

A Systematic Review of Facial Data Augmentation for Deepfake Detection

Naveed Ur Rehman Ahmed ¹, Afzal Badshah ², Ayesha Tajammul ³, Omar Alghushairy ⁴,
Riad Alharbey ⁴, and Ali Daud ^{5,*}

¹ Department of Computing, Hamdard University, Islamabad Campus, Islamabad, Pakistan

² Department of Software Engineering, University of Sargodha, Sargodha, Punjab, Pakistan

³ U.S.-Pakistan Center for Advanced Studies in Water, Mehran University of Engineering and Technology,
Jamshoro, Sindh, Pakistan

⁴ Department of Information Systems and Technology, College of Computer Science and Engineering,
University of Jeddah, Jeddah, Saudi Arabia

⁵ Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates

Email: nraalvi@gmail.com (N.U.R.A.); afzalbadshahkhattak@gmail.com (A.B.); drayetalvi@gmail.com (A.T.);
oalghushairy@uj.edu.sa (O.A.); ralharbi@uj.edu.sa (R.A.); alimsdb@gmail.com (A.D.)

*Corresponding author

Abstract—Deepfake technology has emerged as a major challenge in combating misinformation and manipulated digital content, posing serious risks to individual privacy, institutional credibility, and public trust. The increasing accessibility of generative models has significantly amplified the production and dissemination of highly realistic synthetic media, thereby intensifying the need for robust detection mechanisms. This study investigates facial data augmentation techniques employed to enhance machine learning and deep learning models for Deepfake detection. It presents a structured and comprehensive survey of augmentation strategies, highlighting their methodological foundations, practical implementations, and reported effectiveness. The review systematically examines geometric and photometric transformations, noise injection methods, occlusion strategies, cutout and mixing-based augmentations, and commonly used Python-based augmentation libraries, assessing their contribution to improving model generalization and mitigating overfitting. In addition, leading Deepfake detection approaches are summarized alongside the datasets and performance metrics frequently reported in the literature, including accuracy and Area Under the Curve (AUC), to provide a clearer understanding of existing evaluation practices. The study further discusses key challenges, dataset dependencies, and unresolved research issues that influence augmentation effectiveness across different experimental settings. By synthesizing current findings and identifying methodological limitations, this review aims to guide researchers in selecting and designing augmentation strategies that support the development of more robust, reliable, and generalizable Deepfake detection frameworks.

Keywords—deepfake detection, machine learning, data augmentation, image preprocessing

I. INTRODUCTION

The ability to create and manipulate visual content has advanced significantly with the rise of Artificial Intelligence (AI)-powered imaging processes. Previously, generating or disseminating fake content required substantial professional expertise; however, modern tools and techniques driven by machine learning and deep learning approaches [1–4] have made the creation of Deepfake content far more accessible. Fig. 1 highlights the growing prevalence of Deepfakes, illustrating how frequently users encounter such content and how human detection accuracy varies. These findings underscore the increasing reliance on AI-generated media and reveal both the challenges associated with detecting Deepfakes and the need for more advanced detection methods [5–8]. Consequently, further investigation into Deepfake techniques, particularly in video and image manipulation, has become essential.

The term “Deepfake” originated in late 2017, when a pseudonymous Reddit user employed deep learning algorithms to create highly realistic face-swapped videos, often in inappropriate or non-consensual contexts [9–11]. Although Deepfake technology offers potential applications in various domains [12–15], it also presents serious risks, including identity theft, revenge pornography, defamation of public figures, hoaxes, and the spread of misinformation during critical events such as political elections [16, 17]. Recent incidents highlight its real-world impact. In 2023, a Maryland school official was imprisoned for circulating a racist Deepfake audio targeting a principal [18]. In 2024, a fabricated video of Ukrainian President Zelenskyy urging soldiers to surrender went viral, drawing global attention to the

dangers of manipulated media [19]. By 2025, Deepfake controversies had expanded across entertainment and politics, including manipulated recordings of Chandrababu Naidu (Chief Minister of Andhra Pradesh,

India) falsely threatening farmers, further intensifying concerns regarding synthetic media used for disinformation [20].

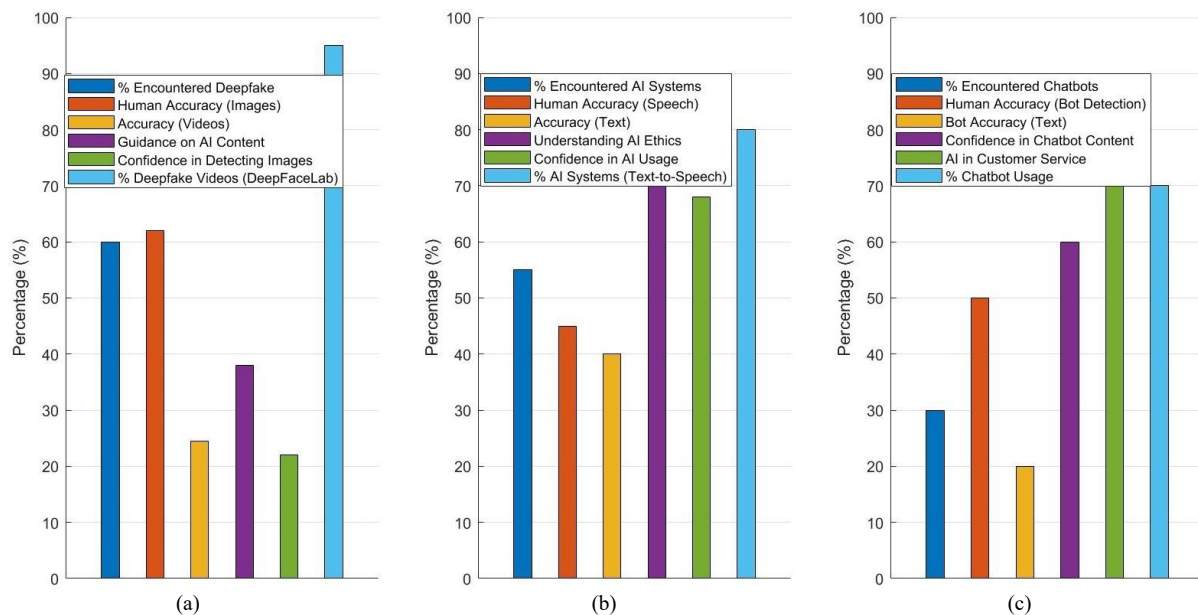


Fig. 1. Deepfake statistics: (a) Deepfake statistics; (b) AI system statistics; (c) Chatbot statistics.

Deepfake media is predominantly generated using Generative Adversarial Network (GAN) [21]. A GAN consists of two neural networks: a generator and a discriminator. The generator employs autoencoders and decoders to synthesize fake samples, while the discriminator utilizes machine and deep learning-based classifiers to distinguish real content from manipulated content [22, 23]. With rapid advancements in Deepfake generation tools, including mobile applications and web-based platforms [24–26], the demand for robust detection techniques has increased substantially. However, designing effective Deepfake detection models requires addressing key challenges such as overfitting, in which models memorize noisy or complex patterns, and underfitting, in which insufficient learning leads to poor predictive performance [27]. Both issues negatively affect model generalization [28]. To mitigate these challenges, researchers aim to balance the bias–variance tradeoff using strategies such as early stopping, pruning, regularization, ensembling, and data augmentation [23, 29].

Despite the growing body of research on Deepfake detection, many approaches continue to face fundamental limitations, including data scarcity, class imbalance, overfitting, and weak generalization to unseen manipulation techniques and real-world conditions. Although architectural innovations and increasingly complex models have been extensively explored, comparatively less attention has been devoted to data-centric strategies that directly influence model robustness. In this context, facial data augmentation has emerged as a critical yet underexamined component of Deepfake detection pipelines. Augmentation provides practical mechanisms to increase data diversity, reduce

overfitting, and improve cross-dataset performance without increasing model complexity.

Moreover, Deepfake detection models require large and diverse datasets. However, acquiring sufficiently varied real-world data is often constrained by data silos, legal restrictions, and other practical barriers. These challenges further justify a focused and systematic examination of augmentation techniques rather than treating them merely as secondary preprocessing steps. Data augmentation, which involves generating additional samples by introducing controlled variations to existing datasets, helps address these limitations. Such techniques can improve model performance, particularly in terms of accuracy and Area under the Curve (AUC), by enhancing generalization and effectively balancing bias and variance [30, 31].

Since 2019, numerous studies on Deepfake detection have been published [30–39]; however, relatively few have specifically examined the role of facial data augmentation techniques. This article therefore provides a focused and comprehensive overview of these methods, offering both theoretical and practical insights to strengthen understanding and support the development of more robust Deepfake detection systems. The primary contributions of this literature review are as follows:

- **Comprehensive Review of Augmentation Techniques (as preprocessing) in Deepfake Detection:** This article presents an analysis of augmentation techniques employed in Deepfake detection as preprocessing strategies, highlighting their applications and reported effectiveness.
- **Summary of Widely Used Deepfake Detection Techniques:** The study systematically reviews prominent Deepfake detection approaches,

summarizing their associated datasets and performance metrics such as accuracy and AUC, thereby guiding researchers in identifying suitable methodologies.

- **Identifying Challenges and Open Issues:** This work examines key challenges and unresolved issues in Deepfake detection, particularly in relation to data augmentation, and provides insights to support future advancements in the field.

The remainder of this paper is organized as follows. Section II presents the research methodology, detailing the search process, search strategy, quality assessment criteria, reporting guidelines, and the inclusion and exclusion criteria applied to the reviewed studies. Section III analyzes visual-based Deepfake detection techniques that incorporate data augmentation, summarizing associated datasets and reported performance metrics such as accuracy and AUC. Section IV discusses the key challenges and open issues related to facial data augmentation in Deepfake detection. Finally, Section V concludes the paper by summarizing the main findings and outlining their implications.

II. METHODOLOGY

This study adopts a systematic review methodology to collect, analyze, and synthesize research literature on Deepfake detection using facial data augmentation. The review process was structured into three main stages: planning, conducting, and reporting. The methodological procedures followed in each stage are described in detail below.

A. Planning Review

The planning stage focused on establishing the need for a comprehensive review of Deepfake detection techniques. This phase was guided by the study's primary objectives: reviewing augmentation techniques in Deepfake detection, analyzing existing detection approaches, and identifying the challenges associated with their implementation.

To achieve these objectives, the study addressed the following Research Questions (RQ):

- **RQ1:** What augmentation techniques are most commonly used in Deepfake detection?
- **RQ2:** How do these techniques affect model performance and generalization?
- **RQ3:** What datasets and evaluation metrics are typically employed?
- **RQ4:** What challenges are commonly associated with augmentation strategies?
- **RQ5:** What emerging research directions are shaping this domain?

These research questions formed the foundation of the systematic review and guided the search strategy, inclusion and exclusion criteria, and analysis of the selected literature.

B. Search Strategy

The search strategy guided the identification and selection of relevant studies using predefined keywords and criteria. The objective of this stage was to collect reliable and relevant literature aligned with the study's research contributions. Indexes and keywords were defined during this phase, as presented in Table I. The search strategy was structured into three components: population, methodology or technique, and outcomes.

TABLE I. SEARCH STRATEGY FOR RELEVANT RESEARCH STUDY

Area	Keywords	Synonyms
Population	Deepfake	Videos Deepfake OR Images Deepfake
Methodology	Deepfake Detection Techniques	Machine Learning OR Deep Learning OR Based on Facial Data Augmentation Mechanisms
Outcomes	Deepfake Detection Applications	Facial Augmentation to Deepfake Detection OR Fake Face Detection OR Facial Forgery Detection OR Face Swapping Detection OR Face Edit Attributes Detection OR Face Reenactment Detection OR Face Synthesis Detection OR Face Forensics OR Facial Transformation OR Facial Noise Detection OR Facial Manipulation Detection

1) Population

The research population was clearly defined prior to initiating the review process. "Deepfake" served as the primary keyword of this study. This review focuses on scenarios involving facial augmentation techniques within the context of recent advancements in Deepfake generation and detection technologies.

2) Methodology or technique

This research article examines various facial data augmentation techniques applied to image- and video-based Deepfake detection. Multiple mechanisms were considered to extract and analyze relevant information from the selected studies. In addition, keywords such as "Deepfake Detection", "Machine Learning", "Deep Learning", and "Facial Data Augmentation" were used to define the methodological scope. The empirical evidence extracted from the literature

was based on: (1) classifiers and techniques employed in Deepfake detection models, (2) datasets used for model training and evaluation, (3) facial augmentation strategies aimed at improving detection performance, and (4) reported performance metrics, particularly accuracy and AUC, associated with augmentation techniques.

Accuracy ranges from 0% to 100% are considered in the studies, and the formula for accuracy is:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Prediction} \times 100$$

- 100% = Perfect classification
- 50% = Random guess (binary case)

The AUC is commonly used to evaluate the performance of binary classification models. An AUC value of 50% indicates random classification. Values between 60%–70% represent poor performance, whereas

70%–80% indicates acceptable discrimination ability. An AUC range of 80%–90% is considered good, while values between 90%–100% demonstrate excellent classification capability.

3) Outcomes

The study outcomes were primarily influenced by the selected data collection and evaluation strategies, which directly shaped the analysis and conclusions. The following keywords were used to capture relevant application domains: “Facial Augmentation to Deepfake Detection”, “Fake Face Detection”, “Facial Forgery Detection”, “Face Swapping Detection”, “Face Reenactment Detection”, “Face Synthesis Detection”, “Face Edit Attributes Detection”, “Face Forensics”,

“Facial Transformation”, “Facial Noise Detection”, and “Facial Manipulation Detection”.

To retrieve relevant literature, combinations of search queries were applied, including: (Videos Deepfake OR Images Deepfake OR Deepfake) AND (Machine Learning OR Deep Learning OR Facial Data Augmentation Mechanisms) AND (Facial Augmentation to Deepfake Detection OR Fake Face Detection OR Facial Forgery Detection OR Face Swapping Detection OR Face Edit Attributes Detection OR Face Reenactment Detection OR Face Synthesis Detection OR Face Forensics OR Facial Transformation OR Facial Noise Detection OR Facial Manipulation Detection). After applying the above search queries, 130 relevant studies were selected. The screening process is illustrated in Fig. 2.

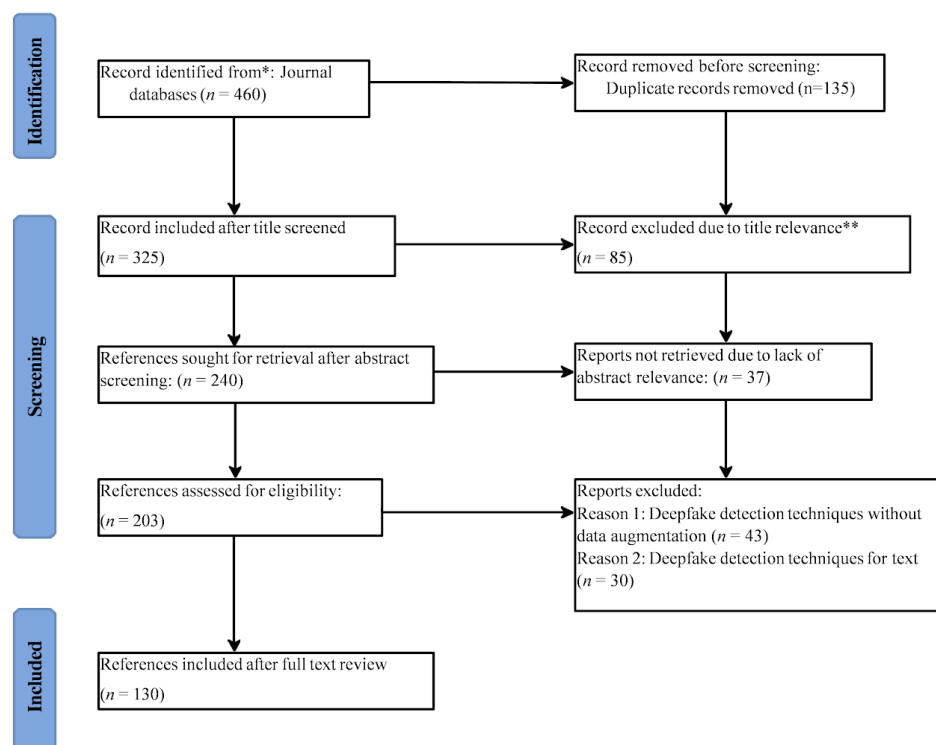


Fig. 2. PRISMA flow diagram for study collection.

