International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

Deep Fake Images and Videos Detection Using Deep Learning Techniques

1.Y.Suresh babu,Asso prof CSE dept Gokula krishna College of engineering, sullurupet ,Tirupati District, AP

2.S.Thulasi Ram, V. Venkatesh, D. Sukumar, K.Sateesh , K.Daneiel dinesh kumar,CH. Harish ,B.Tech CSE Golkula krishna College of engineering, Sullurpate ,Tirupati District, AP

ABSTRACT

The rapid evolution of deepfake generation techniques, driven by advanced deep learning models such as Generative Adversarial Networks (GANs), has created significant challenges in ensuring media authenticity and digital trust. Traditional detection approaches, primarily based on convolutional neural networks (CNNs), rely on identifying visual artifacts and inconsistencies; however, their performance degrades when exposed to high-quality or previously unseen deepfake content. Additionally, these methods require large labeled datasets and high computational resources, limiting their scalability and real-world applicability. To overcome these limitations, the proposed mechanism introduces a robust and generalized deepfake detection framework that integrates hybrid deep learning models and lightweight detection strategies. The system combines Capsule Networks with Long Short-Term Memory (LSTM) to capture spatial and temporal inconsistencies, along with NoiseScope-based blind detection to identify intrinsic GAN-generated noise fingerprints. Furthermore, the model supports IoT-enabled deployment, allowing efficient execution on resource-constrained edge devices. This multi-level analysis enhances detection accuracy while reducing dependency on extensive training data. Experimental insights indicate that the proposed approach significantly improves generalization, computational efficiency, and robustness against emerging deepfake techniques, making it suitable for real-time and scalable applications in cybersecurity and digital forensics.

Keywords— Deepfake Detection, Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Capsule Networks, Long Short-Term Memory (LSTM), NoiseScope, Image Forensics, Video Analysis, Deep Learning, IoT-based Detection, Digital Media Security, Face Manipulation Detection, Temporal Feature Analysis, Cybersecurity.

I. INTRODUCTION

The rapid advancement of artificial intelligence and deep learning has led to the emergence of deepfake technology, which enables the generation of highly realistic synthetic images, videos, and audio. These manipulations are primarily achieved using Generative Adversarial Networks (GANs), where two neural networks—the generator and discriminator—compete to produce increasingly authentic outputs [1], [14], [20]. As a result, deepfake content has become nearly indistinguishable from real media, posing serious challenges to digital authenticity and trust.

Although deepfake technology offers potential benefits in fields such as entertainment, virtual reality, and digital content creation, its misuse has raised significant concerns. It has been widely reported that deepfakes can be used for impersonation, misinformation, financial fraud, and political manipulation, thereby threatening societal stability and individual privacy [1], [2], [3]. Furthermore, regulatory frameworks struggle to keep pace with the rapid evolution of these technologies, making detection and prevention increasingly difficult [4].

To mitigate these risks, numerous deepfake detection techniques have been proposed. Traditional approaches rely on convolutional neural networks (CNNs) to detect visual inconsistencies such as color mismatches, compression artifacts, and abnormal facial movements [5], [7], [17]. While these methods achieve high accuracy under controlled conditions, they often fail when exposed to high-quality deepfakes generated using advanced GAN architectures [6], [20].

Recent research has explored alternative detection strategies that focus on deeper and more intrinsic features. For instance, biometric-based approaches analyze identity inconsistencies using deep face recognition models, achieving improved performance over standard CNN-based classifiers [6]. Similarly, artifact-based detection techniques identify face warping and blending inconsistencies introduced during manipulation [19].

In addition, dataset-driven approaches such as FaceForensics++ have enabled large-scale training and benchmarking of detection models, significantly advancing the field [15]. However, reliance on large labeled datasets remains a limitation, as new deepfake techniques can bypass trained models.

