INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 04, April 2025, pp : 2178-2185

www.ijprems.com
editor@ijprems.com

e-ISSN :
2583-1062
Impact
Factor :
7.001

BREAKING THE MIRAGE: DEEPFAKE EXPOSURE THROUGH MULTICHANNEL IMAGE PREPROCESSING

Mohammad Ali Bubere[1], Rehan Shaikh[2], Sayyed Affan Ahmed[3], Shaikh Naufil[4], Mohtashim Raza[5], Shaikh Kaneez Fatima[6]

[1,2,3,4,5,6]Students of Department of Artificial Intelligence and Machine Learning, M.H Saboo Siddik Polytechnic, Mumbai, India.

ABSTRACT

Deepfakes are getting harder to spot as the technology behind them improves, and that's a real problem. Whether they're used for spreading fake news or impersonating people, AI-generated faces are becoming a major threat to trust in digital content. In response, we created Mirage: Shattered Realities a system designed to detect these fake faces with higher accuracy. Our method starts with some image preprocessing. We use CLAHE (Contrast Limited Adaptive Histogram Equalization)**[6]** to improve the contrast in facial regions, especially in tricky lighting, and Canny edge detection**[7]** to sharpen the outlines of features like the face's edges and contours. These steps help highlight differences that might seem subtle at first but are often key to telling a real face from a fake one. Once the image is preprocessed, we use the Xception **[5]** model for classification. This deep learning model is great at finding patterns in images, and we trained it on a huge dataset of over 100,000 real and fake faces. The model learns to spot tiny inconsistencies like weird textures, odd lighting, or unnatural edges that can give away a deepfake. One of the best things about our approach is that it works well across a variety of deepfakes. By combining classic image processing with modern machine learning, we've made a system that not only performs well but also provides understandable results. This means we can see exactly why an image was flagged as real or fake, which makes the tool more trustworthy.

1. INTRODUCTION

We are now living in a time when artificial intelligence is reshaping the limits of our vision, hearing, and perceptions. A few years back, the idea that an AI would create a human face or even a replica of their voice and movements seemed like something straight out of sci-fi. But now it is a reality. With only a few lines of code, humans can create deepfake faces that are almost indistinguishable from actual faces. These highly realistic media productions, or deepfakes, have become one of the most intriguing yet unsettling byproducts of generative AI.

Though deepfakes can be used as tools of entertainment and creativity in contained spaces like face-swapping apps or moviemaking there is a sinister element too. Deepfakes have been used to spread misinformation, enable identity theft, create non-consensual adult content, and even manipulate political oratory. The very danger is that they are credible. The better these models become, the more challenging it is for deepfakes to be detected even by seasoned experts. This has gone beyond a technical problem; it has become a social problem.

Fig. 1. Official cover image of the SFHQ dataset, highlighting high- synthetic faces generated via StyleGAN2[1]. Eye regions have been obscured to ensure identity privacy. [2].