We utilized digital data repositories covering publications from 2019 to 2025, as illustrated in Fig. 3. To support the empirical analysis, visualization tools such as histograms, pie charts, and tables were employed to present the extracted research data more clearly. A total of 460 records were initially identified from journal databases. After removing 135 duplicate records, 325 studies remained for title screening. During this stage, 85 records were excluded due to lack of relevance in their titles. Subsequently, 240 studies were considered for abstract screening, of which 37 reports were not retrieved due to insufficient abstract relevance. The remaining 203 studies were assessed for eligibility through full-text review. During this stage, 73 reports were excluded, including 43 studies focusing on deepfake detection techniques without data augmentation and 30 studies related to deepfake detection for text. Ultimately, 130

studies were included in the final review as shown in Fig. 2.

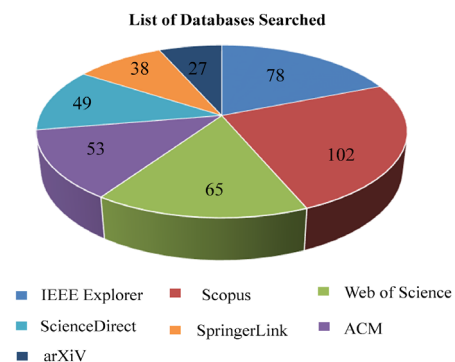


Fig. 3. List of databases searched for Deepfake detection using facial data augmentation.

C. Quality Assessment Criteria

In terms of the study selection criteria, we focused on clearly defined inclusion and exclusion parameters, including language, digital repositories, publication type (e.g., journals, magazines, newsletters, and technical reports), author credibility and authenticity, and theses from Doctor of Philosophy (PhD) or other degree programs that met the defined requirements, with particular emphasis on the specified time frame. The structured application of these selection criteria ensured the retrieval of relevant, credible, and high-quality research literature. Initially, we considered only publications from 2019 onward, as Deepfake detection techniques based on data augmentation began gaining significant research attention during this period. We further ensured that the selected studies were directly relevant to the manuscript's title and scope, free from bias, and evaluated using clearly defined selection parameters, as illustrated in Fig. 2.

In conclusion, the following inclusion criteria, listed in priority order, were applied to select the literature:

- (1) The predefined keywords were required to appear in the title, abstract, or keywords.
- (2) Studies published in 2019 or later.
- (3) Research focusing on Deepfake detection techniques based on facial data augmentation.
- (4) Studies reporting quantitative performance metrics.
- (5) Studies exclusively addressing visual Deepfake detection techniques.
- (6) Articles published in the English language only.

The following exclusion criteria were applied to remove irrelevant studies:

- (1) Patents, PhD theses, newsletters, master's project

documents, and technical or undergraduate reports were excluded.

- (2) Studies that did not clearly specify augmentation techniques or lacked empirical validation were excluded.

- (3) Audio- and text-based Deepfake studies were excluded.

Due to the heterogeneity of datasets, augmentation configurations, and experimental protocols across studies, numerical quality weighting or meta-analytic scoring was not applied. Instead, study quality was ensured through predefined inclusion criteria and empirical validation requirements.

D. Reporting Review

The reviewed literature highlights the importance of facial data augmentation techniques in improving Deepfake detection performance. Fig. 3 illustrates the yearly distribution of related publications across various journals, while Fig. 2 presents the literature selection process. The reviewed studies employ heterogeneous datasets, augmentation configurations, and evaluation protocols: statistical aggregation or meta-analysis was not feasible; therefore, augmentation effects were synthesized descriptively through categorical grouping and cross-study comparison. Furthermore, Fig. 4 identifies the most commonly employed facial augmentation methods. The key methodological categories are summarized in the corresponding tables: Geometric Transformations (Table II), Random Cutout or Dropout and Mixing Augmentations (Table III), Photometric Transformations (Table IV), Python Libraries (Table V), Occlusion Techniques (Table VI), and Random Noise (Table VII).

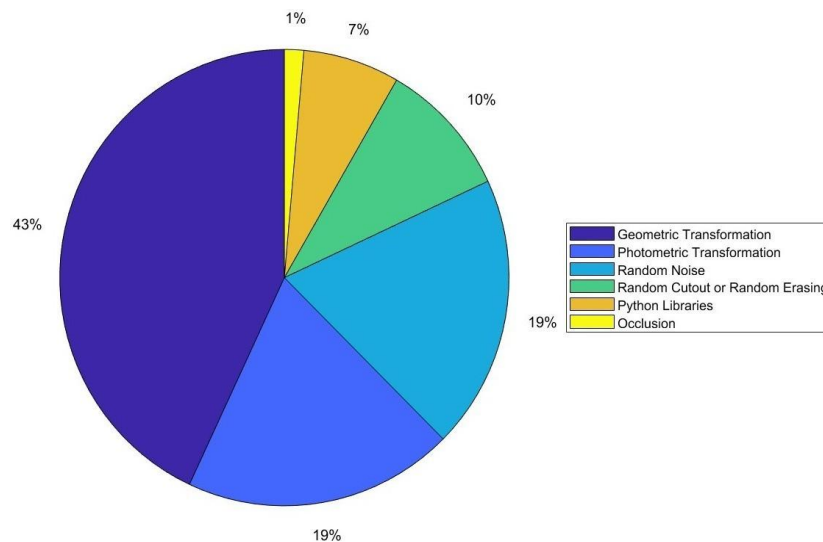


Fig. 4. Most commonly used facial augmentation techniques.

III. AN OVERVIEW OF FACIAL DATA AUGMENTATION FOR DEEPAKE DETECTION

The accuracy of Deepfake detection techniques largely depends on the quality and diversity of the training datasets. Since many training datasets contain limited variations, model performance may be adversely affected.

To improve detection accuracy, training datasets are typically processed using one or more augmentation techniques. Image preprocessing refers to the transformation of raw image data into a structured format suitable for further processing and analysis [40]. It helps eliminate unwanted distortions and enhances specific features required for computer vision applications. In

Deepfake detection, image preprocessing is employed to extract relevant frames from videos or images while discarding irrelevant raw data. During this stage, augmentation techniques are applied to facial image data prior to the training process. The primary objective of image preprocessing is to improve generalization, prevent

overfitting and underfitting, and ensure robust performance on unseen data through the application of augmentations. Facial augmentation techniques are commonly applied using facial landmarks, as facial features and orientations can be accurately detected. The facial augmentation techniques are illustrated in Fig. 5.

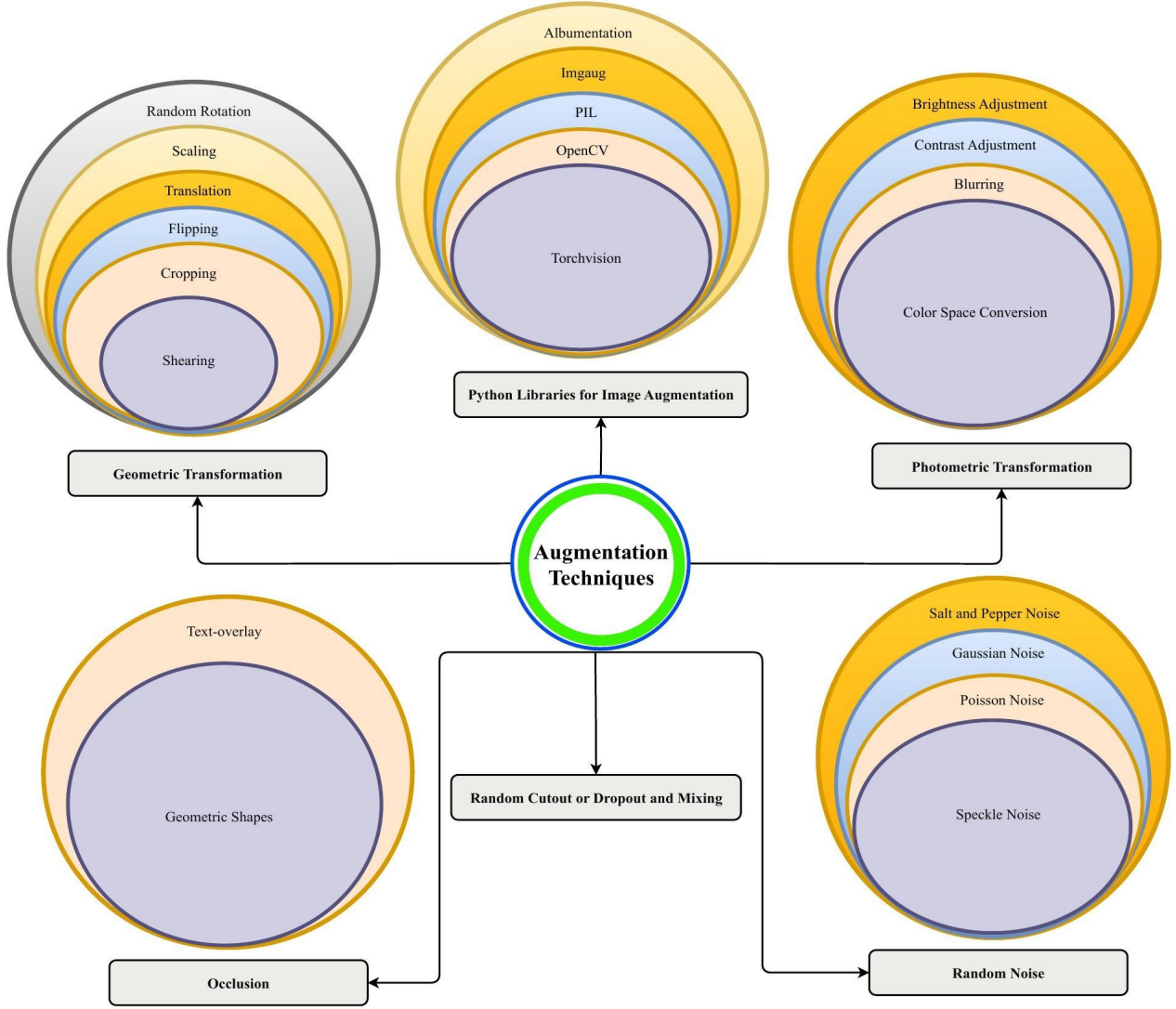


Fig. 5. Facial augmentation techniques.

A. Geometric Transformation

Geometric transformation is a type of image data augmentation technique that modifies the geometric structure of images by relocating pixels from their original positions to new coordinates without altering their intensity values [26, 41]. Geometric transformation involves two main steps [42, 43]: spatial transformation and gray-level interpolation. The spatial transformation changes pixel coordinates while preserving their intensity values. It is defined by the following formula.

$$x = r(x, y)$$

$$y' = s(x, y)$$

where (x, y) are the pixel coordinates of the geometrically distorted image are used to extract the original image, and r and s are functions that depend upon the values of x and y . A Gray level interpolation defines an image into Gray level to the transformed image. A common bilinear interpolation technique uses the following scheme.

$$v(x, y) = c_1x + c_2x + c_3xy + c_4$$

where the values of $v(x, y)$ are the Gray values depending upon the Gray values of the pixel at the location of (x, y) , and c_1, c_2, c_3 , and c_4 are the coefficients, which are the surrounding pixel values to generate the grayscale image.

Non-affine and affine transformations are the two primary categories of geometric transformations. Non-affine transformations modify an image's geometric

properties without preserving straight lines, thereby significantly altering its overall appearance. Examples of non-affine transformations include projective (perspective), warping, Bézier, polynomial, radial distortion, and nonlinear scaling transformations [44, 45]. These transformations are widely used in fields such as weather forecasting and medical imaging, where new images are generated based on mapping functions derived from the original image. However, non-affine transformations have not been extensively explored in the context of facial data augmentation for Deepfake detection. The second category is affine transformation, which preserves linear relationships (e.g., points lying on a line remain on that line after transformation) and maintains relative distance ratios [46, 47]. In this context, the term “affine” indicates that geometric relationships are preserved within the same plane.

1) Random rotation

In this technique, the original image is randomly rotated by a specified number of degrees, either clockwise or counterclockwise, to alter its orientation within the frame. This changes the angles at which objects appear during training. For example, if the image acquisition process captures primarily horizontally aligned objects, rotation can simulate variations in orientation that may occur in real-world scenarios. This augmentation method enhances model performance without requiring additional data collection or annotation [48, 49].

2) Scaling

Image pixels represent intensity values [50]. Scaling commonly refers to normalizing pixel values to a predefined range (e.g., 0 to 1 or -1 to 1), which stabilizes training and prevents large numerical values from dominating computations. Resizing ensures that all images in a dataset have consistent dimensions, thereby simplifying model training and reducing computational complexity [13, 51]. This consistency enables the model to identify patterns reliably across multiple samples.

The primary objective of resizing is to adjust the input image size while preserving salient regions and minimizing visual distortion. Many resizing approaches begin with saliency detection to identify critical areas of the image. An appropriate resizing operator is then applied to generate the resized image while maintaining the integrity of these significant regions and introducing minimal artifacts.

3) Translation

In image processing, translation refers to shifting image pixels along the x- or y-axis without altering their size, shape, or orientation [52]. Spatial translation is commonly used for data augmentation, where an image is shifted vertically, horizontally, or in both directions without modifying its content. The purpose of this operation is to simulate natural variations in object positioning and camera alignment that commonly occur in real-world scenarios.

In contrast, image-to-image translation generates a new image in a different style or domain, whereas spatial translation repositions the original image content. By

shifting the facial region by a few pixels, the model becomes more robust to misalignment, off-center faces, and slight camera movements. This augmentation is particularly valuable in Deepfake detection, where manipulated videos often exhibit inconsistent facial positioning or jitter caused by generation artifacts.

There are two main types of image translation: supervised and unsupervised. Supervised translation involves learning conditional mappings from paired samples drawn from a joint probability distribution [53]. In contrast, unsupervised translation learns conditional mappings using samples from marginal distributions across two or more unrelated domains [54].

4) Flipping

Flipping refers to reversing the orientation of an image either horizontally or vertically, as illustrated in Fig. 6. In a vertical flip, the image is inverted along the vertical axis; in a horizontal flip, the image is mirrored along the horizontal axis [55]. The input images are transformed into their mirrored counterparts, thereby increasing dataset diversity [56].

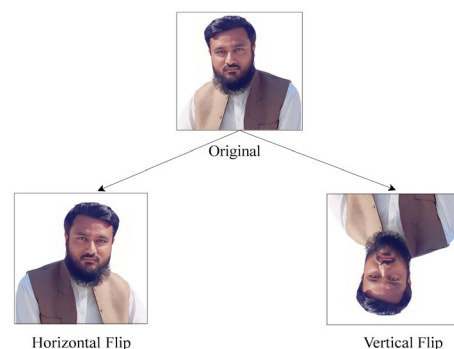


Fig. 6. Flipping.