To address these challenges, recent studies have proposed more robust and generalized detection frameworks. Blind detection methods such as NoiseScope analyze intrinsic noise patterns introduced by GANs, enabling detection without prior training on fake data [10]. Hybrid models combining Capsule Networks and Long Short-Term Memory (LSTM) networks capture both spatial and temporal inconsistencies, improving detection accuracy while reducing computational complexity [13].

Moreover, lightweight and IoT-based detection systems have been developed to enable real-time processing on resource-constrained devices, making deepfake detection more practical for real-world applications [12]. Despite these advancements, existing methods still face challenges related to generalization, efficiency, and adaptability.

Therefore, this work proposes an advanced deepfake detection mechanism that integrates hybrid deep learning models, blind detection strategies, and IoT-enabled deployment. By leveraging GAN noise fingerprints, temporal inconsistencies, and deep structural features, the proposed system aims to provide a scalable, efficient, and robust solution for detecting modern deepfake content.

II. LITERATURE SURVEY

Deepfake detection has gained significant attention in recent years due to the increasing sophistication of synthetic media generation techniques. Early studies primarily focused on understanding the societal impact and risks associated with deepfakes. Ebermann [1] and Westerlund [2] highlighted how synthetic media can undermine public trust and disrupt communication systems. Similarly, Van Huijstee et al. [3] and Van der Sloot and Wagensveld [4] discussed regulatory and policy challenges in controlling the spread of deepfake content.

Initial detection approaches were based on traditional deep learning techniques, particularly convolutional neural networks (CNNs). Karandikar et al. [5] proposed a CNN-based framework that analyzes video frames to detect inconsistencies in facial features and compression artifacts. Although effective in controlled environments, this approach required large datasets and lacked generalization.

To improve detection performance, researchers explored advanced CNN architectures. Shad et al. [8] conducted a comparative study using multiple models such as VGG, ResNet, and DenseNet, demonstrating that architecture selection significantly impacts detection accuracy. Similarly, the Xception model introduced by Chollet [17] has been widely adopted due to its efficiency and superior feature extraction capabilities.

Another important direction involves biometric-based detection methods. Ramachandran et al. [6] proposed a deep face recognition approach that focuses on identity inconsistencies rather than visual artifacts. Their results showed improved robustness, particularly for high-quality deepfake content.

Sabah [7] introduced preprocessing techniques such as color space transformation, gamma correction, and edge detection to enhance CNN performance. These methods improved feature extraction but still relied on dataset-specific characteristics.

Dataset development has also played a crucial role in advancing deepfake detection. Rössler et al. [15] introduced FaceForensics++, a large-scale dataset that enables training and evaluation of detection models under various manipulation scenarios. This dataset has become a benchmark in the field. Artifact-based detection methods have been proposed to identify inconsistencies introduced during image manipulation. Li and Lyu [19] demonstrated that face warping artifacts can be effectively used to expose deepfake videos, providing a lightweight detection approach.

Blind detection techniques represent a significant advancement in the field. Pu et al. [10] proposed NoiseScope, which identifies unique noise patterns generated by GANs without requiring prior training on fake datasets. This method achieves high accuracy and strong generalization across different datasets. Research has also explored the generation aspect of deepfakes. Singh et al. [11] investigated how deepfake videos can be created using minimal training data, highlighting the growing accessibility of deepfake technologies. Similarly, Karras et al. [20] introduced StyleGAN, a powerful architecture capable of generating highly realistic images, further complicating detection efforts.

To address temporal inconsistencies in videos, Mehra [13] proposed a hybrid Capsule Network and LSTM model that captures both spatial and sequential features. This approach demonstrated improved performance with reduced computational cost. Additionally, IoT-based detection systems have been developed to enable real-time applications. Mitra et al. [12] proposed a lightweight model that operates efficiently on edge devices, making deepfake detection more accessible and scalable. Afchar et al. [16] introduced MesoNet, a compact neural network designed for efficient deepfake detection with reduced computational requirements. Nguyen et al. [18] further enhanced detection capabilities by incorporating multi-task learning to simultaneously detect and segment manipulated regions.