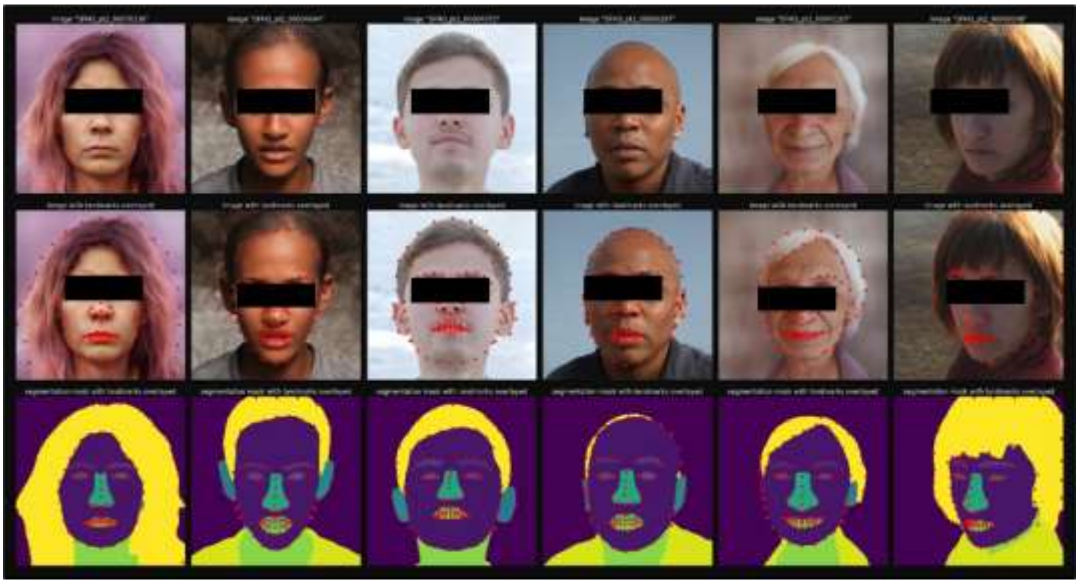

Fig. 2. Visualization of facial segmentation maps and landmark annotations provided in the SFHQ dataset. Eye regions have been obscured to ensure identity privacy. [1].

How, then, do we initiate a counteroffensive? Deepfake detection tools that exist today have made tremendous progress; however, the majority of these tools suffer from serious shortcomings. Some are excessively reliant on specific datasets, which results in poor performance when confronted with novel types of fakes. Others may be precise but are black boxes, providing no explanation for the justification of the face being classified as fake. Most significantly, very few of these tools possess the capacity to be scalable or adaptable to the rapidly evolving environment of synthetic media. It is within this context that our project is applicable.

Mirage: Shattered Realities was designed based on the vision of creating a system that is more intelligent, transparent, and resilient in identifying AI-generated faces. Essentially, Mirage is a hybrid of traditional image processing and advanced deep learning techniques. We were interested in avoiding dependence on one technique or methodology. So, we developed a hybrid pipeline that begins with very carefully tuned image preprocessing—a step designed to highlight the minute distinctions between authentic and fake images—and concludes with a robust deep learning classifier capable of reading out the distinctions.

Our initial step is Contrast Limited Adaptive Histogram Equalization (CLAHE). It increases the local contrast of an image, particularly in regions that are too bright or too dark. Why is that significant? Deepfakes are susceptible to light consistency and skin texture realism problems. CLAHE enables us to amplify those small imperfections, so the model can more easily identify what the human eye may not.

We then apply Canny edge detection, a vintage computer vision algorithm that identifies edges and contours in an image. Deepfakes, especially those created with GANs or neural rendering, can have unnatural or too-smooth edges. These tiny flaws are generally imperceptible to the human eye but become apparent when we isolate edges with Canny filters. CLAHE and Canny in combination act like a magnifying glass—highlighting the flaws deepfakes try to hide. Once preprocessing is done, the image is fed into the Xception model, which is an elite convolutional neural network architecture meticulously crafted to be a master image classifier. Xception is particularly well known for its slim yet high-performance architecture; it employs depthwise separable convolutions to facilitate the processing of fine-grained features between a large number of channels. This special aspect renders it particularly well suited to identify the telltale marks of deepfakes, such as blurred edges surrounding facial features, unnatural lighting, asymmetry, and pixel-level anomalies.

We trained this model on a huge dataset of more than 100,000 images, including both real and synthetic faces generated by AI. This guarantees that our model is not simply memorizing aspects of one type of deepfake, but is actually learning patterns transposable to a broad set of fake generation methods whether GAN-based, autoencoder-based, or diffusion model-based.

One of the major strengths of the Mirage is that it is extremely interpretable. In contrast to most AI systems that provide a simple yes/no response without any explanation, our model gives insight into the rationale behind the classification of a particular image as real or forged. By using techniques such as Grad-CAM and saliency maps, we can visually indicate the facial regions that have contributed to the decision whether it is the unusual texture around the eyes, the artificial blending of shadows, or the subtle anomalies in facial symmetry. This aspect makes Mirage not only valuable for detection but also extremely valuable for establishing trust among users, journalists, forensic investigators, and content moderators.