5) Cropping

Image cropping is frequently applied to emphasize specific regions of an image while removing distracting elements. By eliminating irrelevant content, cropping enables the model to focus on targeted regions, as shown in Fig. 7. This improves the model’s ability to identify relevant features, particularly in tasks where object localization is critical [57]. Cropping is essentially an internal operation, as the resulting image is always a subset of the original frame.

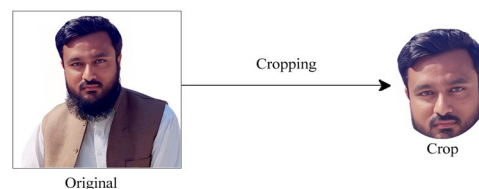


Fig. 7. Cropping.

Several automated image cropping techniques, such as attention-based and aesthetics-based methods, are commonly employed [58]. Cropping can simulate zooming or focusing on different parts of an image. In Deepfake detection, specific facial regions such as the

nose, eyes, lips, and cheeks are often cropped and recombined into a single frame to enhance the identification of manipulated content.

6) *Shearing*

Shearing skews an image along the horizontal or vertical axis. It is typically applied within a small range (e.g., 0.2), introducing slight geometric distortion. This approach improves model robustness against perspective variations and distortions commonly encountered in real-world conditions [59].

Khan *et al.* [60] restricted their augmentation strategies to random cropping and horizontal flipping during the training phase on the FaceForensics++ (FF++) dataset. Distortions used during training and testing were carefully selected to avoid potential interactions. The results demonstrated that the proposed method outperformed both frame-based and sequence-based approaches when common distortions were applied. Furthermore, the model's robustness was significantly enhanced by incorporating adversarial training as a regularization component within the loss function.

Iqbal *et al.* [61] discussed various augmentation techniques that are crucial for improving the effectiveness of deep learning models for Deepfake detection. Images were typically resized to 256×256 pixels to ensure input size consistency, which is essential for training deep learning models; a similar approach was adopted in another study [62].

A rotation range of 25° was applied to improve the model's ability to generalize to synthesized data and recognize images from different perspectives. Images were also transformed using a shear range of 0.2, enabling the model to become invariant to distortions that may occur in real-world scenarios. Additionally, a zoom range of 0.2 was applied to allow the model to learn from both zoomed-in and zoomed-out images, thereby improving recognition across different scales. Width and height shift ranges (both within 0.1) were employed to help the model learn from slightly displaced images, increasing robustness to positional variations. Another step in the augmentation process involved normalizing pixel values to a predefined range (e.g., 0 to 1), which stabilizes training and enhances performance. One or more augmentation techniques were combined to create a more diverse training dataset, thereby improving the model's generalization ability and accuracy in identifying Deepfake images.

The study conducted by Nadimpalli and Rattani [63] emphasized the importance of selecting optimal augmentation strategies for Deepfake detection. An ablation study was performed to determine the ideal number of augmentations (top-k) that yield the most significant performance improvements. The results indicated that, when applied to the FF++ dataset, the top three augmentations achieved the best performance metrics, with an AUC of 99.4% and an Equal Error Rate (EER) of 0.022. Applying multiple transformations to training images improved generalization and reduced the risk of model overfitting.

The challenge of limited key-frames for training was

addressed by augmenting key-frame data using techniques such as rotation, translation, and both horizontal and vertical flipping. Although key-frames are highly detailed, they are often limited in number. Naz *et al.* [14] and Wang *et al.* [64] demonstrated that augmentation helped create a more balanced dataset, enabling effective model training. However, they also highlighted a potential vulnerability: attackers could enhance the realism of key-frames, potentially reducing the model's ability to detect manipulated content if such frames are intentionally modified. Patel *et al.* [59] further implemented a zoom factor of 0.2, enabling the model to train on images at various scales, which is crucial for recognizing Deepfakes of different sizes. Additionally, width and height shift ranges of 0.2 were applied, allowing the model to learn from slightly off-center images, a common occurrence in real-world data. Random rotations of up to 360° were also utilized, enabling the model to detect features regardless of orientation and improving generalization. Overall, these augmentations strengthened the model's robustness, enabling reliable detection across diverse conditions.

To reduce computational cost without compromising efficiency, Sharma *et al.* [65] proposed a novel approach for extracting counterfeit features, specifically targeting face-swap images generated using deep learning techniques such as Deepfake. Their study integrated deep learning with Error Level Analysis (ELA) techniques and various augmentation strategies, including horizontal and vertical flipping.

This approach enhanced the training effectiveness of Convolutional Neural Networks (CNN) models by identifying differences in image compression artifacts between real and manipulated regions, a key characteristic of face-swapped images. The proposed technique achieved approximately a 90% reduction in computational cost without degrading accuracy.

Khormali and Yuan [66] employed two primary data augmentation techniques to enhance Deepfake detection. The face-scaled enhancement method emphasized enlarging specific facial regions, such as the lips, forehead, and eyes, enabling the model to focus on forgery-prone areas while capturing fine-grained details. Conversely, the face shut-off enhancement method obscured key facial regions, compelling the model to analyze alternative features and thereby extract discriminative information. Attention-based augmentations proved more effective than random augmentations, as they reduced background noise and enhanced the model's feature learning capability.

To improve flexibility and generalization capability in the Noise-Augmented-Visual Geometry Group (NA-VGG) detection model, Chang *et al.* [67] proposed a novel framework that incorporates both noise extraction and augmentation strategies. Specifically, a Spatial Rich Model (SRM) filter layer was used to extract noise features from the original Red, Green, Blue (RGB) images. The resulting noise maps were then subjected to additional augmentation processes. By reducing dependence on facial features and emphasizing tampering artifacts that may not be visible in the conventional RGB domain,

augmentation was applied directly to the noise maps. Through diversified augmentation of training data, the model was exposed to a broader range of noise patterns and possible manipulations, enabling it to learn more generalized and discriminative features for distinguishing between real and fake images. This approach is particularly important because modern Deepfake images exhibit high visual fidelity and often do not display obvious artifacts in the RGB domain.

Gram-Net, a novel architecture proposed in Ref. [68], used ResNet-18 as the baseline model and was enhanced with a manually designed texture descriptor known as Gray Level Co-occurrence Matrix (GLCM) for the RGB channels. Images were randomly resized between 64×64 and 256×256 pixels, and data augmentation was applied to improve generalization. These random bilinear resizing strategies helped evaluate the model's robustness.

Ding *et al.* [69] employed deep transfer learning to achieve high detection rates for face-swapped images,

reporting minimal errors and true positive rates above 96%. This success was supported by the construction of a large-scale dataset, where each participant contributed approximately 1000 genuine images, facilitating effective model training. The authors also proposed a framework to assess human recognition capability by developing a website for collecting pairwise image comparisons. This approach helped evaluate prediction uncertainty, which is critical for estimating system confidence. Additionally, test-time data augmentation, including random cropping and horizontal flipping, was utilized to quantify data-dependent variance. An active learning strategy was further applied to identify informative image pairs for comparison. This approach encouraged the model to learn generalized representations rather than memorizing training samples, thereby improving its ability to distinguish between real and fake images.

The common types of geometric transformations are outlined below and detailed in Table II.

TABLE II. SUMMARY OF DEEFAKE DETECTION TECHNIQUES BASED ON GEOMETRIC TRANSFORMATION AUGMENTATIONS

Ref.	Classifier/Techniques	Augmentation Techniques	Dataset	Accuracy (%)	AUC (%)
[60]	Adversarial Feature Similarity Learning (AFSL)	Flipping and cropping	FF++, FaceShifter and DeeperForensics	-	89.7
[70]	Vision Transformers (ViT)	Random resizing and cropping	FF++, FaceShifter, FaceShifter (FSh), Deep-erForensics (DFo), WildDeepfake (WDF), and Deepfake Detection Challenges (DFDC)	-	92.3
[61]	VGG16, InceptionV3, and VGG19	Rotation, flipping, and scaling	190,335 RGB images	90	-
[71]	D3 (Discrepancy Deepfake Detector)	Rotation and flipping	UniversalFakeDetect (UFD)	97.6	-
[64]	Convolutional Pooling Transformer	Rotation and flipping	FF++, DFDC, celebrities-Deepfake (Celeb-DF), and DF-1.0	92.11	97.66
[59]	Deep-CNN	Flipping, Scaling, and shear	CelebA, and Flickr Faces High Quality (FFHQ)	99.33	-
[65]	CNN	Flipping and rotating	140k Real/Fake Faces Dataset	98.79	-
[66]	VGGNet, Residual Network Net (ResNet), Xception, and MobileNet	Scaling	Celeb-DF-V2, and DFDC	98.30	-
[72]	Multiscale CNN, and Deep Neural Network (DNN)	Rotation and shearing	Kaggle Deepfake Detection Challenge	85.8	-
[73]	ResNetV2, Xception, and ShallowNetV3	Zoom, Cutout, and rotation	CelebA, PGGAN, and FaceForensics	90.29	93.99
[68]	ResNet-18	Scaling	CelebA-HQ, StyleGAN, and PGGAN	99.76	-
[69]	CNNs	Rotation and flipping	420k images (real + swapped)	99.9	-
[74]	Multi Discriminator Deep (MDP)-CNN	Flipping, cropping, and rotating	70k StyleGAN + 70k real FFHQ	93.63	-
[63]	CNN	Rotate and translate	FF++, Celeb-DF, and DeeperForensics-1.0	-	99.7
[67]	VGG based CNN	Horizontal and vertical flipping	FF++	-	85.7

B. Random Cutout or Dropout and Mixing

Cutout is a regularization technique that randomly masks square or circular regions of input images during training, as illustrated in Fig. 8. This method generates partially occluded versions of the original images, effectively augmenting the dataset with additional samples, as demonstrated in Table III. The primary objective of Cutout is to improve accuracy and robustness. By simulating occlusions, it encourages the model to consider the entire image context rather than relying solely on a limited set of discriminative visual cues [75]. The size of the occluded regions is also randomized.

CutMix is a local-level augmentation technique that improves generalization and reduces overfitting. It

replaces a rectangular region of one image with a corresponding patch from another image, generating new training samples. Unlike Random Erasing, where the masked region is filled with random values, CutMix inserts semantically meaningful content from another image. This encourages the model to learn from multiple regions across different samples and to focus on distributed discriminative features rather than specific localized patterns. In Deepfake detection, CutMix reduces reliance on isolated forgery artifacts and promotes learning of more holistic facial representations.

MixUp is another image augmentation technique that blends two different images and their corresponding labels to generate synthetic training samples centered around original data distributions. Unlike CutMix, which replaces

a distinct image patch, MixUp combines the entire pixel space of two images, producing a globally interpolated representation. This method improves robustness, reduces overfitting, and mitigates limitations caused by insufficient training data. In Deepfake detection, MixUp enhances generalization across both subtle and high-frequency facial manipulations by providing more diverse training examples.

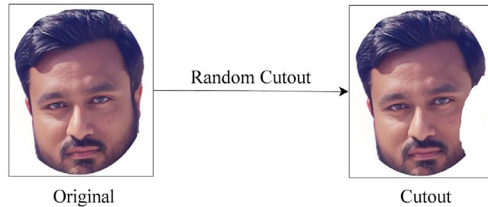


Fig. 8. Random cutout or dropout and mixing.

Omar *et al.* [76] focused on augmentation techniques, particularly emphasizing the CutMix method. This approach increased the model's exposure to diverse

samples, thereby enhancing generalization. The proportion of pixels contributed by each image was controlled by a coefficient sampled from a Beta distribution, and a binary mask determined which regions were retained or replaced. These strategies collectively improved model flexibility and adaptability to novel data. Notably, the study revealed that integrating CutMix significantly improved performance, as the Convolution and Attention Network (CoAtNet) model augmented with CutMix outperformed both the conventionally augmented CoAtNet model and its non-augmented counterpart. The study also explored aggregation strategies such as Sum, Product, and Majority Voting to combine predictions from multiple models. Among these, Majority Voting selected the most frequent class prediction. This ensemble strategy improved detection robustness and performance across datasets. The integration of CutMix with the CoAtNet architecture achieved high accuracy in detecting manipulated faces across multiple benchmarks, highlighting the substantial impact of augmentation strategies on model performance.

TABLE III. SUMMARY OF DEEFAKE DETECTION TECHNIQUES BASED ON RANDOM CUTOUT OR DROPOUT AND MIXING AUGMENTATIONS

Ref.	Classifier/Techniques	Augmentation Techniques	Dataset	Accuracy	AUC
[76]	CoAtNet	CutMix and Dropout	Celeb-DF and FF++	98.90%	99.70%
[77]	EfficientNet and Xception	Cutout and Patchup	DFDC, Deepfake (DF)-1.0, and Celeb-DF	-	99.82%
[78]	MesoNet, Xception, and Efficient-Net	Mixing	DFDC and Celeb-DF	74.20%	-
[79]	Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and 3DCNN	Temporal dropout	Celeb-DF-v2, DFDC, and FF++	82.64%	88.83%
[80]	Multiscale Convolutional Neural Network (MTCNN)	Dropout	DFDC and Celeb-DF-v2	-	97.80%
[81]	Deep Distribution Transfer (DDT)	Spatial augmentation and mixup	FF++ and Celeb-DF	90.01%	-
[82]	MCDM and GANs	Random Erasing	FF++ and Deepfake Detection (DFD)	-	99.31%
[83]	CNNs and VITs	Patch embedding	DFDC	91.50%	91.00%
[84]	VIT	Random cutout and random erasing	DFDC	-	97.80%
[85]	VIT	Vulnerability-driven Cutout	FF++ and Celeb-DF	85.30%	-

Cheng *et al.* [77] applied augmentation strategies to mitigate overfitting in Deepfake detection models. These included CutMix, which combines two images to produce augmented samples and encourages learning from mixed features. Additional techniques included DropBlock, which removes contiguous regions of feature maps during training to improve robustness; PatchUp, which regularizes feature blocks from random sample pairs; and Feature Statistics Mixing Regularization (FSMR), which ensures invariance to stylistic variations in input images. Label Smoothing was applied to prevent overconfidence in predictions, while Primary Region Regularization (PRR) enhanced generalization by removing dominant regions identified by the model during training.

Bao *et al.* [78] addressed the challenge of insufficient individualized fake video data for training Deepfake detectors. Their work focused on generating fake videos closely resembling specific individuals. Mixing synthesized fake videos with real videos of the same individuals enhanced the model's ability to differentiate between authentic and manipulated content. The study

demonstrated that incorporating identity-specific fake samples improved cross-dataset generalization. Autoencoders were employed to generate realistic fake videos. However, the authors acknowledged that models trained on dataset-specific manipulations may still suffer from overfitting and reduced performance when exposed to unseen manipulation techniques.

Zhang *et al.* [79] introduced a Temporal Dropout (TD) module integrated into a Three-Dimensional Convolutional Neural Network (3DCNN) architecture, enabling the sampling of diverse frame sequences from the same video during training. This enhancement allowed the model to better capture both local temporal dependencies and global sequence-level information. The TD module functioned as a form of data augmentation by exposing the classifier to varied frame sequences in each iteration, thereby reducing overfitting and improving generalization. This approach is particularly important for Deepfake video detection, where high photorealism makes manipulation difficult to identify.

Tran *et al.* [80] focused on key facial regions, including

the lips, nose, and eyes, to enhance discriminative learning between real and manipulated videos. This method involved randomly dropping significant facial regions during training, thereby increasing data heterogeneity and reducing overfitting. As a result, the model demonstrated improved generalization and detection performance.