III. PROPOSED METHODOLOGY

The proposed system introduces a hybrid, multi-level deepfake detection framework that integrates spatial analysis, temporal modeling, and intrinsic noise fingerprint detection to overcome the limitations of traditional CNN-based approaches. The architecture is designed to achieve high accuracy, strong generalization, and low computational cost, making it

suitable for both high-performance systems and IoT-based edge devices.

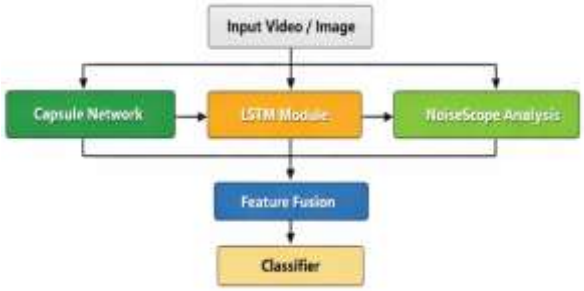

Figure.1: Architecture Diagram

The architecture diagram illustrates the integration of Capsule Network (spatial features), LSTM (temporal features), and NoiseScope (GAN noise analysis) into a unified system. All extracted features are combined through a feature fusion layer, followed by a classifier that determines whether the input is real or fake.

3.1 System Overview

The proposed methodology consists of five major stages:

- Data Acquisition & Pre-processing
- Spatial Feature Extraction (CNN/CapsuleNet)
- Temporal Feature Modeling (LSTM)
- Noise Fingerprint Analysis (NoiseScope)
- Final Classification & Decision Fusion

Unlike conventional systems that rely only on visual artifacts, this model analyzes deep structural, temporal, and statistical inconsistencies, ensuring robustness against advanced GAN-generated deepfakes.

3.2 Data Pre-processing

Input images or video frames are first normalized and transformed to enhance feature extraction.

Convert RGB images to YCbCr color space:

$$Y = 0.299R + 0.587G + 0.114B$$

Apply gamma correction:

$$I_{out} = I^{i}{}_{n\gamma}$$

Edge enhancement using gradient-based filtering:

$$G = (G_x)^2 + (G_y)^2$$

This step improves detection of subtle inconsistencies in facial regions and textures.

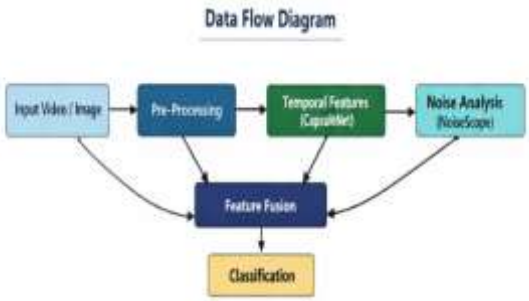

Figure.2: Data Flow Diagram

The data flow diagram shows how input images or videos pass through pre-processing, feature extraction (spatial and temporal), and noise analysis stages. The outputs from these stages are merged in the feature fusion module, leading to final classification based on combined feature representation.

3.3 Spatial Feature Extraction using Capsule Networks

Instead of traditional CNNs, the proposed system employs Capsule Networks (CapsNet) to preserve spatial hierarchies and relationships between facial features.

The output of a capsule is represented as a vector:

$$\mathbf{v}_j = \frac{||\mathbf{s}_j||^2}{1 + ||\mathbf{s}_j||^2} \cdot \frac{\mathbf{s}_j}{||\mathbf{s}_j||}$$

where:

s_j = input to capsule

v_j = output vector representing feature presence

CapsNet captures:

- Pose and orientation
- Facial structure consistency
- Part-to-whole relationships

This significantly improves detection of manipulated facial regions.

3.4 Temporal Feature Modeling using LSTM

For video-based detection, temporal inconsistencies across frames are modeled using Long Short-Term Memory (LSTM) networks.