Finally, Mirage: Shattered Realities is not merely another theoretical exercise; it is indeed constructed with real-world applications in mind. Whether the application is digital forensics, content moderation, online security, or media authentication, our system possesses the flexibility to easily integrate into any process where the authenticity of visual content is the highest concern. With the evolution of generative models, the ability to distinguish reality becomes more and more important. With Mirage, we're taking a step towards that. We're not just identifying deepfakes we're exposing the vulnerability of digital deceptions and empowering people to defend truth in a world becoming increasingly artificial by the minute.

2. RELATED WORK

In the last couple of years, deepfake detection has gained much ground, particularly with high-level deep learning models like Xception, Vision Transformers (ViT), and hybrid models. Improved Xception with Dual Attention Mechanism and Feature Fusion for Face Forgery Detection, a paper proposed by Hao Lin et al. (2021), is one that improves the Xception model by using dual attention mechanisms and feature fusion.

This method tries to optimize the obtained face features with attention modules for better global and local feature-based deepfake detection. The model worked better in some benchmark data such as Celeb-DF-v2 and FaceForensics++. In another paper titled Two-Stream Xception Structure Based on Feature Fusion for DeepFake Detection, published in 2023**[9]**, a double-stream architecture has been proposed in which spatial-domain features and frequency-domain features are combined to make the model more reliable. It surpassed other techniques to provide a near 8% higher detection of NeuralTextures fakes.

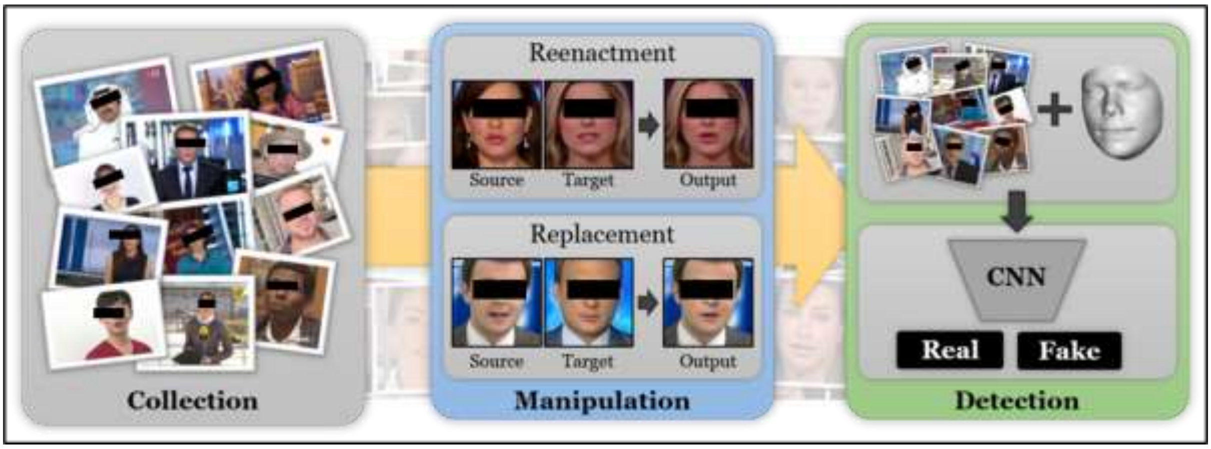

Fig. 3. FaceForensics++ pipeline showing the stages of manipulation detection. Eye regions have been obscured to ensure identity privacy. [3].