A spatial MixUp augmentation strategy proposed by Aneja *et al.* [81] combined images from source and target domains to enhance cross-dataset and cross-manipulation generalization. This method improved performance not only on the same manipulation type but also on previously unseen forgery techniques. Fine-tuning with spatial augmentation resulted in measurable improvements, including a 3.81% increase in binary classification accuracy in the Neural Textures (NT)-to-Deepfake transfer scenario.

Chen *et al.* [82] introduced a Masked Conditional Diffusion Model (MCDM) framework for data augmentation in Deepfake detection. The model manipulated original images while preserving intrinsic semantic structures during reconstruction. A mask-based strategy was used to modify facial regions, ensuring reconstructed outputs remained realistic while promoting diversity. Both pixel-level and feature-level reconstruction losses were employed to generate high-quality forged samples. Experimental results demonstrated that the synthesized images significantly improved model robustness and generalization.

Wodajo and Atnafu [83] proposed a patch-embedding approach based on a hybrid architecture combining CNN and ViT features for Deepfake video detection. By representing inputs through complementary feature embeddings, the model improved generalization on unseen data. The method achieved an AUC of 91.50% and an F1-Score of 91.9%, surpassing previous state-of-the-art approaches.

Heo *et al.* [84] proposed a ViT-based model that processes video frames by dividing them into patches and mapping each patch to a feature vector for efficient spatial representation learning. To enhance performance, EfficientNet was incorporated at the patch-embedding stage to leverage its strong feature extraction capabilities. A knowledge distillation strategy transferred information from EfficientNet to the transformer-based ViT model, improving accuracy while reducing false negatives. Additionally, random cutout and random erasing were introduced to enhance generalization. When evaluated on the DFDC dataset, the model achieved an F1-Score of 91.9% and an AUC of 97.80%. The patch-based framework enabled focused attention on critical frame regions, and confusion matrix analysis demonstrated improved classification stability and reduced false negatives compared to earlier models.

C. Photometric Transformation

Photometric transformations are based on color manipulation, where the original data are modified through brightness and contrast adjustments, as well as blurring. These techniques are useful for data augmentation and have been applied in Deepfake detection, as shown in Table IV.

1) Brightness adjustment

Brightness refers to the perceived intensity of light emitted or reflected by an object. It plays a critical role in image processing, as it affects how effectively features can be detected within an image. Adjusting brightness can improve visibility and performance in image analysis tasks, including Deepfake detection [86]. Reducing brightness may conceal features in highlighted regions, whereas increasing brightness can reveal intricate details in shadowed areas. In real-world scenarios, videos often undergo brightness adjustments, which may introduce distinct visual artifacts [87].

2) Contrast adjustment

Adjusting contrast in images enhances the visibility of important features. This is particularly significant in Deepfake detection, where even subtle differences between real and manipulated images are essential for accurate identification. Increasing contrast enables models to better distinguish authentic from fake content by emphasizing facial features. It sharpens the boundaries between key facial components such as the nose, mouth, and eyes [88]. Deepfake generation models often introduce artifacts along these edges, which become more pronounced after contrast enhancement. By modifying contrast, subtle inconsistencies in lighting and color between genuine and manipulated regions become more apparent. Many Deepfake models struggle to generate realistic skin textures under varying lighting conditions, and these textural irregularities become more visible when contrast is increased. Additionally, Deepfake videos often exhibit color, lighting, or shadow inconsistencies across frames, resulting in unnatural flickering or abrupt illumination changes. These artifacts become more detectable when contrast adjustments are applied across frame sequences [89].

3) Blurring

Inconsistencies in facial artifacts, texture, and motion patterns often indicate manipulated content and can be highlighted through blurring, which is helpful in Deepfake detection. When blurring is applied to an image, certain fine details are smoothed, thereby exposing structural flaws introduced during the Deepfake generation process, such as irregular blending at facial boundaries. These inconsistencies may be less noticeable in sharp images but become more apparent after blurring [90]. Gaussian blurring reduces high-frequency details, minimizing variations in texture and alignment [91]. This technique highlights anomalies in pixel distribution and color blending, thereby enhancing the detection of identity manipulation and face swapping in images.

4) Color space conversion

Color representation is fundamentally based on color spaces, which are essential concepts in computer vision and image processing. Color spaces provide a standardized framework for representing digital image colors, facilitating efficient processing and analysis. In image processing, colors are numerically represented and manipulated using color spaces. These spaces are

mathematical models that define the range of colors that can be displayed or printed. All other colors in a given color space are derived from a specific set of primary colors. The most commonly used color spaces in image processing include RGB, Cyan, Magenta, Yellow, and Black (CMYK), and Hue, Saturation, and Lightness (HSL) [92].

Normalization: Pixel values are standardized by normalizing them to a mean of 0 and a standard deviation of 1. This process maintains values within a specified range, facilitating convergence and improving model stability during training. For grayscale images, normalization is applied to a single channel, whereas for RGB images, normalization is applied across all three-color channels, as shown in Fig. 9.

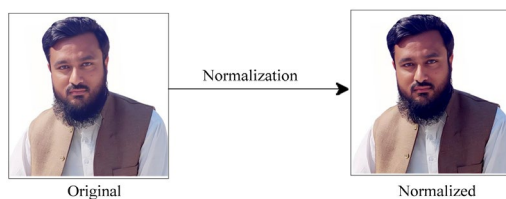


Fig. 9. Normalization.

RGB Color Space: The RGB color space utilizes three primary colors: Red, Green, and Blue, to generate all other colors. In the RGB color space, each color is represented by a combination of RGB values, with each component ranging from 0 to 255.

CMYK Color Space: The CMYK color space follows a subtractive color model commonly used in print media. This model relies on four primary components: Cyan, Magenta, Yellow, and Black, to create all other colors. Each component in the CMYK color space is represented by values ranging from 0 to 100.

HSL Color Space: The HSL color space represents colors using three components: hue, saturation, and lightness. Hue denotes the actual color of the object, saturation indicates the intensity of the color, and lightness refers to the brightness level. In the HSL color model, hue is represented by a value between 0° and 360° , while saturation and lightness are expressed as percentages ranging from 0% to 100%. In image processing, conversion between color spaces is often necessary to improve the efficiency of Deepfake detection methods. Color space conversion is an important preprocessing step that enables techniques to better detect discrepancies in images, thereby enhancing robustness and accuracy in Deepfake detection tasks [93].

Various augmentation techniques were addressed by Wang *et al.* [94] to enhance model accuracy. To simulate more realistic attack scenarios, they applied Boundary Blur Mask Generator (BBMG), which introduces boundary-blurred manipulation traces at both global and local stages. This technique facilitates the generation of pseudo-fake samples that closely resemble real samples, thereby improving model training. The Noise Refinement Strategy (NRS) method reduces the impact of noisy samples, improving generalization by controlling noise during augmentation. They also utilized common

augmentation techniques such as RGB shift, Hue, Saturation, and Value (HSV) transformation, and contrast adjustments. Data augmentation techniques, including blurring and color adjustments [95], were employed during training to improve model generalizability by creating diverse image variations.

Fung *et al.* [96] utilized random color jittering and grayscale conversion. Applying these augmentations led to significant improvements in detection performance, with FF++ dataset results increasing from 52% to 59%, and UADFV dataset performance improving to 98.90%. However, these augmentations resulted in a slight decrease in performance on the Celeb-DF dataset. Various techniques, including color jittering and Gaussian blur, were used to improve detection performance under low-contrast or inconsistent visual conditions [89]. Lu and Luo [97] implemented contrast modifications at the beginning of the augmentation process, each with a 50% probability. These adjustments were non-linear, with a random factor selected from [0.5, 1.5] to replicate diverse lighting conditions. A 50% probability of image blurring was also introduced, employing varying kernel sizes for average or Gaussian blur to simulate visual degradation and enhance model generalization. Similarly, Tariq *et al.* [98] applied brightness adjustments, color channel shifts, and zooming to improve robustness under different real-world conditions.

Lu and Ebrahimi [99] proposed the Stochastic Degradation-Based Augmentation (SDAug) approach to enhance the robustness of Deepfake detectors. This method begins with an image enhancement operation, where either brightness or contrast is randomly adjusted with a 50% probability. Subsequently, non-linear modifications are applied and randomly selected from a predefined range. The results demonstrated that models trained with SDAug achieved significant robustness improvements while maintaining high performance on original high-quality data. The study compared SDAug with traditional augmentation methods such as Augmentation Mixture (Augmix), which applies multiple transformations in parallel. However, Augmix demonstrated only modest robustness improvements and performed inadequately when trained with low-quality data. In contrast, the proposed SDAug approach consistently enhanced detector accuracy across various processed Deepfake videos.

To replicate the range of distortions and alterations that Deepfake videos may encounter in real-world scenarios, the authors implemented multiple augmentations during the training stage, as suggested by Bondi *et al.* [100]. The augmentation process incorporated random modifications in the HSV color space to provide color diversity, along with contrast adjustments to simulate different lighting conditions. Each probabilistic augmentation had a 50% chance of being applied to a given training image. The researchers aimed to increase the model's resilience to subtle manipulations and degradations by exposing it to a broader range of data variations, similar to the strategies discussed by Ali *et al.* [101].

Zhang *et al.* [102] combined deep learning with ELA to

develop a counterfeit feature extraction method for face-swap image detection. Although the primary focus was on ELA and CNNs, considerable emphasis was placed on data augmentation during training. Various post-processing scenarios that Deepfake content may experience were simulated using augmentation techniques, including color jittering, sharpness adjustment, contrast modification, and color distortion. Random values were selected for these transformations to ensure exposure to diverse image variations during training. This approach improved generalization and reduced overfitting by diversifying the dataset. Before processing with the ELA approach, the augmented images were resized to 128×128 pixels. The model achieved a high detection accuracy of 97.6% by integrating comprehensive augmentation strategies, which improved its ability to identify Deepfake images subjected to different resolutions and editing artifacts. Lin *et al.* [103] simulated various lighting conditions by randomly adjusting contrast within a specified range. Gaussian blur, with kernel sizes between 3 and 15, was applied to simulate low-quality or blurred frames. Additionally, brightness was modified nonlinearly through gamma adjustments within the range [0.5, 2.5] to replicate diverse image sources. However, GAN-based augmentation can be computationally expensive due to the cost of generating synthetic data.

The study employed three CNN architectures: XceptionNet, InceptionV3, and VGG16, to determine whether video frames were real or fake [104]. A fusion strategy was used to combine predictions from different

models, improving robustness against adversarial attacks. Image augmentation techniques, including Gaussian blur and contrast adjustments, were used to diversify the training data and reduce overfitting. However, the study acknowledged that Deepfake detection models still struggle to generalize when confronted with unseen manipulation techniques. This indicates that while augmentation supports robustness, it may not entirely resolve generalization challenges. The study also emphasized that adversarial attacks remain a critical issue, potentially degrading the effectiveness of individual techniques within an ensemble.

Li *et al.* [105] proposed data augmentation methods to enhance the robustness of a one-class detection framework for identifying GAN-generated synthetic faces. Gaussian blur and homomorphic filtering were applied to preserve grayscale consistency while enhancing image contrast. Motion blur simulated movement effects, enabling the model to recognize faces that appear blurred due to motion. The combination of these approaches yielded the best experimental results. Furthermore, a weakly supervised augmentation strategy based on attention maps was introduced to focus on relevant image regions. This approach cropped or suppressed less important areas, improving the model's focus on discriminative features. A self-attention layer was incorporated to preserve pixel-level relationships and improve recognition accuracy. A two-step training process was implemented using binary and one-class classification loss functions to selectively filter fake features and refine the model's discriminative capability.

TABLE IV. SUMMARY OF DEEPPFAKE DETECTION TECHNIQUES BASED ON PHOTOMETRIC TRANSFORMATION AUGMENTATIONS

Ref.	Classifier/Techniques	Augmentation Techniques	Dataset	Accuracy	AUC
[89]	Xception based neural networks	Random color jittering, and Gaussian blur	FF++, Celeb-DF-v2, and VGGFace2	99.70%	-
[94]	BBMG and NRS	RGB Shift, HSV, and Random Brightness Contrast	FF++ and Celeb-DF	-	84.50%
[95]	EfficientNet and VIT	Blur and color adjustment	DFDC and FF++	-	95.10%
[96]	Unsupervised contrastive learning	Random color jittering and grayscale conversion	FF++, UADFV, and Celeb-DF	-	98.90%
[97]	Capsule-Forensics and XceptionNet	Brightness or contrast modifications	FF++ and Celeb-DF	99.80%	-
[98]	LSTM and RNN	Brightness adjustment and channel shift	FF++	91.30%	99.50%
[99]	CNNs	Brightness or contrast adjustments	FF++ and Celeb-DF-v2	-	80.67%
[100]	EfficientNet and Xception	Color space conversion brightness and contrast adjustment	Celeb-DF, DF, DFD, and DFDC	-	99.8%
[102]	VGG16, ResNet50, ResNet101, and ResNet152	Color jittering, brightness, contrast, sharpness, and color distortion.	Milborrow University of Cape Town (MUCT)	97.0%	97.60%
[103]	GAN	Contrast and brightness adjustment	FF++	65%	-
[104]	VGG16, InceptionV3, and XceptionNet	Gaussian blur and contrast adjustments	DFDC and DeepfakeTIMIT	91.88%	-
[105]	Multi-Channel Convolutional Neural Network (MCCNN)	Gaussian blur, grayscale and Homomorphic Filter Enhancement	FFHQ and Progressive Growing GAN (ProGAN)	-	99.20%
[106]	Dual-branch CNN with multi-scale fusion	Contrast adjustment and blurring	FF++, Celeb-DF, and DFD	-	94.70%
[107]	Fourier transform and VIT	Contrast enhancement, and blurring	FF++, Celeb-DF, and WildDeepfake	-	98.18%
[108]	CNNs and VITs	Masked Eye Pretraining and Brightness adjustments	Celeb-DF v2	-	97.82%
[109]	CNNs	Affine transformations	UADFV and DFDM	95%	-
[110]	Siamese neural network	Contrast enhancement, and Gaussian blur	FFHQ and StyleGAN2	-	99.30%

D. Python Libraries for Image Augmentation

Python is a programming language widely used for

image processing, and it provides built-in libraries that support various image augmentation techniques, as shown in Table V.

1) *Albumentation*

Albumentation is a fast and versatile library for image augmentation, well suited for tasks requiring extensive image processing. It offers a broad range of augmentation techniques and is optimized for high performance, making it an excellent choice for various image-related tasks. Albumentation is widely integrated with popular deep learning frameworks, including TensorFlow and PyTorch.

The Albumentations library provides comprehensive support for various computer vision tasks, including object detection, image classification, instance segmentation, pose estimation, and semantic segmentation. It offers more than 70 unique augmentation techniques, enabling the efficient creation of diverse and representative training samples from existing datasets. As a result, Albumentation has become a widely adopted tool in industry applications, open-source projects, and research endeavors spanning machine learning and deep learning [111, 112].

2) *Imgaug*

The open-source Python module Imgaug is a powerful tool for image augmentation in machine and deep learning research. It supports a wide range of augmentation methods, including the modification of segmentation maps, heatmaps, bounding boxes, landmarks, and images through an easy-to-use interface. Imgaug is particularly useful when models need to be trained to handle missing data, as it can randomly erase portions of an image. These techniques are optimized for speed and can be applied either randomly or in a controlled manner, depending on the specific task.

For machine and deep learning experiments, Imgaug is an effective image augmentation library. It supports methods such as rotation, cropping, and flipping, and also enables more advanced techniques, including image blurring and the addition of Gaussian noise [113].

