program, for three years. Upon completion of his doctorate, he joined Birmingham City University as a Lecturer in computer science with the Department of Computing and Data Science. He has played a leading role in several funded research projects, working as an effective research coordinator and making substantial contributions to innovation and knowledge in his field. He has authored numerous peer-reviewed publications in high-impact international journals and conferences. His research interests include artificial intelligence (AI), the Internet of Things (IoT), wireless sensor networks, fog computing, and indoor localization. With a solid interdisciplinary foundation, he actively bridges the gap between academia and industry, fostering impactful research and practical innovation.

SHADI BASURRA received the B.Sc. degree (Hons.) in computer science from the University of Exeter, the M.Sc. degree in distributed systems and networks from the University of Kent, and the Ph.D. degree from the University of Bath in collaboration with the University of Bristol. Following the completion of his Ph.D., he joined Sony Corporation, where he contributed to the development of goal decision systems. He later took on a Research Fellow role at the Zero Carbon Laboratory, Birmingham City University (BCU). Currently, he is a Professor of intelligent systems and the Head of the Department of Computing and Data Science, BCU. His research interests span across a wide range of areas, including the simulation and emulation of networks (vehicular, mesh, and sensor ad hoc networks), game theory and multi-agent systems, multi-objective optimization and model calibration, and dynamic simulation of zero-carbon design and building retrofit. He is particularly passionate about the intersection of intelligent systems and sustainable technologies.

ADEL ANEIBA received the B.Sc. degree in computer science from the University of Benghazi, Libya, in 1997, and the M.Sc. and Ph.D. degrees in computer networks from Staffordshire University, U.K., in 2003 and 2008, respectively. He is currently the Head of the College of Computing, Birmingham City University. He is also the Creator of the Cyber-Physical Systems (CPS) Research Group, Birmingham City University. This group specializes in software-defined networking (SDN), network function virtualization (NFV), high computing performance (HCP), future networks (5G and LPWAN), AI/ML, blockchain, and trustworthy protocols and systems. He is also the Experienced Academic Leader with 20 years of service in the industry and academia. He is a member of the research bids reviewers at the EU H2020 Evaluation Panel for their Digital Innovation Hubs (DIH) Federation for large-scale adoption of digital technologies by European SMEs, ''DigiFed.'' He has secured and led several research projects in the area of networking and the IoT funded by UKRI worth £1.2M. His research contribution has reached out to and impacted key global challenges, such as clean energy and climate change. His CoP26 Project was funded by British Council, Japan Sector, in partnership with a top global ranking university in Japan (The University of Tokyo) and the Universitas Negeri Gorontalo, Indonesia, around bioenergy using smart sensing technologies and state-of-the-art connectivity and intelligent solutions (5G/IoT/AI/ML).

JACK DEVEY is currently a Research Assistant in robotics with Birmingham City University. His research interests include multi-agent UAV swarms, computer vision, and machine learning. His work particularly focuses on the analysis of multi-agent UAVs through simulational and ad hoc networking developments using co-simulation approaches. He is also a key member of the Computing Research Team, BCU, working on various projects aimed at leveraging robotics to enhance search and rescue operations, making impacts through advancements in computer vision and machine learning. In addition to his research contributions, his current explorations involve the deployment of large language models (LLMs) in adversarial image explanation and classification, aiming to enhance the robustness and interpretability of AI systems in this domain.