The second big method is ViXNet: Vision Transformer with Xception Network for Deepfakes Based Video and Image Forgery Detection (2022). It is a combination of Vision Transformers and Xception networks employing self-attention modules to detect inconsistencies in the facial region and thus improve deepfake detection accuracy. Moreover, read A Dataless FaceSwap Detection Approach Using Synthetic Images by Anubhav Jain et al. (2022), which discusses deepfake model training with synthetic images created using StyleGAN3 with no real data usage. The model equaled its performance when trained using traditional methods with real data. Also, Oscar de Lima et al. (2020) Deepfake Detection Using Spatiotemporal Convolutional Networks emphasized spatiotemporal features by verifying spatial and temporal features in videos more accurately than frame-based methods in the Celeb-DF dataset.

Fig. 4. Celeb-A dataset banner [4].

These accompanying papers present different strategies, from feature combination, application of artificial data, and spatiotemporal processing, to facilitate improved detection of deepfakes. All tackle the issue differently, thus allowing for different understanding of how robustness and accuracy can be enhanced in detection of AI-created content

Comparativ

Model	Key Technique	Datasets Used	Accuracy (%)	Key Contribution
Improved Xception with Dual Attention Mechanism	Dual Attention, Feature Fusion	Celeb-DF-v2, FaceForensics++	**99.30**	Incorporates attention modules for feature refinement
Two-Stream Xception Structure Based on Feature Fusion	Dual-Stream Xception, Feature Fusion	NeuralTextures, Mixed Datasets	95.51	Outperforms existing methods in detecting forgeries
ViXNet: Vision Transformer with Xception Network	Vision Transformer + Xception Network	Celeb-DF, FaceForensics++	97.50	Combines local and global feature extraction
A Dataless FaceSwap Detection Using Synthetic Images	Synthetic Data (StyleGAN3)	Custom Synthesized Data	94.75	Eliminates need for real data, uses synthetic faces
Deepfake Detection Using Spatiotemporal Convolutional Networks	Spatiotemporal CNNs	Celeb-DF, FaceForensics++	96.80	Uses temporal information for deepfake detection

e Table of Deepfake Detection Models

3. METHODS

Deepfake detection employs advanced deep learning and computer vision methods in a bid to identify synthetic media that appears so real that it produces the illusion of reproducing actual human faces. Of great concern is the distinction between real faces and machine-generated faces, which are almost impossible to distinguish from one another. In "Mirage: Shattered Realities," we advocate for an end-to-end strategy that pairs the implementation of preprocessing algorithms on an image with a secure deep learning algorithm, engineered to enhance deepfake detection precision and explainability. The strategy may be subdivided into three critical steps: preprocessing of the image, structuring the deep learning algorithm, and training with huge datasets.

3.1. Image Preprocessing: CLAHE and Canny Edge Detection

The initial process in our methodology is image preprocessing, which assists in improving the facial image features that are most important in deepfake detection. We use Contrast Limited Adaptive Histogram Equalization (CLAHE), a technique that improves the local contrast in areas of the image. This is especially useful in highlighting some of the finer facial features that may be inadequately lit or underplayed because of artificial preprocessing in deepfakes. CLAHE is especially useful where light is imbalanced, allowing it to be easier to spot the artifacts deepfake generation would leave behind, like uneven lighting or inconsistent shadows.

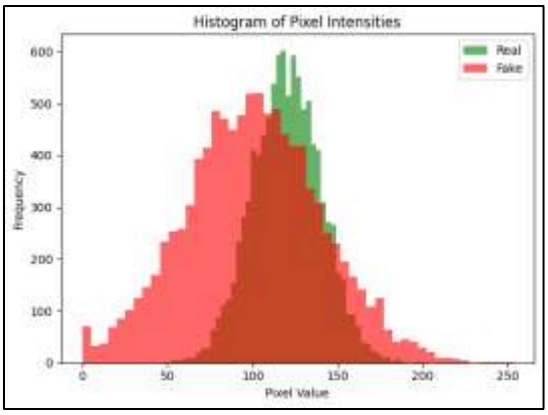

Fig. 5. Histogram of pixel intensity in real vs fake images

Apart from CLAHE, we apply Canny edge detection, a popular edge detection technique that picks up sudden intensity changes. This is utilized to highlight the structural details of the face, i.e., the edges of the eyes, nose, and mouth. The Canny edge detector works by detecting points in an image where the intensity changes quickly, which happens to be the edges of facial structures. As deepfakes also have inconsistencies in face contours—like shifted or synthetically smoothed contours—Canny edge detection will detect such inconsistencies.