3) *Pillow*

A popular Python image processing library [114], Python Imaging Library (PIL) was developed as an extension of the original PIL. It offers strong performance and flexibility across multiple image formats. Pillow is a preferred tool for developers and researchers working on image-related projects, as it supports basic image manipulation, color and contrast adjustments, filtering and effects, text and drawing capabilities, and image transformation for image enhancement purposes.

4) *Open Source Computer Vision Library (OpenCV)*

OpenCV is an extensive open-source library widely used for computer vision, machine learning, and image processing applications [115]. It is a powerful tool suitable for tasks ranging from simple image transformations to advanced real-time video analysis, including object detection and tracking. Originally developed by Intel, OpenCV is now maintained by an active open-source community. The library provides more than 2500 optimized algorithms for image and video processing.

5) *Torchvision*

Torchvision, a module within the PyTorch ecosystem, was specifically developed for computer vision

applications [116]. It enables the easy implementation of an end-to-end vision framework within PyTorch by providing tools that streamline image processing, augmentation, and dataset handling. Torchvision offers pre-trained models for transfer learning and integrates seamlessly with PyTorch's DataLoader, facilitating efficient data preparation and augmentation. Additionally, Torchvision includes several widely used datasets that are readily available, such as MNIST, CIFAR-10/100, ImageNet, and COCO.

The methodology proposed by Bakliwal and Joshi [117] leverages multiple CNN classifiers, specifically 2D and 3D models, to combine their complementary strengths in capturing high-level semantic information. The primary objective of this approach is to improve overall Deepfake detection accuracy. The spatio-temporal properties captured by the 3DCNN are crucial for detecting subtle abnormalities in Deepfake videos that conventional 2D models may overlook. The EfficientNet architecture, used in the 2D model, is well known for its efficiency in image classification tasks. To further enhance model robustness, captured faces are subjected to a range of data augmentation techniques during training and validation, including horizontal flipping, shearing, rotations, noise addition, contrast adjustments, random downscaling, and normalization. These strategies are implemented using the Albumentation library, contributing to a more diverse training dataset and thereby improving model robustness. However, a key challenge of this approach lies in the integration of 2D and 3D components, which increases training complexity compared to single-model systems.

An ensemble method using EfficientNet b4 and other CNN models, incorporating various attention layers and Siamese training strategies, was proposed by Bonettini *et al.* [118] to improve detection accuracy. The models were trained using specific hyperparameters and the Adam optimizer to ensure robust performance. To enhance robustness, facial data augmentation techniques were applied, including random downscaling, horizontal flipping, adjustments to saturation, color, and contrast, noise addition, and JPEG compression. These augmentations were implemented using the Albumentation library, helping to reduce overfitting and diversify the training dataset. However, due to hardware and time constraints during the DFDC challenge, the study restricted the number of frames per video to 32 during both training and testing. This limitation may affect the model's ability to capture temporal dynamics in longer videos and may result in the loss of contextual information crucial for detecting manipulations, as the approach primarily focuses on facial regions extracted from video frames.

Lu and Ebrahimi [119] also explored augmentation strategies aimed at improving Deepfake detection in real-world scenarios. SDAug focuses on simulating compression artifacts and noise commonly encountered in image and video processing. This approach strengthens Deepfake detectors' resistance to various distortions. The augmentation process begins with image enhancement, adjusting contrast with a 50% probability, followed by the addition of Gaussian noise. These augmentation

techniques are applied using the Torchvision library. SDAug significantly improves Deepfake detection performance, particularly when evaluated on unseen datasets such as Celeb-DF v2. The results demonstrate that

models trained with SDAug exhibit superior generalization capabilities compared to detectors trained without this augmentation.

TABLE V. SUMMARY OF DEEFAKE DETECTION TECHNIQUES BASED ON PYTHON LIBRARIES AUGMENTATIONS

Ref.	Classifier/Techniques	Augmentation Techniques	Dataset	Accuracy	AUC
[117]	CNN	Albumentation library	DFDC	-	89.69%
[118]	EfficientNet b4	Albumentation library	FF++ and DFDC	-	94.44%
[120]	Xception Net and EfficientNet b4	Random cut-out from Imgaug Python library	FF++ and DFDC	98.90%	-
[119]	Capsule-Forensics and Xception Net	Torchvision	Celeb-DF-V2 and FF++	-	73.88%

E. Occlusion

Occlusion refers to techniques (as shown in Table VI) used to disrupt the detection of Deepfake content by introducing additional elements into video frames. These distractions make it more challenging for detection algorithms to accurately distinguish between real and fake videos, thereby improving model generalization [121]. The blending and synthesis processes involved in creating Deepfakes often result in certain facial areas, such as the lips, cheeks, or eyes, being more susceptible to artifacts. By selectively occluding or revealing specific facial features, occlusion-based data augmentation allows researchers to focus on particular facial regions. This technique helps models identify the precise areas where Deepfake artifacts are most likely to appear, as illustrated in Fig. 10. The various methods of occlusion are as follows:

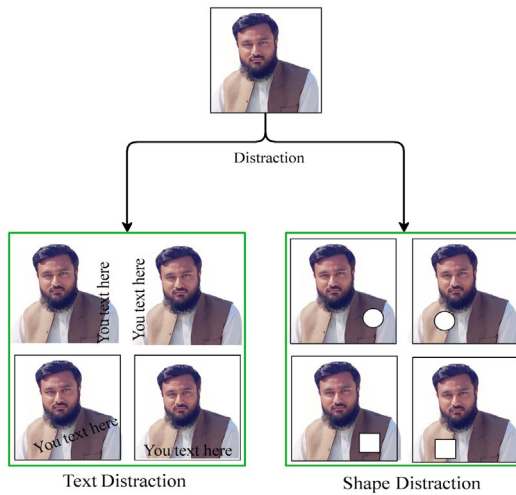


Fig. 10. Occlusion.

1) Text overlays

Random alphanumeric text is generated and overlaid onto video frames. The text can be either static or dynamic, scrolling across the screen in various directions. The font size, thickness, and color are randomly selected to enhance the distraction effect.

2) Geometric shapes

Shapes such as circles, squares, or even hand shadows are placed over the image to obscure facial features, similar to text overlays. This technique enhances model robustness and improves generalization.

The MTCNN pretrained model, as utilized by Prajapati and Pollett [121], is used to detect and extract faces from video frames. By analyzing every tenth frame and extracting faces from those frames, the method reduces computational costs. Features are subsequently extracted from the detected faces using EfficientNet b0. To determine whether a face is authentic or a Deepfake, these features are classified using a Multilayer Perceptron (MLP) model. This innovative approach leverages GANs to identify perceptual differences in images, with particular emphasis on Structural Similarity Index (SSIM) for evaluating image quality. The DFDC training set undergoes various augmentation techniques, including blurring, brightness and contrast adjustments, and the addition of different noise types (such as Gaussian, speckle, and salt-and-pepper noise). The study also employs several distraction strategies to simulate real-life video scenarios, including static, rolling, and random text overlays, as well as shape-based distractions (e.g., circles and rectangles). As a result of these extensive augmentations, the plain frame-based model achieves an accuracy of 91.0%, significantly outperforming the Magnetic Resonance Imaging (MRI)-GAN framework, which achieves a test accuracy of 74%.

TABLE VI. SUMMARY OF DEEFAKE DETECTION TECHNIQUES BASED ON OCCLUSION AUGMENTATIONS

Ref.	Classifier/Techniques	Augmentation Techniques	Dataset	Accuracy	AUC
[121]	MTCNN, EfficientNet b0, and MLP	Occlusion (overlying static, rolling, and spontaneous text)	DFDC, Flickr Diverse Faces (FDF), FFHQ, and Celeb-DF-v2	91.0%	-

F. Random Noise

One highly effective technique for data augmentation is the addition of noise. This method is particularly useful for handling blurry images, as introducing noise can improve model robustness. Random changes in pixel values, referred to as noise, can mislead models during training

and potentially affect accuracy. However, this challenge can be mitigated by controlling the noise level to ensure that the model focuses on significant features. By adding noise to the dataset, the Deepfake detection classifier is exposed to more diverse patterns, thereby enhancing generalization and making the technique more adaptable to real-world scenarios. Common data augmentation

techniques involving random noise include salt-and-pepper, Gaussian, Poisson, and speckle noise. These techniques improve model robustness and generalizability by introducing controlled randomness into the training data. Various noise-based techniques are illustrated in Fig. 11 and summarized in Table VII, respectively.

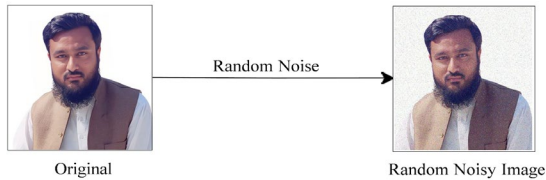


Fig. 11. Random noise.

1) Salt and pepper noise

Salt-and-pepper noise refers to a phenomenon in which a certain percentage of an image's pixels are randomly set to either black or white, resembling salt (white dots) and pepper (black dots). This type of noise is commonly used in image augmentation for deep learning model training. By introducing salt-and-pepper noise, models are forced to focus on partial features, thereby improving robustness. It essentially acts as a random mask, disrupting the image while maintaining the structural integrity of key features. This technique enhances model generalization by training the model to recognize important patterns even in the presence of random pixel distortions [122].

2) Gaussian noise

Gaussian noise, associated with the Gaussian distribution, is a statistical noise model that follows a normal distribution and exhibits a symmetric, bell-shaped probability density function. The normal distribution is one of the most important and widely used statistical distributions. It is frequently observed in real-world scenarios, as many natural phenomena and random events exhibit Gaussian characteristics. It arises from the sum of multiple independent random variables, resulting in a distribution that approximates normality. In image augmentation, Gaussian noise is introduced by adding random values based on this distribution, thereby increasing dataset variability and enhancing model robustness to natural randomness in input data [123].

3) Poisson noise

Photon noise, also known as Poisson noise, arises due to the quantized nature of light and the discrete detection of photons. It is a type of signal-dependent noise in which the amplitude fluctuates according to signal strength. Poisson noise is typically observed in low-light conditions or in scenarios where the number of detected photons is limited. In data augmentation, Poisson noise is intentionally added to the training data to simulate real-world noise conditions, thereby enhancing model robustness and improving generalization in noisy environments. By incorporating this type of noise, the model becomes more capable of handling variations and uncertainties encountered in practical scenarios, ultimately improving detection accuracy in tasks such as Deepfake

detection [124, 125].

4) Speckle noise

Speckle noise is a common noise type inherent in images obtained through coherent imaging modalities, including laser, Synthetic Aperture Radar (SAR), and ultrasound imaging [126]. Notably, speckle noise exhibits multiplicative behavior, influencing the variance of pixel values rather than their mean intensity, which distinguishes it from additive noise types such as Gaussian or salt-and-pepper noise. This type of noise is generated by multiplying random values with pixel intensities in the image, resulting in a grainy and textured appearance that can obscure fine details and reduce image contrast. To address speckle noise effectively, specialized filtering techniques are required to reduce noise while preserving edges and structural information. In the context of Deepfake detection, speckle noise is often incorporated as part of data augmentation to improve model generalization capabilities [127]. The study proposed a strategy for enhancing Deepfake detection through data augmentation [128]. By introducing variations into the training dataset, the model adapts more effectively to different types of Deepfake manipulations. Using autoencoders, the authors generated slightly modified images from the original dataset, maintaining image quality while adding noise or frequency-based perturbations. This approach significantly enhanced detection performance, with all tested models achieving over 99% AUC on the FF++ dataset. However, augmentation was applied only to static frames, and attempts to incorporate temporal networks such as LSTMs or transformers did not yield improved results. The findings indicated that the efficiency of the autoencoder models had no significant impact on overall Deepfake detection performance, suggesting that even lower-performing models provided beneficial augmentation effects.

To enhance generalization, Abdelminaam *et al.* [129] added Gaussian noise to simulate artifacts arising from low-quality recordings or video compression. However, excessive augmentation may produce misleading results or cause overfitting to specific artifacts, even though noise addition generally improves robustness. Real-time applications remain challenging due to the high computational costs associated with face extraction, augmentation, and model training. Similarly, various augmentation techniques, such as contrast adjustment and compression, were implemented by Waseem *et al.* [130]. Their findings revealed that these techniques significantly improved model performance on unseen datasets. A comparative analysis demonstrated that models trained with augmentation (w-aug) outperformed those trained without augmentation (w/o-aug).

To increase the robustness of Deepfake detection techniques, Lu and Ebrahimi [131] proposed a novel data augmentation strategy. By applying different probabilities of noise addition and JPEG compression, the augmentation process replicates real-world degradation effects. This approach aims to train the model to handle distortions that arise during image capture, compression,

and processing. Regarding limitations, the study notes that although the augmentation technique improves resilience to various distortions, it may still be insufficient when addressing complex or unpredictable manipulations not covered by the augmentation procedure.

Coccomini *et al.* [132] described transformations applied randomly during training, including Poisson and speckle noise. These augmentations were implemented after face extraction using the MTCNN detector, focusing on facial regions critical for Deepfake detection. The training dataset included images manipulated using various techniques, with more than 36 perturbations applied to over 4300 objects. This diversity enabled the models to effectively identify Deepfake content generated through different manipulation methods.

Ganiyusufoglu *et al.* [133] utilized both frame-level and sequence-level features, combining data augmentation techniques to enhance generalization and proposing an effective Deepfake detection method based on spatio-temporal analysis. The method identified image-level artifacts typical of Deepfake manipulations by extracting spatial features from each frame using a CNN, and analyzed temporal inconsistencies between frames using an RNN [134]. Nevertheless, the model's performance declined when confronted with high-quality Deepfakes exhibiting natural temporal coherence, as it primarily focused on detecting frame-to-frame irregularities. Real-time Deepfake detection remains

challenging due to the computational complexity of integrating CNN and RNN architectures.

Geometric transformations (Table II) demonstrate reported accuracy ranging approximately from 85% to 99.9%, with AUC values frequently exceeding 97%, particularly when evaluated on large benchmark datasets such as FF++, DFDC, and Celeb-DF. These augmentation strategies, including rotation, flipping, cropping, scaling, and translation, enhance dataset diversity and help models learn invariant spatial features, thereby improving robustness to variations in facial orientation and positioning.

Cutout and mixing strategies (Table III) show improved robustness by randomly masking or mixing regions of the input images during training. Several studies report accuracy values above 97% and AUC values above 91%, indicating that these augmentation methods effectively reduce overfitting and enhance generalization across datasets such as DFDC, FF++, and Celeb-DF. Photometric transformations (Table IV), including brightness adjustment, contrast modification, color jittering, and blurring, demonstrate strong cross-dataset generalization. Multiple studies achieve accuracy levels above 90% and AUC values ranging between approximately 97% and 99.8%, highlighting the effectiveness of color and illumination-based augmentation strategies for improving detection performance under varying lighting and visual conditions.

