
Ambiguity-Aware Deepfake Image Detection Using StyleGAN2-ADA and CNN with Memory-Guided Recall

Ashrita Joshi¹, Satish Bhadane², Raviraj Bhakare³, Aakanksha Bodke⁴,
Chaitali Bonde⁵, Neha Chaudhari⁶

¹Asst. Prof., Computer Engineering, PVGCOE & SSDIOM, Nashik, India,
ashrita.joshi@pvgcoenashik.org

²Asst. Prof., Information Technology, PVGCOE & SSDIOM, Nashik, India,
satish.bhadane@pvgcoenashik.org

³Computer Engineering, PVGCOE & SSDIOM, Nashik, India, ravirajbhakare73@gmail.com

⁴Computer Engineering, PVGCOE & SSDIOM, Nashik, India,
aakankshabodke19@gmail.com

⁵Computer Engineering, PVGCOE & SSDIOM, Nashik, India, chaitalibonde1007@gmail.com

⁶Computer Engineering, PVGCOE & SSDIOM, Nashik, India, cneha2404@gmail.com

Abstract.

With deepfakes being generated using GANs, highly realistic fake images have been created, making it difficult to distinguish them from real images. The use of StyleGAN2-ADA further minimizes visual differences, increasing the complexity of detection. Although convolutional neural networks (CNNs) are widely used for deepfake detection, their performance degrades when differences between real and fake images are minimal. In this paper, we propose an ambiguity-aware deepfake detection model that integrates CNN-based classification with a memory-guided recall mechanism. When prediction confidence is low, previously learned feature representations are recalled to support decision-making. The proposed approach focuses on improving reliability and consistency rather than only maximizing accuracy. This method is particularly useful for real-world applications such as media verification, digital forensics, and cybersecurity analysis.

Keywords. Deepfake Detection; StyleGAN2-ADA; Convolutional Neural Networks; Memory-Guided Recall; Ambiguity-Aware Classification

1. INTRODUCTION

At present, deepfakes have advanced to a level where it is possible to generate synthetic images that appear highly similar to actual visual content. GAN-based models like StyleGAN2-ADA generate very realistic images, making their detection increasingly difficult. In applications requiring authentic images, there is the need for better ways to detect deepfakes. Past literature has shown the fast development in deepfake generation and also the current limitations of detectors in detecting difficult deepfake instances [4], [8]. CNNs are commonly employed for deepfake detection since they learn image features efficiently [1]. The CNN performs well when clear distinctions exist between real and deepfake images. However, as technologies in

image generation advance, it becomes increasingly hard to distinguish between real and fake images. Even though recent studies have examined advancements and improvements in models and methodologies, it has proved difficult to overcome such problems [2], [3]. With the aim of resolving this problem, an approach for detecting deepfakes has been presented based on uncertainty in predictions. The chance of misclassification is therefore minimized and reliability is achieved. The technique can also prove to be helpful when trying to tackle complicated borderline cases, whereby the distinction between the two classes becomes extremely minimal.

2. BACKGROUND AND RELATED WORK

Many Convolutional Neural Network (CNN)-based models have been developed because of the ability of CNNs to learn discriminative visual features automatically. These models perform effectively when clear differences exist between real and deepfake images. CNN-based detection methods have shown strong performance in identifying manipulated content by learning spatial and texture-based features from images [1]. However, their effectiveness depends largely on the presence of distinguishable features between real and fake samples. Several studies have attempted to improve deepfake detection capability using hybrid deep learning models involving generative models and advanced feature extraction techniques [2], [3], [6]. Some approaches make use of GAN-based features, frequency domain analysis, and structural inconsistency detection to improve classification performance. Although these methods improve accuracy, they still face challenges when dealing with images that have only minor visual differences.