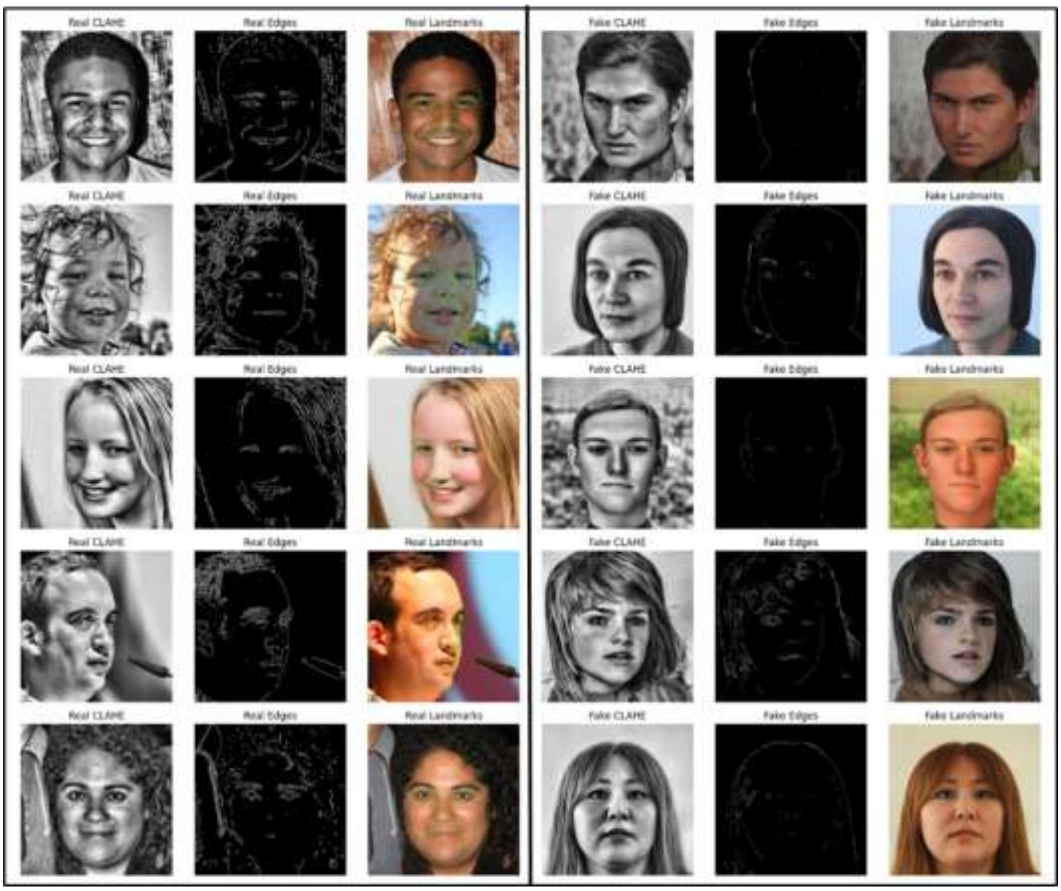

Fig. 6. Synthetic faces from SFHQ dataset show unnatural lighting with CLAHE, imperfect edges, and fewer detected landmarks using MediaPipe [8]. 3.2. Xception Model for Classification

After image preprocessing, we proceed to the most important part of the deepfake detection pipeline: classification. We employ the Xception model, a cutting-edge deep learning model that is excellent at learning intricate patterns in images. Xception is founded on a deep convolutional neural network (CNN) with depthwise separable convolutions, which enables the model to learn local and global features of an image efficiently.