TABLE VII. SUMMARY OF DEEFAKE DETECTION TECHNIQUES BASED ON RANDOM NOISE AUGMENTATIONS

Ref.	Classifier/Techniques	Augmentation Techniques	Dataset	Accuracy	AUC
[129]	CNNs and VGG	Random noise and color correction	DFDC and FF++	94.0%	-
[130]	CNNs and VIT	Gaussian noise and compression	FF++, Celeb-DF, and DFDC	92.50%	97.0%
[135]	CNN-LSTM and RNN	Gaussian noise	FF++	99%	-
[131]	Capsule Forensics and XceptionNet	Gaussian noise	FF++ and DFDC	-	99.56%
[128]	Autoencoder, XceptionNet, and EfficientNet b4	Salt and pepper Noise addition	Celeb-DF, FF++, and DFDC	82.87%	99.0%
[132]	VIT and EfficientNet-V2	Poisson and speckle noise	ForgeryNet	81.10%	-
[133]	RNNs and 3DCNN	Compression and Gaussian noise	FF++	88.57%	-
[136]	Texture and Artifact Detector (TAD)	Random noise	FF++ and DFDC	97.11%	-
[137]	ResNet, EfficientNet, DenseNet, and MobileNet (spatial based)	Noise, exposure, and blur	FF++, StyleGANv2, StarGAN, and ProGAN	99.3%	-

Python-based augmentation libraries (Table V), such as Albumentations, Torchvision, PIL, and Imgaug, provide flexible frameworks for implementing multiple augmentation operations within deep learning pipelines. Studies utilizing these libraries report accuracy typically ranging between 90% and 98%, demonstrating that integrated augmentation pipelines can improve training efficiency and model robustness across diverse datasets. Occlusion-based augmentation techniques (Table VI) focus on masking or obscuring specific facial regions to encourage models to learn more generalized facial representations.

Experimental results show accuracy commonly ranging from approximately 88% to 97%, suggesting that occlusion strategies improve resilience to partial facial obstruction and manipulation artifacts. Random noise augmentation techniques (Table VII), including Gaussian noise, salt-and-pepper noise, Poisson noise, and speckle

noise, introduce controlled perturbations to simulate real-world distortions. Studies report accuracy typically between 85% and 96% with strong AUC performance, indicating that noise-based augmentation helps models become more robust to compression artifacts, sensor noise, and other visual degradations encountered in real-world Deepfake content.

G. Quantitative Synthesis and Quality Assessment

To provide clearer benchmarking and assess the impact of augmentation, we synthesized performance across 130 studies (Tables II–XI). Due to heterogeneity in datasets, model architectures, and evaluation protocols, the results are presented as ranges rather than through formal meta-analysis. This approach aligns with standard practices in systematic reviews, where heterogeneous studies are synthesized using comparative and range-based analysis rather than controlled experimental evaluation.

Although a full Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) or A MeaSurement Tool to Assess systematic Reviews (AMSTAR) checklist is not strictly applied, the review follows a structured study selection and synthesis process, including predefined inclusion criteria, systematic screening, and comparative analysis, which aligns with standard practices in systematic literature reviews.

Although a full PRISMA or AMSTAR checklist is not strictly applied, the review follows a structured study selection and synthesis process, including predefined inclusion criteria, systematic screening, and comparative analysis, which aligns with standard practices in systematic literature reviews.

1) Performance benchmarks by augmentation group

Table VIII summarizes the performance of

augmentation-based methods in comparison to baseline configurations. Baselines represent models trained without augmentation or with minimal preprocessing, as reported in the original studies.

Across the reviewed studies, geometric and mixing-based augmentations are associated with relatively higher performance improvements, while photometric and noise-based methods contribute to improved robustness under varying visual conditions. Overall, augmentation techniques are generally associated with measurable performance improvements relative to baseline models reported in the literature, with improved robustness under standard evaluation conditions, but limited resilience in adversarial or security-critical scenarios. The synthesis is supported by a structured study selection process, as illustrated in Fig. 2, ensuring transparency and reproducibility.

TABLE VIII. PERFORMANCE SYNTHESIS BY AUGMENTATION TYPE ($N = 130$ STUDIES). BASELINES CORRESPOND TO NON-AUGMENTED MODELS REPORTED IN THE SAME STUDIES. REPORTED IMPROVEMENTS ARE APPROXIMATELY 5–15% ACROSS DIFFERENT SETTINGS

Augmentation Group	Datasets	Accuracy Range (Augmented)	AUC Range	Baseline Accuracy
Geometric	FF++, Celeb-DF, DFDC	92–99%	0.97–0.99	~78–86%
Photometric	FF++, Celeb-DF	90–99%	0.95–0.99	~75–85%
Cutout/Mixing	DFDC, Celeb-DF	91–99%	0.91–0.99	~70–82%
Noise/Occlusion	FF++, DFDC	88–97%	0.92–0.99	~72–85%

TABLE IX. QUALITY SCORING FRAMEWORK APPLIED TO $N = 130$ STUDIES: 45 HIGH (4–6), 65 FAIR (2–3), AND 20 LOW (0–1)

Criteria	0 (Low)	1 (High)	Max
Datasets	Single dataset	≥3 datasets cross-dataset evaluation	1
Augmentation Analysis	No ablation	Includes comparison with baseline	1
Models	Single model	Multiple models or ensemble	1
Validation	Basic split	K-fold or real-world testing	1
Reporting	Limited metrics	Comprehensive (Acc/AUC/EER)	1
Generalization	In-domain only	Cross-domain or adversarial testing	1
Total	-	-	6

2) Quality scoring system

To assess methodological rigor, a structured quality scoring framework was developed, as shown in Table IX. Each criterion is assigned a binary score (0 or 1), resulting in a total score ranging from 0 to 6. Studies are categorized as high quality (4–6), fair (2–3), or low quality (0–1).

High-quality studies consistently report stronger and more reliable performance, particularly for geometric augmentation methods achieving AUC values above 0.95, whereas low-quality studies often overestimate performance due to limited datasets and lack of generalization. Importantly, the core conclusions of this review are primarily supported by high-quality studies (scores 4–6), which demonstrate consistent improvements in performance and generalization. In contrast, findings from low-quality studies are interpreted with caution due to limitations such as small datasets, lack of baseline comparison, and insufficient validation. This quality-based grouping enables a structured interpretation of the literature, ensuring that conclusions are weighted according to methodological rigor rather than treated uniformly.

3) Grouped analysis and trends (2023–2025)

Table X presents performance trends based on high-quality studies. The results indicate a consistent

improvement in reported performance over time, driven by the integration of hybrid augmentation strategies and improved model architectures. This trend reflects the increasing adoption of hybrid augmentation strategies and more robust model architectures in recent studies.

TABLE X. PERFORMANCE TREND BY PUBLICATION YEAR (HIGH-QUALITY STUDIES ONLY). VALUES REPRESENT TYPICAL RANGES DERIVED FROM REPORTED RESULTS

Year	Typical Accuracy Range	Typical AUC Range	Studies
2023	90–94%	0.93–0.95	18
2024	93–96%	0.95–0.97	32
2025	95–98%	0.97–0.99	25

To further support quantitative synthesis, a frequency-based analysis of the reviewed studies was conducted. Among the 130 selected papers, geometric augmentation techniques were the most commonly used, appearing in approximately 70% of the studies, followed by photometric methods (around 55%), noise and occlusion-based approaches (approximately 40%), and cutout or mixing strategies (about 35%).

In terms of datasets, FF++ was the most frequently utilized benchmark, followed by DFDC and Celeb-DF. Additionally, convolutional neural network (CNN)-based architectures were dominant across the literature, often

combined with hybrid augmentation strategies in recent studies. This frequency-based distribution highlights the prevailing research trends and supports the grouped quantitative analysis presented in this review.

These values are derived from high-quality studies and are presented as representative ranges rather than exact statistical averages. Together, the frequency analysis, grouped performance comparison, and temporal trends provide a structured quantitative synthesis of augmentation effects beyond descriptive reporting as described in Table XI.

TABLE XI. PERFORMANCE GROUPED BY QUALITY SCORE

Quality Level	Accuracy Range	Key Datasets	Observations
4–6 (High)	94–99%	FF++, Celeb-DF, DFDC	Strong gains with geometric and hybrid augmentation
2–3 (Fair)	88–95%	FF++, DFDC	Improved robustness with photometric methods
0–1 (Low)	82–92%	Single dataset	Risk of overfitting and limited generalization

4) Critical limitations analysis

Cross-Dataset Generalization Failure: Models achieving AUC values above 0.97 on FF++ often experience a decline to 0.80–0.92 when evaluated on Celeb-DF, which contains higher-fidelity manipulations. These observations indicate that, although augmentation enhances performance under controlled conditions, it does not fully address robustness challenges in adversarial environments. Therefore, claims regarding robustness must be interpreted within the context of standard evaluation settings rather than security-critical deployments. These failure cases highlight that augmentation is not universally beneficial and must be carefully designed to avoid performance degradation.

Solution: Incorporating cross-dataset training and paired real-fake augmentation strategies. Augmentation improves robustness primarily under standard and data-limited conditions, but its effectiveness decreases in cross-dataset and high-variation scenarios.

Cause: Variations in dataset characteristics and manipulation artifacts.

Contradictory Findings: Some studies report limited or negligible improvements from photometric augmentation under controlled lighting conditions. This inconsistency is primarily attributed to reduced variability and dataset bias.

Failure Cases of Augmentation: Several studies report cases where augmentation does not lead to performance improvement and may even degrade detection accuracy. For instance, excessive geometric transformations or inappropriate mixing strategies can distort key facial features, leading to reduced model sensitivity. Similarly, photometric augmentation under already controlled lighting conditions has been shown to introduce noise rather than useful variability, resulting in marginal or negative performance gains.

In some cases, augmentation applied without proper balance across classes leads to biased learning, further degrading performance. These findings indicate that the effectiveness of augmentation is highly dependent on

dataset characteristics, parameter selection, and model architecture, and that improper application can adversely affect detection performance.

Security Failures: Adversarial attacks can reduce AUC by approximately 10–30%, with noise-based augmentation particularly vulnerable to adversarial perturbations.

Security-Oriented Robustness Limitation: While augmentation techniques improve performance under standard evaluation settings, their effectiveness in adversarial or security-critical scenarios remains limited. Several studies report that Deepfake detection models, including those trained with augmentation, are vulnerable to adversarial manipulations designed to evade detection.

In particular, perturbation-based attacks, compression artifacts, and unseen manipulation techniques can significantly degrade model performance, indicating that augmentation alone is insufficient to ensure robustness. This highlights a gap between reported performance improvements and real-world security requirements.

Implication: Augmentation should be complemented with security-aware training strategies, such as adversarial training, anomaly detection mechanisms, and cross-domain generalization techniques, to improve resilience against malicious attacks.

Mitigation Strategies: Adversarial training and incorporation of diverse augmentation techniques.

Risk-of-Bias Considerations: Several sources of bias are identified across the reviewed studies. Dataset bias is common, where models trained on specific datasets (e.g., FF++) may not generalize well to others such as Celeb-DF. Reporting bias is also observed, as some studies emphasize improved results without providing complete baseline comparisons or statistical validation. Additionally, overfitting bias arises in studies using limited datasets or lacking cross-validation. These biases are more prevalent in low-quality studies and are explicitly considered when interpreting the overall findings of this review.

Overall, the synthesized evidence indicates that augmentation is consistently associated with measurable performance improvements (typically 5–15%) relative to baseline models reported in the literature. However, the magnitude of improvement varies depending on dataset characteristics, model architecture, and evaluation settings. These findings are presented in a study-relative manner, consistent with comparative analyses reported in Ref. [138]. It is important to note that these observations are derived from reported results in the literature and may be influenced by variations in experimental design, dataset composition, and reporting practices across studies.

IV. KEY CHALLENGES AND OPEN ISSUES

Deepfake detection methods have demonstrated strong performance in controlled experimental settings, particularly through the integration of facial data augmentation techniques. However, despite these advances, significant limitations persist that restrict their reliability in real-world deployment. Many reported improvements are benchmark-dependent and do not

consistently translate to unconstrained environments characterized by diverse capture conditions, compression pipelines, and evolving manipulation strategies. Key challenges include the limited realism of existing augmentation procedures, insufficient dataset diversity, and the rapid evolution of Deepfake generation techniques, which collectively undermine long-term robustness and cross-dataset generalization. Moreover, the trade-off between model complexity and computational efficiency poses a critical barrier to real-time and resource-constrained deployment. Although data augmentation is widely used to improve model robustness and generalization, it may also introduce potential failure cases when improperly applied. Excessive or poorly designed augmentation strategies can distort facial structures, alter subtle manipulation cues, or introduce artificial artifacts that are not representative of real-world Deepfake characteristics. In such situations, models may learn misleading patterns or overfit to augmented artifacts rather than genuine manipulation signals, resulting in degraded detection performance. Furthermore, augmentation strategies that are effective for one dataset may not generalize well to others due to differences in manipulation techniques, compression levels, and visual realism. These observations highlight the importance of carefully selecting augmentation techniques and evaluating their impact across multiple datasets to ensure reliable Deepfake detection performance. This section critically examines these limitations and highlights unresolved issues that must be addressed to bridge the gap between laboratory performance and practical applicability.

A. Labeling

A major challenge in data augmentation is ensuring the correct labeling of augmented data. When operations such as scaling, cropping, or rotation are applied to the original data, it becomes crucial to update the labels accurately to reflect these transformations [139]. Mislabeling augmented data can lead to confusion during model training, potentially resulting in performance degradation due to errors and inconsistencies in the learning process.

B. Aggressive Augmentation Leading to Overfitting

Overfitting is a common challenge in machine learning and deep learning, where a model performs well on training data but poorly on unseen data. While data augmentation is commonly used as a mitigation strategy, excessive augmentation can introduce a different form of overfitting [140]. Excessive augmentation may cause the model to overly specialize in patterns present in augmented data while neglecting the characteristics of the original data. This, in turn, can impair the model's ability to generalize effectively to real-world scenarios.

C. Cross-Dataset Generalization and Augmentation Limitations

Cross-dataset generalization remains a major challenge in Deepfake detection [141], as models trained on one dataset, such as FF++, often exhibit significantly reduced performance when evaluated on others, including

Celeb-DF, DFDC, or DFD. Performance inconsistencies across commonly used datasets such as FF++ and Celeb-DF highlight an important challenge in Deepfake detection research, particularly when evaluating the effectiveness of data augmentation techniques. Several studies report very high accuracy on FF++, while the same models experience noticeable performance degradation when evaluated on Celeb-DF. This discrepancy is largely attributed to differences in dataset characteristics. FF++ contains varying compression levels and visible manipulation artifacts that detection models can exploit during training, and common augmentation techniques such as compression simulation, geometric transformations, and noise injection often reinforce these artifact-based cues. In contrast, Celeb-DF was designed to produce more visually realistic Deepfakes with fewer detectable artifacts, making such augmentation strategies less effective. As a result, models trained with augmentation strategies optimized for FF++ may inadvertently learn dataset-specific artifacts rather than generalized manipulation patterns. These observations indicate that while augmentation can improve robustness within a specific dataset, its effectiveness may remain dataset-dependent. Consequently, cross-dataset evaluation is essential to determine whether augmentation techniques truly enhance generalizable Deepfake detection performance rather than merely amplifying dataset-specific features. While augmentation techniques can enhance model robustness within a single dataset, they frequently fail to capture the distributional diversity necessary for effective cross-domain generalization. In some cases, advanced augmentation methods may introduce synthetic artifacts that inadvertently cause models to overfit to unrealistic patterns, further hindering transferability.

Prominent industry benchmarks and datasets, such as Meta's DFDC [142], Google and Jigsaw's DFD [143], and collaborative efforts by DeepMind, underscore the urgent need for scalable and generalizable detection systems capable of performing reliably under varied real-world conditions. Despite ongoing progress, most existing augmentation techniques remain static and manually engineered. Future research should explore dynamic, context-aware augmentation strategies that adapt in response to emerging Deepfake generation techniques. Moreover, there is a growing need for augmentation methods that more accurately replicate the natural imperfections present in user-generated content, as well as frameworks that jointly optimize detection accuracy and generalization to previously unseen domains.

These findings suggest that many current augmentation strategies primarily improve intra-dataset robustness rather than true generalization, raising concerns about their effectiveness in real-world scenarios where manipulation techniques and data distributions are unknown in advance.