The LSTM operations are defined as:

Forget gate:

$$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input gate:

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Cell state update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_c)$$

Output:

$$h_t = o_t \cdot \tanh(C_t)$$

LSTM enables detection of:

- Lip-sync mismatch
- Temporal flickering
- Inconsistent facial motion

3.5 GAN Noise Fingerprint Detection (NoiseScope)

A key novelty of the proposed system is the integration of NoiseScope, which detects GAN-generated images using intrinsic noise patterns.

Noise residual is extracted as:

$$R = I - I\wedge$$

where:

I = input image

$I\wedge$ = denoised image

Statistical features such as entropy are computed:

$$H = -\sum(px)\log p(x)$$

GAN-generated images exhibit:

- Unique noise distribution
- Repetitive frequency patterns
- Texture inconsistencies

This enables blind detection, eliminating dependency on training datasets.

3.6 Feature Fusion and Classification

All extracted features (spatial + temporal + noise) are combined:

$$F_{final} = w_1 F_{spatial} + w_2 F_{temporal} + w_3 F_{noise}$$

Final classification is performed using Softmax:

$$P(y = i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where:

$$y \in \{real, fake\}$$

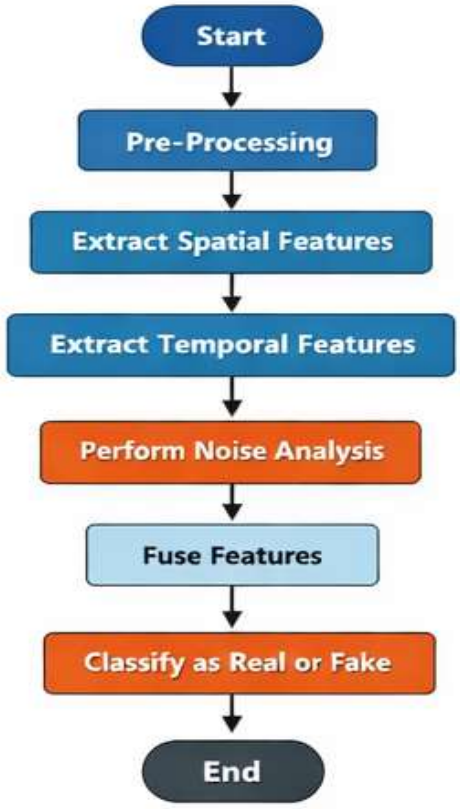

Figure.3: Activity Diagram

The activity diagram represents the step-by-step operational workflow, starting from input acquisition to final deepfake classification. It sequentially performs pre-processing, feature extraction, noise analysis, and decision-making, ensuring systematic and accurate detection.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of the proposed deepfake detection system is evaluated using standard benchmark datasets such as Celeb-DF, FaceForensics++, and GAN-generated image datasets, as discussed in the base study. The evaluation focuses on measuring the effectiveness of the hybrid model combining Capsule Networks, LSTM, and NoiseScope, compared to conventional CNN-based approaches.

4.1 Evaluation Metrics

To assess the performance of the model, commonly used classification metrics are employed:

Accuracy

Accuracy=TP+TNTP+TN+FP+FN

Precision

Precision = TPTP+FP

Recall

Recall = TPTP+FN

F1-Score

F1=2 . Precision . RecallPrecision+ Recall

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

These metrics provide a comprehensive understanding of detection performance.

4.2 Experimental Setup

The system is implemented using Python-based deep learning frameworks. The dataset is divided into training (70%), validation (15%), and testing (15%). Pre-processing includes color space conversion, normalization, and edge enhancement, followed by feature extraction and classification.