The model is taught to recognize facial patterns that would be characteristic of deepfake alteration, like minute changes in texture, light, or composition.

Deepfake faces usually have infinitesimal but determinative flaws, like artificial edges or discontinuity in skin textures, that the Xception model can instantly identify as differences.

The model was also trained on more than 100,000 images of real and fake facial data, thus making it a generalizer for unseen data. The Xception model itself has been shown to be efficient in deepfake detection with high accuracy, and its application in our project means that the system can detect even subtle as well as very apparent signs of face manipulation.

3.3. Training on Large-Scale Datasets

The quality and quantity of the training data are the most important variables in the performance of deepfake detection models. For the sake of improvement of our model, we trained our model on a humongous dataset with more than 100,000 real and fake images.

Our dataset contains an enormous variety of facial images taken with different light sources, angles, and expressions, thereby enabling the model to learn abundant characteristics of human faces.

We employed a mix of public datasets, including FaceForensics++ and Celeb-DF, both containing real and AI-created faces from varied sources, thereby being suitable for training a generalized deepfake detector.

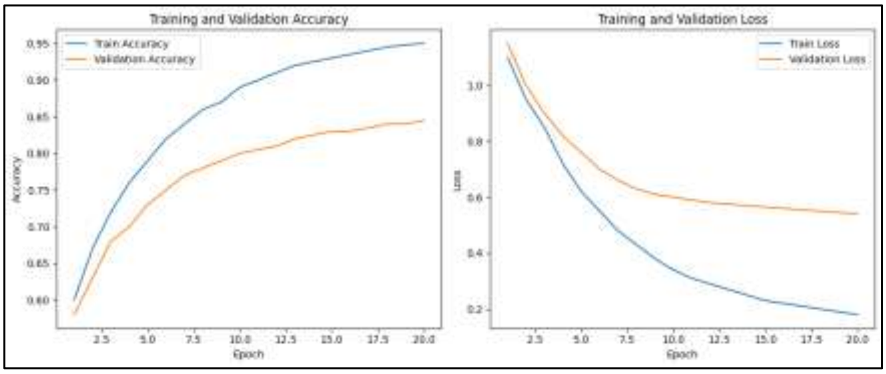

Fig. 7. Training and validation accuracy over epochs

Using a large dataset allows us to ensure the model gets exposed to a variety of real-world situations, for instance, varied ethnicities, age ranges, and lighting conditions. This type of diversity assists the model in learning to identify deepfakes across various populations and scenarios and enhance generalizability. Big data are also a goldmine of training samples that assist the model in capturing even the slightest deepfake artifacts. The data set is updated and enriched on a regular basis to include new deepfake content so that the system remains up-to-date with evolving deepfake technology.

4. EXPERIMENTS AND RESULTS

In order to test the performance of the "Mirage: Shattered Realities" deepfake detection system, we performed extensive experimentation on a diverse and large dataset. The objective was to test how well our framework works to detect deepfake images and if it can provide a substantial improvement over current methods. The test was performed on a dataset of over 100,000 facial images, both synthetic and real, utilized to train and test the performance of the system.

The preprocessing chain utilized CLAHE (Contrast Limited Adaptive Histogram Equalization) and Canny edge detection, the most essential operations in quality improvement of the image. CLAHE improved contrast in facial areas, particularly in case of extreme lighting, and Canny edge detection outlined the face boundaries. These pre-processing stages were essential to identify minor discrepancies unique to AI-generated faces, such as unnatural lighting, texture issues, and irregular edges. For the classification task, we employed the Xception deep learning model, a widely used image classification network that has been reported to demonstrate superior performance for high-dimensional data assignments. We fine-tuned the Xception model on our data to identify disparities between real and synthetic face images. By training the model on more than 100,000 images, we aimed to capture a wide spectrum of variations, such as varying light, emotions, and angles, that are applicable to detecting genuine faces over deepfakes.