D. Domain Specific Issues

There are specific challenges associated with data augmentation across different domains and applications [144]. For example, when augmenting medical images, it is crucial to maintain data integrity and

avoid introducing artifacts that could compromise diagnostic accuracy. Similarly, in natural language processing, text generated through augmentation techniques must remain logically coherent and contextually relevant to ensure its usefulness.

E. Adversarial Robustness and Security Evaluation Challenges

While many augmentation-based Deepfake detection frameworks report improvements in robustness and accuracy across benchmark datasets, relatively few studies evaluate their resilience under security-driven threat models. In most cases, models are tested using standard datasets such as FF++, Celeb-DF, or DFDC without considering adversarial manipulation, evasion attacks, or intentionally crafted perturbations designed to bypass detection systems. As a result, improvements attributed to data augmentation often reflect robustness to dataset variability rather than true resistance to adversarial or security-oriented attacks. This limitation highlights the need for future research to incorporate cybersecurity-aware evaluation protocols, including adversarial testing, attack simulation, and real-world threat modeling, to ensure that Deepfake detection systems remain reliable in adversarial environments.

F. Data Diversity and Balance

Balancing the extent of data augmentation is crucial. Insufficient augmentation may fail to introduce enough diversity to improve model performance, while excessive augmentation can lead to a different form of overfitting [145]. Achieving the optimal balance requires a deep understanding of the problem domain, the dataset, and the specific machine learning or deep learning models being used.

G. Computational Cost

When working with large datasets, data augmentation can impose substantial computational demands [146]. Simple transformations such as rotation, flipping, and color jittering are relatively cost-effective and can typically be applied on-the-fly during training with minimal impact on processing speed. In contrast, complex augmentations such as attention-guided region masking, face reenactment, or GAN-based synthetic data generation require considerably greater computational resources. For instance, GAN-based augmentations often result in increased storage and memory consumption, as well as higher computational overhead during model training [132]. Similarly, attention-guided methods (e.g., region enlargement or face occlusion) typically depend on preprocessing steps involving object detectors or facial landmark extractors, which further increase time and memory requirements.

These challenges become more severe in low-resource environments, such as real-time or edge-device scenarios. Therefore, while advanced augmentation strategies have the potential to improve detection performance, their computational costs must be carefully considered when deploying models in practical, resource-constrained applications.

As a result, several high-performing detection models remain impractical for deployment in large-scale monitoring systems or on edge devices, where strict latency, memory, and energy constraints are unavoidable.

H. Requirement for Original and High-Quality Data

The effectiveness of data augmentation is directly influenced by the quality of the original dataset [147]. Data augmentation cannot fully compensate for issues such as noise, labeling errors, or insufficient diversity. Regardless of how advanced augmentation techniques are, they cannot resolve fundamental problems related to data quality. Therefore, ensuring that the base dataset is of high quality is essential for the successful application of data augmentation.

In Deepfake detection, an augmentation strategy can be considered effective when it preserves manipulation-relevant visual cues while enhancing robustness to realistic data variations, thereby supporting improved generalization across datasets. Conversely, augmentation may be detrimental if it suppresses or distorts forgery-related artifacts, introduces unrealistic synthetic patterns, or biases the learning process toward dataset-specific features, ultimately degrading cross-dataset performance. The impact of augmentation is therefore context-dependent and influenced by dataset characteristics, manipulation techniques, and the underlying detection architecture.

V. CONCLUSION

This review highlights the critical role that facial data augmentation techniques play in enhancing the accuracy, generalization, and robustness of Deepfake detection models. Our analysis demonstrates that basic transformations (such as rotation, flipping, and color jittering) can yield consistent performance improvements with minimal computational cost. In contrast, more advanced approaches, including region masking and GAN-based synthetic augmentation, offer greater detection improvements but often require substantial computational resources.

However, direct comparisons across studies remain challenging due to variations in datasets, augmentation strategies, and evaluation metrics. The apparent superiority of certain methods may be highly dependent on specific experimental conditions, thereby emphasizing the need for cautious interpretation of reported results. Nevertheless, augmentation techniques consistently demonstrate effectiveness in mitigating overfitting, increasing training data diversity, and enhancing model resilience to subtle or previously unseen manipulations. Importantly, their success is closely linked to the quality and representativeness of the underlying datasets.

Key challenges persist, particularly with respect to computational efficiency, the standardization of evaluation protocols, and the preservation of high-quality source data. While advanced augmentation techniques can improve detection performance, practitioners must strike an appropriate balance between algorithmic complexity and practical deployment constraints, such as limited

computational resources and real-time application requirements.

In conclusion, this review synthesizes the current progress and ongoing challenges in facial data augmentation for Deepfake detection. By providing this structured overview, we aim to support the development of more robust, scalable, and generalizable detection systems and to guide researchers and practitioners in making informed decisions when applying augmentation techniques in real-world scenarios.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION

Naveed Ur Rehman Ahmed conceptualized the study, designed the methodology, conducted the systematic literature review, performed data extraction and analysis, and drafted the original manuscript. Afzal Badshah and Ayesha Tajammul contributed to the literature screening, methodological refinement, and validation of the extracted data. Omar Alghushairy and Riad Alharbey assisted in result interpretation, critical technical review, and manuscript editing. Ali Daud provided supervision, conceptual guidance, and critical revisions to improve the overall quality of the manuscript. All authors reviewed and approved the final manuscript.

REFERENCES

- [1] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4490–4499.
- [2] P. Kumar, M. Vatsa, and R. Singh, "Detecting face2face facial reenactment in videos," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2589–2597.
- [3] Z. Xu *et al.*, "Facecontroller: Controllable attribute editing for face in the wild," in *Proc. AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 3083–3091.
- [4] S. Zobaed *et al.*, "Deepfakes: Detecting forged and synthetic media content using machine learning," in *Proc. Artificial Intelligence in Cyber Security: Impact and Implications: Security Challenges, Technical and Ethical Issues, Forensic Investigative Challenges*, 2021, pp. 177–201.
- [5] C. Chipeta. (2025). Deepfake statistics (2025): 25 new facts for CFOs. [Online]. Available: <https://www.eftsure.com/statistics/deepfake-statistics/>
- [6] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood, "Towards deep learning prospects: insights for social media analytics," *IEEE Access*, vol. 7, pp. 36958–36979, 2019.
- [7] N. U. R. Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud, and T. Alsahfi, "Visual deepfake detection: Review of techniques, tools, limitations, and future prospects," *IEEE Access*, pp. 1923–1961, 2024.
- [8] M. A. Hossain and M. S. Islam, "Ensuring network security with a robust intrusion detection system using ensemble-based machine learning," *Array*, vol. 19, 100306, 2023.
- [9] O. for Social Media Safety. (2025). Deepfake technology. [Online]. Available: <https://www.socialmediasafety.org/advocacy/deepfake-technology/>
- [10] BBC. (2019). Deepfakes: What are they and why would I make one? [Online]. Available: <https://www.bbc.co.uk/bitesize/articles/zfkwcqt>
- [11] A. Badshah *et al.*, "Smart cities' big data: Performance and cost optimization with regional computing," *IEEE Access*, vol. 12, pp. 128896–128908, 2024.
- [12] R. Javed, T. Abbas, A. H. Khan, A. Daud, A. Bukhari, and R. Alharbey, "Deep learning for lungs cancer detection: A review," *Artificial Intelligence Review*, vol. 57, no. 8, 197, 2024.
- [13] U. Zara, K. Ayub, H. U. Khan, A. Daud, T. Alsahfi, and S. Gulzar, "Phishing website detection using deep learning models," *IEEE Access*, vol. 12, pp. 167072–167087, 2024.
- [14] A. Naz, H. U. Khan, S. Alesawi, O. I. Abouola, A. Daud, and M. Ramzan, "AI knows you: Deep learning model for prediction of extroversion personality trait," *IEEE Access*, vol. 12, pp. 159152–159175, 2024.
- [15] A. Badshah *et al.*, "Big data applications: overview, challenges and future," *Artificial Intelligence Review*, vol. 57, no. 11, 290, 2024.
- [16] O. Debroy and B. D. Hemmige, "Psycho-social impact of deepfake content in entertainment media," *International Journal for Innovative Research in Multidisciplinary Field*, vol. 10, no. 5, pp. 235–249, 2024.
- [17] C. M. S. Faisal, A. Daud, F. Imran, and S. Rho, "A novel framework for social web forums' thread ranking based on semantics and post quality features," *The Journal of Supercomputing*, vol. 72, pp. 4276–4295, 2016.
- [18] L. Skene. (2025). Former school athletic director gets 4 months in jail in racist AI deepfake case. [Online]. Available: <https://apnews.com/article/racist-ai-recording-maryland-high-school-487ea673b0449077cb23e7970546cb9f>
- [19] B. Allyn. (2025) Deepfake video of zelenskyy could be 'tip of the iceberg' in info war, experts warn. [Online]. Available: <https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia>
- [20] K. Kurmanath. (2025) Deepfake videos of CMS of two Telugu states used to peddle fraudulent financial schemes. [Online]. Available: <https://www.thehindubusinessline.com/news/national/cms-of-two-telugu-states-fall-victims-to-deepfake-fraud/article69826567.ece>
- [21] C. Yang, L. Ding, Y. Chen, and H. Li, "Defending against GAN-based deepfake attacks via transformation-aware adversarial faces," in *Proc. 2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9533868
- [22] W. Khan *et al.*, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, 100026, 2023.
- [23] S. Kazi, S. Khoja, and A. Daud, "A survey of deep learning techniques for machine reading comprehension," *Artificial Intelligence Review*, vol. 56, no. Suppl 2, pp. 2509–2569, 2023.
- [24] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu, "A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, 2022.
- [25] A. Badshah, A. Daud, H. U. Khan, O. Alghushairy, and A. Bukhari, "Optimizing the over and underutilization of network resources during peak and off-peak hours," *IEEE Access*, vol. 12, pp. 82549–82559, 2024.
- [26] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait Journal of Science*, vol. 43, no. 4, pp. 95–113, 2016.
- [27] H. Jabbar and R. Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)," *Computer Science, Communication and Instrumentation Devices*, vol. 70, no. 10.3850, pp. 978–981, 2015.
- [28] M. Banoula. (2025). Everything you need to know about bias and variance. [Online]. Available: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/bias-and-variance>
- [29] Amazon. (2025). What is overfitting? [Online]. Available: <https://aws.amazon.com/what-is/overfitting/>
- [30] A. Deshmukh and S. B. Wankhade, "Deepfake detection approaches using deep learning: A systematic review," in *Proc. Intelligent Computing and Networking*, 2020, pp. 293–302.
- [31] M. Albahar and J. Almalki, "Deepfakes: Threats and countermeasures systematic review," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 22, pp. 3242–3250, 2019.
- [32] V. K. Sharma, R. Garg, and Q. Caudron, "A systematic literature review on deepfake detection techniques," *Multimedia Tools and Applications*, vol. 84, no. 20, pp. 22187–22229, 2025.

- [33] D. Gamage, K. Sasahara, and J. Chen, "The emergence of deepfakes and its societal implications: A systematic review," *TTO*, pp. 28–39, 2021.
- [34] M. S. Rana, M. N. Nobil, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.
- [35] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "Deepfake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.
- [36] L. Stroebel, M. Llewellyn, T. Hartley, T. S. Ip, and M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," *Journal of Cyber Security Technology*, vol. 7, no. 2, pp. 83–113, 2023.
- [37] T. Rajput and B. Arora, "A systematic review of deepfake detection using learning techniques and vision transformer," in *Proc. International Conference on Cognitive Computing and Cyber Physical Systems*, 2023, pp. 217–235.
- [38] N. U. R. Ahmed *et al.*, "From brain scans to classifiers: A systematic review of ML-based autism diagnostic frameworks," *Digital Health*, vol. 12, 2026.
- [39] L. A. Passos, D. Jodas, K. A. 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 Systems*, vol. 41, no. 8, e13570, 2024.
- [40] G. L. E. Team. (2025). Introduction to image pre-processing—What is image pre-processing? Available: <https://www.mygreatlearning.com/blog/introduction-to-image-pre-processing/>
- [41] Geometric transformation in image processing. (2025). [Online]. Available: <https://www.geeksforgeeks.org/geometric-transformation-in-image-processing-1/>
- [42] K. R. Castleman, "Geometric transformations," in *Proc. Microscope Image Processing*. Elsevier, 2023, pp. 47–54.
- [43] R. Alsin, A. Naz, H. U. Khan, A. Bukhari, A. Daud, and M. Ramzan, "Using deep learning and word embeddings for predicting human agreeableness behavior," *Scientific Reports*, vol. 14, no. 1, 29875, 2024.
- [44] Y. Wang, T. L. Ainsworth, and J. S. Lee, "On the geometrical dependency of the polarimetric bistatic SAR observation," in *Proc. 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 4935–4938.
- [45] A. Badshah *et al.*, "Towards smart education through internet of things: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–33, 2023.
- [46] E. W. Weisstein. (2004). Affine transformation. [Online]. Available: <https://mathworld.wolfram.com/>
- [47] A. Badshah *et al.*, "Orchestrating model to improve utilization of IaaS environment for sustainable revenue," *Sustainable Energy Technologies and Assessments*, vol. 57, 103228, 2023.
- [48] H. Mo and G. Zhao, "RIC-CNN: Rotation-invariant coordinate convolutional neural network," *Pattern Recognition*, vol. 146, 109994, 2024.
- [49] R. A. Abbasi, O. Maqbool, M. Mushtaq, N. R. Aljohani, A. Daud, J. S. Alowibdi, and B. Shahzad, "Saving lives using social media: analysis of the role of twitter for personal blood donation requests and dissemination," *Telematics and Informatics*, vol. 35, no. 4, pp. 892–912, 2018.
- [50] N. G. Pratiwi, Y. Nabila, R. Fiqraini, and A. W. Setiawan, "Effect of CT-scan image resizing, enhancement and normalization on accuracy of covid-19 detection," in *Proc. 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2021, pp. 17–22.
- [51] D. Danon, M. Arar, D. Cohen-Or, and A. Shamir, "Image resizing by reconstruction from deep features," *Computational Visual Media*, vol. 7, no. 4, pp. 453–466, 2021.
- [52] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2021.
- [53] A. Daud, N. ul Islam, X. Li, I. Razzak, and M. K. Hayat, "Identifying rising stars via supervised machine learning," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2374–2384, 2022.
- [54] P. Coscia, A. Genovese, F. Scotti, and V. Piuri, "Applications and limits of image-to-image translation models," in *Proc. 2023 24th International Conference on Digital Signal Processing (DSP)*, 2023, pp. 1–5.
- [55] Welcome to Corel PaintShop pro 2018 help. (2025). [Online]. Available: <https://help.corel.com/paintshop-pro/v20/main/en/documentation/>
- [56] I. Ito, "Flipping data augmentation of convolutional neural networks using discrete cosine transforms," in *Proc. 2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1501–1505.
- [57] Y. Xu, W. Xu, M. Wang, L. Li, G. Sang, P. Wei, and L. Zhu, "Saliency aware image cropping with latent region pair," *Expert Systems with Applications*, vol. 171, 114596, 2021.
- [58] S. Kakaei, "Automatic image cropping and semantic object selection," *Journal of Applied Intelligent Systems and Information Sciences*, vol. 2, no. 2, pp. 22–32, 2021.
- [59] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I. E. Davidson, and T. F. Mazibuko, "An improved dense cnn architecture for deepfake image detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023.
- [60] S. Khan, J. C. Chen, W. H. Liao, and C. S. Chen, "Adversarially robust deepfake detection via adversarial feature similarity learning," in *Proc. International Conference on Multimedia Modeling*, 2024, pp. 503–516.
- [61] F. Iqbal, A. Abbasi, A. R. Javed, A. Almadhor, Z. Jalil, S. Anwar, and I. Rida, "Data augmentation-based novel deep learning method for deepfaked images detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 11, pp. 1–15, 2024.
- [62] J. Jenkins, K. Roy, and J. Shelton, "Using deep learning techniques and genetic-based feature extraction for presentation attack mitigation," *Array*, vol. 7, 100029, 2020.
- [63] A. V. Nadimpalli and A. Rattani, "On improving cross-dataset generalization of deepfake detectors," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 91–99.
- [64] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–20, 2023.
- [65] J. Sharma, S. Sharma, V. Kumar, H. S. Hussein, and H. Alshazly, "Deepfakes classification of faces using convolutional neural networks," *Traitement du Signal*, vol. 39, no. 3, pp. 1027–1037, 2022.
- [66] A. Khoramali and J.-S. Yuan, "Add: Attention-based deepfake detection approach," *Big Data and Cognitive Computing*, vol. 5, no. 4, 49, 2021.
- [67] X. Chang, J. Wu, T. Yang, and G. Feng, "Deepfake face image detection based on improved VGG convolutional neural network," in *Proc. 2020 39th Chinese Control Conference (CCC)*, 2020, pp. 7252–7256.
- [68] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8060–8069.
- [69] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, "Swapped face detection using deep learning and subjective assessment," *EURASIP Journal on Information Security*, vol. 2020, pp. 1–12, 2020.
- [70] Y.-H. Han, T.-M. Huang, K.-L. Hua, and J.-C. Chen, "Towards more general video-based deepfake detection through facial component guided adaptation for foundation model," in *Proc. Computer Vision and Pattern Recognition Conference*, 2025, pp. 22995–23005.
- [71] Y. Yang, Z. Qian, Y. Zhu, O. Russakovsky, and Y. Wu, "D³: Scaling up deepfake detection by learning from discrepancy," in *Proc. Computer Vision and Pattern Recognition Conference*, 2025, pp. 23850–23859.
- [72] A. Ajoy, C. U. Mahindrakar, D. Gowrish, and A. Vinay, "Deepfake detection using a frame based approach involving CNN," in *Proc. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 1329–1333.
- [73] H. Jeon, Y. Bang, and S. S. Woo, "Fdfnet: Facing off fake images using fake detection fine-tuning network," in *Proc. IFIP International Conference on ICT Systems Security and Privacy Protection*, 2020, pp. 416–430.
- [74] W. Bian, "Real-fake face detection based on joint multi-layer CNN structure and data augmentation," in *Proc. 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, 2024, pp. 25–29.