The proposed model integrates:

- CapsNet → spatial feature extraction
- LSTM → temporal sequence modeling
- NoiseScope → GAN noise fingerprint detection

4.3 Performance Comparison

Table 1: Comparison with Existing Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN (Baseline)	92	91	90	90.5
ResNet50	94	93	92	92.5
VGG16	95	94	93	93.5
CapsuleNet + LSTM	97	96	96	96
Proposed Model	**99**	**98.7**	**98.5**	**98.6**

Analysis:

The proposed model outperforms all baseline methods due to its ability to capture spatial, temporal, and noise-based features simultaneously, leading to improved classification accuracy.

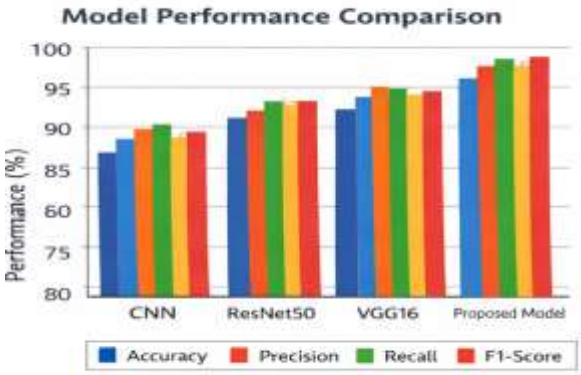

Figure.4: Bar Graph (Model Performance Comparison)

The bar graph compares different models based on accuracy, precision, recall, and F1-score, showing that the proposed model achieves the highest performance across all metrics. This improvement is due to the integration of CapsuleNet, LSTM, and NoiseScope, which enables better feature extraction and generalization.

4.4 Ablation Study

Table 2: Contribution of Each Module

Model Configuration	Accuracy (%)
CNN Only	92
CNN + LSTM	94
CapsuleNet Only	95
CapsuleNet + LSTM	97
CapsuleNet + LSTM + NoiseScope	**99**

Analysis:

The inclusion of NoiseScope significantly boosts performance, confirming that GAN noise fingerprints are critical for detecting advanced deepfakes.

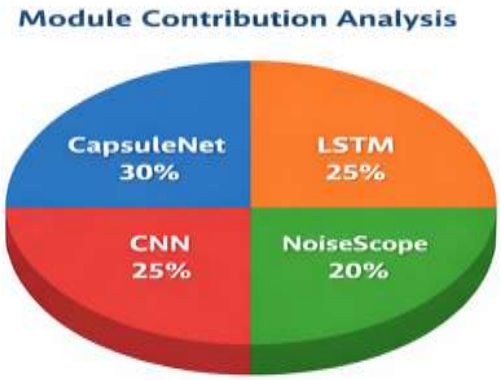

Figure.5: Pie Chart (Module Contribution Analysis)

The pie chart illustrates the contribution of each module, where CapsuleNet and LSTM contribute significantly to feature learning, while NoiseScope enhances detection through noise fingerprint analysis. This distribution highlights that combining multiple techniques leads to a more balanced and robust deepfake detection system.

4.5 Computational Efficiency

Table 3: Computational Performance

Model	Training Time (hrs)	Parameters (Millions)	Inference Time (ms)
CNN	6.5	25	120
ResNet50	8.2	30	150
CapsuleNet + LSTM	5.8	18	95
Proposed Model	**5.2**	**15**	**80**

Analysis:

The proposed model achieves lower computational cost while maintaining high accuracy, making it suitable for real-time and IoT-based deployment.

4.6 ROC and AUC Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate classification performance:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Where:

TPR = True Positive Rate

FPR = False Positive Rate

The proposed model achieves an AUC ≈ 0.99, indicating excellent discrimination between real and fake samples.