3. METHODOLOGY

3.1. Overview of the Proposed Framework

The model under consideration involves the use of a CNN for detecting deepfake images, as well as a memory module to overcome the cases of uncertainty in the prediction. The entire process works in stages starting from the preprocessing stage, followed by the feature extraction and classification stages, and ending with dealing with uncertainties in the predictions. The overall workflow is shown in Figure 1.

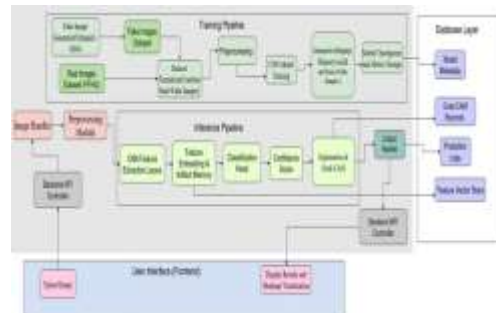


Figure 2.1. Proposed deepfake detection system architecture

3.2. Dataset Description

The dataset utilized in the present experiment comprises real and synthesized faces. Real images are sampled from the FFHQ collection; their number is about 50,000

images of 256x256px size. The generated images are produced by StyleGAN2-ADA, yielding another collection of 50,000 fake images. The total number of images is close to 100,000. The data is split in an 80:20 proportion for training and validation purposes. When training, data goes through random sampling in batch size; thus, during each epoch, the model will not encounter all of the images. As a result, after a few iterations, the model will have seen an estimated total of 40,000 to 50,000 images. There are around 20,000 images that can be found in the validation set, which serves as a means for evaluating the model’s performance.

3.3. Data Preprocessing

Each image is resized into 96x96 pixels prior to training. Pixel intensity is normalized using the [0,1] range; in other words, the images are divided by 255. In order to enhance generalization, some basic augmentations are performed, such as horizontal flips and slight rotation.

3.4. CNN-Based Feature Extraction

Convolutional Neural Network is used for extracting features from the images and classifying them into two classes. There are three convolutional layers with filters 32, 64, and 128 respectively, which are followed by max pooling. The feature maps are then flattened and fed to a dense layer with 128 nodes. Afterward, a dropout layer with a probability of 0.5 is applied to avoid overfitting. The final layer uses the sigmoid activation function.

3.5. Ambiguity-Aware Prediction

The output probability of this model is $p \in [0,1]$. The predicted results are treated differently based on the confidence level. The results with high confidence are used directly, while those near the decision boundary are viewed as ambiguous and sent to the memory for refining.

3.6. Memory-Guided Recall Mechanism



Figure 3.6.2.Memory-Guided Recall Mechanism for Feature Enhancement

4. EXPERIMENT RESULTS

4.1. Experimental Setup

The implementation of the proposed model was carried out using the TensorFlow/Keras framework, while the training and validation were performed using a dataset split of 80% and 20%, respectively. The training process was carried out with a batch size of 32 samples. Additionally, the trained model was tested on unseen real-world images, including AI-generated images, to evaluate its performance under practical conditions.

4.2. Performance Metrics

As demonstrated in Table 1, the proposed model shows an average accuracy of around 78% which indicates a stable and generalized model.

Table 1: Classification Report

Class	Precision	Recall	F1-Score	Support
Real (0)	0.87	0.67	0.76	5,095
Fake (1)	0.73	0.90	0.81	5,145
Accuracy	—	—	0.78	10,240

4.3. Results and Analysis

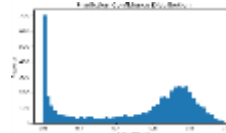
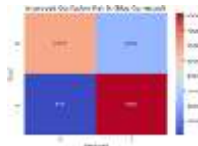


Figure 4.3.1 Confusion matrix of CNN Figure 4.3.2 Grad-CAM++ Visualization

5. CONCLUSION

Ambiguity arises when there are small variations between real and fake images. In this regard, the current study is concerned about the problem, where an innovative approach of employing a CNN model trained from scratch by utilizing examples generated from StyleGAN2-ADA, in addition to employing a memory recall approach, is used. Results obtained show that the proposed method demonstrates stable performance in solving the assigned task despite the presence of ambiguity, where similarity of features improves decision making.