- [75] T. DeVries, "Improved regularization of convolutional neural networks with cutout," arXiv preprint, arXiv:1708.04552, 2017.
- [76] K. Omar, R. H. Sakr, and M. F. Alrahmawy, "An ensemble of CNNs with self-attention mechanism for deepfake video detection," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2749–2765, 2024.
- [77] H. Cheng *et al.*, "Towards generalizable deepfake detection by primary region regularization," arXiv preprint, arXiv:2307.12534, 2023.
- [78] H. Bao, L. X. Dang, and L. Zhang, "Improving identity-relevant deepfake video detection in real-world with adversarial data augmentation," in *Proc. 2022 7th International Conference on Multimedia Systems and Signal Processing*, 2022, pp. 14–18.
- [79] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, "Detecting deepfake videos with temporal dropout 3DCNN," in *Proc. Thirtieth International Joint Conference on Artificial Intelligence*, 2021, pp. 1288–1294.
- [80] V.-N. Tran, S.-H. Lee, H.-S. Le, and K.-R. Kwon, "High performance deepfake video detection on CNN-based with attention target-specific regions and manual distillation extraction," *Applied Sciences*, vol. 11, no. 16, 7678, 2021.
- [81] S. Aneja and M. Nießner, "Generalized zero and few-shot transfer for facial forgery detection," arXiv preprint, arXiv:2006.11863, 2020.
- [82] T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang, "Masked conditional diffusion model for enhancing deepfake detection," arXiv preprint, arXiv:2402.00541, 2024.
- [83] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," arXiv preprint, arXiv:2102.11126, 2021.
- [84] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake detection scheme based on vision transformer and distillation," arXiv Preprint, arXiv:2104.01353, 2021.
- [85] R. Sapkota, S. Raza, M. Shoman, A. Paudel, and M. Karkee, "Multimodal large language models for image, text, and speech data augmentation: A survey," arXiv preprint, arXiv:2501.18648, 2025.
- [86] S. B. Son, S. H. Park, and Y. K. Lee, "A comprehensive study on key components of grayscale-based deepfake detection," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 18, no. 8, pp. 2230–2252, 2024.
- [87] Y. Jeong, D. Kim, S. Min, S. Joe, Y. Gwon, and J. Choi, "Bihsf: Bilateral high-pass filters for robust deepfake detection," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 48–57.
- [88] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
- [89] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6111–6121, 2021.
- [90] N. Guhagarkar, S. Desai, S. Vaishyampayan, and A. Save, "Deepfake detection techniques: A review," *VIVA-Tech International Journal for Research and Innovation IJRI*, vol. 1, no. 4, pp. 1–10, 2021.
- [91] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18720–18729.
- [92] N. Phuangsaibai, J. Jakmunee, and S. Kittiwachana, "Investigation into the predictive performance of colorimetric sensor strips using RGB, CMYK, HSV, and CIELAB coupled with various data preprocessing methods: A case study on an analysis of water quality parameters," *Journal of Analytical Science and Technology*, vol. 12, pp. 1–16, 2021.
- [93] Z. Bi and P. Cao, "Color space conversion algorithm and comparison study," in *Proc. Journal of Physics: Conference Series*, 2021, vol. 1976, 012008.
- [94] Y. Wang, S. Chen, T. Yao, L. Ma, Z. Zhang, and X. Tan, "Explore and enhance the generalization of anomaly deepfake detection," in *Proc. International Conference on Computational Visual Media*, 2024, pp. 27–47.
- [95] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *Proc. International Conference on Image Analysis and Processing*, 2022, pp. 219–229.
- [96] S. Fung, X. Lu, C. Zhang, and C.-T. Li, "Deepfakeucel: Deepfake detection via unsupervised contrastive learning," in *Proc. 2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.
- [97] Y. Lu and R. Luo, "A novel framework for assessment of learning-based detectors in realistic conditions with application to deepfake detection," arXiv preprint, arXiv:2203.11797, 2022.
- [98] S. Tariq, S. Lee, and S. S. Woo, "A convolutional LSTM based residual network for deepfake video detection," arXiv preprint, arXiv:2009.07480, 2020.
- [99] Y. Lu and T. Ebrahimi, "Assessment framework for deepfake detection in real-world situations," *EURASIP Journal on Image and Video Processing*, vol. 2024, no. 1, 6, 2024.
- [100] L. Bondi, E. D. Cannas, P. Bestagini, and S. Tubaro, "Training strategies and data augmentations in CNN-based deepfake video detection," in *Proc. 2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2020, pp. 1–6.
- [101] F. Ali *et al.*, "DBP-DeepCNN: prediction of DNA-binding proteins using wavelet-based denoising and deep learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 229, 104639, 2022.
- [102] W. Zhang, C. Zhao, and Y. Li, "A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis," *Entropy*, vol. 22, no. 2, 249, 2020.
- [103] Y. Lin, Y. Qu, Y. Li, and Z. Nie, "Exploring generalization capability for video forgery and detection based on generative adversarial network," in *Proc. 2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2020, pp. 1575–1580.
- [104] S. A. Khan, A. Artusi, and H. Dai, "Adversarially robust deepfake media detection using fused convolutional neural network predictions," arXiv preprint, arXiv:2102.05950, 2021.
- [105] S. Li, V. Dutta, X. He, and T. Matsumaru, "Deep learning based one-class detection system for fake faces generated by GAN network," *Sensors*, vol. 22, no. 20, 7767, 2022.
- [106] Z. ZHOU, J. SUN, and X. LI, "Edfm: An enhanced dual-branch fusion model for face deepfake detection," *J. Appl. Numer. Optim.*, vol. 7, no. 1, pp. 97–111, 2025.
- [107] X. Luo and Y. Wang, "Frequency-domain masking and spatial interaction for generalizable deepfake detection," *Electronics*, vol. 14, no. 7, 1302, 2025.
- [108] J. G. Pho, C. A. Gouw, H. Lucky, and D. Suhartono, "Utilizing data augmentation methods to generalize deepfake classifier," in *Proc. 2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)*, IEEE, 2024, pp. 143–148.
- [109] M. Karaköse, H. Yetiş, and M. Çeçen, "A YOLOv3-based method for detecting deepfake manipulated facial images," *Turkish Journal of Science and Technology*, vol. 19, no. 2, pp. 315–324, 2024.
- [110] J. Wang, B. Tondi, and M. Barni, "An eyes-based siamese neural network for the detection of GAN-generated fake images," *Frontiers in Signal Processing*, vol. 2, 918725, 2022.
- [111] Analytics Vidhya. (2025). [Online]. Available: <https://www.analyticsvidhya.com/blog/2022/04/master-image-augmentation-with-widely-used-python-libraries/>
- [112] A. Alharthi, M. Alaryani, and S. Kaddoura, "A comparative study of machine learning and deep learning models in binary and multiclass classification for intrusion detection systems," *Array*, vol. 26, 100406, 2025.
- [113] Aleju. (2025). Imgaug. [Online]. Available: <https://github.com/aleju/imgaug>
- [114] A. Clark *et al.*, "Pillow (PIL Fork) documentation," *Readthedocs*, 2015.
- [115] I. Culjak, D. Abram, T. Pribanic, H. Dzapov, and M. Cifrek, "A brief introduction to opencv," in *Proc. the 35th International Convention MIPRO*, IEEE, 2012, pp. 1725–1730.
- [116] F. Albardi *et al.*, "A comprehensive study on torchvision pretrained models for fine-grained inter-species classification," in *Proc. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2021, pp. 2767–2774.
- [117] A. Bakliwal and A. D. Joshi, "Deepfake detection: Leveraging the power of 2d and 3d CNN ensembles," arXiv preprint, arXiv:2310.16388, 2023.
- [118] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *Proc. 2020 25th International Conference on Pattern*

- Recognition (ICPR), IEEE, 2021, pp. 5012–5019.
- [119] Y. Lu and T. Ebrahimi, “A novel assessment framework for learning-based deepfake detectors in realistic conditions,” in *Proc. Applications of Digital Image Processing XLV*, 2022, pp. 207–217.
- [120] S. A. Khan and D.-T. Dang-Nguyen, “Hybrid transformer network for deepfake detection,” in *Proc. 19th International Conference on Content-Based Multimedia Indexing*, 2022, pp. 8–14.
- [121] P. Prajapati and C. Pollett, “Mri-gan: A generalized approach to detect deepfakes using perceptual image assessment,” arXiv preprint, arXiv:2203.00108, 2022.
- [122] X. Liu, H. Wang, Y. Zhang, F. Wu, and S. Hu, “Towards efficient data-centric robust machine learning with noise-based augmentation,” arXiv preprint, arXiv:2203.03810, 2022.
- [123] M. Arslan, M. Guzel, M. Demirci, and S. Ozdemir, “Smote and gaussian noise based sensor data augmentation,” in *Proc. 2019 4th International Conference on Computer Science and Engineering (UBMK)*, IEEE, 2019, pp. 1–5.
- [124] K. V. Thakur *et al.*, “Poisson noise reducing bilateral filter,” *Procedia Computer Science*, vol. 79, pp. 861–865, 2016.
- [125] M. Yoshimura, J. Otsuka, A. Irie, and T. Ohashi, “Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14007–14017.
- [126] J. Jaybhay and R. Shastri, “A study of speckle noise reduction filters,” *Signal & Image Processing: An International Journal (SIPIJ)*, vol. 6, no. 3, pp. 71–80, 2015.
- [127] R. Racine, G. A. Walker, D. Nadeau, R. Doyon, and C. Marois, “Speckle noise and the detection of faint companions,” *Publications of the Astronomical Society of the Pacific*, vol. 111, no. 759, 587, 1999.
- [128] D.-C. Stanciu and B. Ionescu, “Autoencoder-based data augmentation for deepfake detection,” in *Proc. 2nd ACM International Workshop on Multimedia AI against Disinformation*, 2023, pp. 19–27.
- [129] D. S. Abdelminaam, N. Sherif, Z. Ayman, M. Mohamed, and M. Hazem, “Deepfakedg: A deep learning approach for deep fake detection and generation,” *Journal of Computing and Communication*, vol. 2, no. 2, pp. 31–37, 2023.
- [130] S. Waseem, S. A. R. S. Abu-Bakar, Z. Omar, B. A. Ahmed, S. Baloch, and A. Hafeezallah, “Multi-attention-based approach for deepfake face and expression swap detection and localization,” *EURASIP Journal on Image and Video Processing*, vol. 2023, no. 1, 14, 2023.
- [131] Y. Lu and T. Ebrahimi, “A new approach to improve learning-based deepfake detection in realistic conditions,” arXiv preprint, arXiv:2203.11807, 2022.
- [132] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, and G. Amato, “Cross-forgery analysis of vision transformers and CNNs for deepfake image detection,” in *Proc. 1st International Workshop on Multimedia AI against Disinformation*, 2022, pp. 52–58.
- [133] I. Ganiyusufoglu, L. M. Ngô, N. Savov, S. Karaoglu, and T. Gevers, “Spatio-temporal features for generalized detection of deepfake videos,” arXiv Preprint, arXiv:2010.11844, 2020.
- [134] W. Khan, A. Daud, F. Alotaibi, N. Aljohani, and S. Ararat, “Deep recurrent neural networks with word embeddings for urdu named entity recognition,” *ETRI Journal*, vol. 42, no. 1, pp. 90–100, 2020.
- [135] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, “Deepfake image forensics for privacy protection and authenticity using deep learning,” *Information*, vol. 16, no. 4, 270, 2025.
- [136] J. Gao *et al.*, “Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection,” *Engineering Applications of Artificial Intelligence*, vol. 133, 108450, 2024.
- [137] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, “Evading deepfake detectors via adversarial statistical consistency,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12271–12280.
- [138] N. Kumar and A. Kundu, “Cyber security focused deepfake detection system using big data,” *SN Computer Science*, vol. 5, no. 6, 752, 2024.
- [139] F. Capital. (2025). Challenges and considerations in data augmentation. [Online]. Available: <https://fastercapital.com/topics/challenges-and-considerations-in-data-augmentation.html>
- [140] A. Mumuni and F. Mumuni, “Data augmentation: A comprehensive survey of modern approaches,” *Array*, vol. 16, 100258, 2022.
- [141] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF (v2): A new dataset for deepfake forensics,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [142] B. Dolhansky *et al.*, “The Deepfake Detection Challenge (DFDC) dataset,” arXiv preprint, arXiv:2006.07397, 2020.
- [143] Y. Li *et al.*, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.
- [144] Data augmentation—advantages, challenges, and instances. 2025. [Online]. Available: <https://snapstack.cz/data-augmentation-advantages-challenges-and-instances/>
- [145] What are the challenges and limitations of data augmentation and synthesis in predictive modeling? 2025. [Online]. Available: <https://www.linkedin.com/advice/3/what-challenges-limitations-data-augmentation>
- [146] S. Hosseini and H. Seilani, “The role of agentic ai in shaping a smart future: A systematic review,” *Array*, vol. 26, 100399, 2025.
- [147] T.-H. Cheung and D.-Y. Yeung, “A survey of automated data augmentation for image classification: Learning to compose, mix, and generate,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13185–13205, 2023.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).