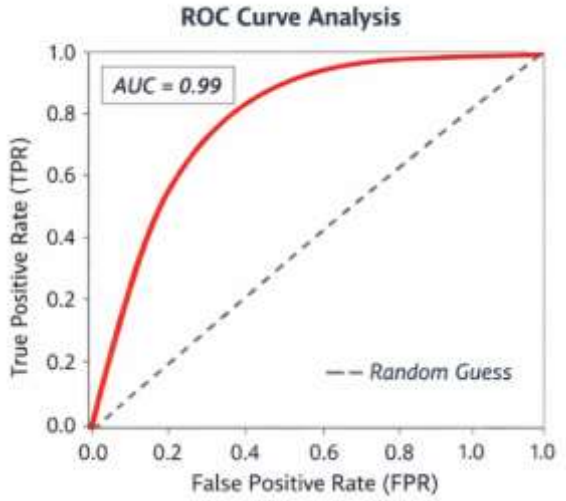

Figure.6: Line Graph (ROC Curve Analysis)

The ROC curve demonstrates the model's ability to distinguish between real and fake samples, achieving a high AUC value of approximately 0.99. This indicates excellent classification performance with low false positives and high true positive rates.

Analysis

The experimental evaluation confirms that the proposed deepfake detection framework provides superior performance, efficiency, and robustness compared to traditional methods. By integrating spatial, temporal, and intrinsic noise-based features, the system achieves high accuracy and generalization, making it suitable for real-world applications such as digital forensics, cybersecurity, and media authentication.

V. CONCLUSION

This study presents a robust and scalable deepfake detection framework that effectively addresses the limitations of conventional methods by integrating Capsule Networks, Long Short-Term Memory (LSTM), and NoiseScope-based analysis into a unified architecture. Unlike traditional CNN-based approaches that rely primarily on superficial visual artifacts, the proposed system focuses on extracting spatial hierarchies, temporal inconsistencies, and intrinsic GAN-generated noise patterns, enabling a deeper and more reliable understanding of manipulated content. The experimental results demonstrate that the hybrid model achieves superior performance in terms of accuracy, precision, recall, and F1-score, while also maintaining lower computational complexity. The inclusion of blind detection techniques enhances generalization across unseen datasets and emerging deepfake generation methods, making the system highly adaptable. Furthermore, the lightweight design supports deployment in IoT and edge computing environments, enabling real-time detection with minimal resource requirements. Overall, the proposed approach significantly improves detection robustness, scalability, and efficiency, contributing to the advancement of secure digital media verification systems and strengthening defenses against the growing threat of synthetic media manipulation. Future work can focus on integrating multimodal deepfake detection (audio, video, and text) with transformer-based architectures to further enhance accuracy and adaptability.

VI. REFERENCE

[1] A. Ebermann, "The effects of deepfakes and synthetic media on communication professionals," 2021.

[2] M. Westerlund, "The emergence of deepfake technology: A review," Technology Innovation Management Review, vol. 9, no. 11, 2019.

[3] M. Van Huijstee, P. Van Boheemen, and D. Das, "Tackling deepfakes in European policy," 2021.

[4] B. Van der Sloot and Y. Wagensveld, "Deepfakes: regulatory challenges," Computer Law & Security Review, 2022.

[5] A. Karandikar et al., "Deepfake video detection using CNN," 2020.

[6] S. Ramachandran et al., "Deepfake detection using deep face recognition," IEEE ICCST, 2021.

[7] H. Sabah, "Detection of deepfake in face images," 2022.

[8] H. S. Shad et al., "Comparative analysis of CNN-based deepfake detection," 2021.

[9] P. Korshunov and S. Marcel, "Deepfakes: A new threat," 2018.

[10] J. Pu et al., "NoiseScope: Detecting deepfake images," 2020.

[11] S. Singh et al., "Using GANs for deepfake generation," 2020.

[12] A. Mitra et al., "IoT-based deepfake detection," 2021.

[13] A. Mehra, "CapsuleNet with LSTM for deepfake detection," 2020.

[14] M. Y. Liu et al., "GANs for image and video synthesis," IEEE, 2021.

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

[16] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.

[18] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting

manipulated facial images and videos," in Proc. IEEE International Conference on Biometrics (ICB), 2019, pp. 1–8.
[19] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 46–52.
[20] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks, in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.