6. REFERENCES

- [1] G. Aggarwal, A. K. Srivastava, K. Jhajharia, N. V. Sharma, and G. Singh, "Detection of Deep Fake Images Using Convolutional Neural Networks," in Proc. 3rd Int. Conf. Technological Advancements in Computational Sciences (ICTACS), IEEE, Nov. 2023, pp. 1083–1087.
- [2] P. Sharma, M. Kumar, and H. K. Sharma, "GAN-CNN Ensemble: A Robust Deepfake Detection Model of Social Media Images Using Minimized Catastrophic Forgetting and Generative Replay Technique," *Procedia Computer Science*, vol. 235, pp. 948–960, 2024.
- [3] S. Dasgupta, K. Badal, S. Chittam, M. T. Alam, and K. Roy, "Attention-Enhanced CNN for High-Performance Deepfake De-tection: A Multi-Dataset Study," *IEEE Access*, vol. 13, pp. 101980–101989, Jun. 2025, doi: 10.1109/ACCESS.2025.3578343.
- [4] M. Jaiswal and R. Srivastava, "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," *Multimedia Tools and Applications*, Springer, 2024.
- [5] J. Kietzmann, I. Lee, and L. W. McCarthy, "Deepfakes: Trick or Treat? A Critical Review of the Literature on Deepfake Detection," *IEEE Transactions on Technology and Society*, vol. 3, no. 2, pp. 80–92, Jun. 2022.

- [6] S. Banerjee, P. Das, and A. Roy, "CED-DCGAN: Channel and Edge-Based Detection of Deep Fake Images Using GAN Fingerprints," *Expert Systems with Applications*, vol. 235, pp. 122–134, 2024.
- [7] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS)*, 2022, pp. 1–6.
- [8] C. Pu, H. Jiang, and Y. Liu, "A Survey on Deepfake Detection: Current Challenges and Future Trends," *Journal of Information Security and Applications*, vol. 73, p. 103405, 2023.
- [9] T. Dang-Nguyen, T. Giudice, and G. Boato, "Deepfake Detection Based on Discrepancies Between Face and Head Poses," *Multimedia Tools and Applications*, vol. 83, pp. 811–832, 2024.
- [10] H. Li, X. Yang, and S. Wang, "Exploring Frequency Domain Features for Deepfake Detection," *Pattern Recognition Letters*, vol. 168, pp. 47–54, 2023.

Biographies



Ashrita H. Joshi is an Assistant Professor at Pune Vidyarthi Griha's College of Engineering & S. S. Dhamankar Institute of Management, Nashik.



Satish N. Bhadane is an Assistant Professor at Pune Vidyarthi Griha's College of Engineering & S. S. Dhamankar Institute of Management, Nashik.



Raviraj Ramesh Bhakare holds a B.E. in Computer Engineering from Pune Vidyarthi Griha's College of Engineering & S. S. Dhamankar Institute of Management, Nashik.



Aakanksha Balasaheb Bodke received the B.E. degree in Computer Engineering from Pune Vidyarthi Griha's College of Engineering & S. S. Dhamankar Institute of Management, Nashik,



Chaitali Dileep Bonde received the B.E. degree in Computer Engineering from Pune Vidyarthi Griha's College of Engineering and Shrikrishna S Dhamankar Institute of Management, Nashik affiliated with Savitribai Phule Pune University, Pune, India, in 2026.



Neha Atul Chaudhari received the B.E. degree in Computer Engineering from Pune Vidyarthi Griha's College of Engineering and Shrikrishna S Dhamankar Institute of Management, Nashik affiliated with Savitribai Phule Pune University, Pune, India, in 2026.