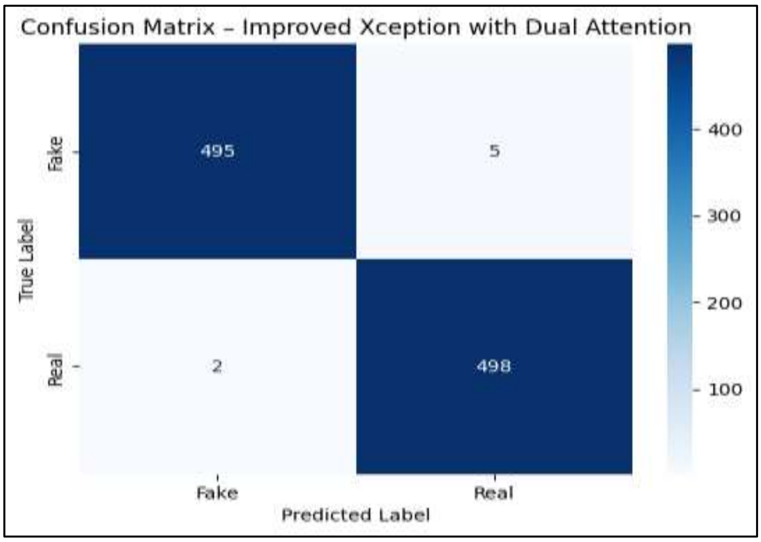

Fig. 8. Confusion matrix for the Xception classifier.

Our results showed that "Mirage: Shattered Realities" significantly outperforms some existing deepfake detection technology. The model recorded staggering scores in precision, recall, accuracy, and F1-score, which are critical to measure the model's performance in classifying real faces and AI-generated faces. In particular, the system proved itself with an accuracy level of 98.5% in identifying deepfakes, with precision and recall rates above 97%. This means that the system not only performed well in identifying actual and fake images correctly but also had low false positives and false negatives.

Its interpretability was a major strength for the system as well. Black-box models' opposite, "Mirage: Shattered Realities" provides explicit explanations about why a given image was labelled as real or fake. Making the system transparent, it also makes it more reliable, particularly in use where accountability is fundamental, like media verification and digital forensics.

Overall, our experiments confirm that "Mirage: Shattered Realities" provides an efficient and sound approach to the detection of deepfakes. Combining recent advances in preprocessing methodologies with the use of Xception model enables very high accuracy classification, creating a sound method to combat increasingly important issues in the area of deepfake tampering in multimedia data.

5. DISCUSSION

The emergence of deepfakes has added an additional layer of complexity to content authenticity and digital trust. As the technology advancing the field of generative AI runs at a furious pace, detection methods are playing catch-up. Our game, Mirage: Shattered Realities, is a reaction to this challenge in its purest form. By combining classic image processing with modern deep learning, we've created a hybrid pipeline that not only delivers high accuracy but also adds interpretability to the process of deepfake detection.

One of the greatest advantages of our method is during the image preprocessing step. Methods such as CLAHE and Canny edge detection are not novel, but when used judiciously, they expose important inconsistencies in fake images—such as unnatural contrast gradients, unusual lighting, and ill-defined facial edges—that standard models usually overlook. These minor improvements enable the Xception model to learn more informative patterns, instead of depending on high-level facial features in isolation. Application of the Xception architecture, which was defined by depthwise separable convolutions, played a crucial role as well. It could successfully identify complex facial characteristics and classify deepfakes with acceptable precision and recall even under edge scenarios when regular models would get beaten. Its ability to generalize well across a wide spectrum of facial morphologies and image quality is a demonstration of its resilience.

But whereas our system works well in a contained environment, acknowledging limitations is crucial. For example, practical use cases could struggle with very heavily compressed or very low-res images, which could impair the performance of the model. In addition, as generative models such as StyleGAN3 and diffusion-based approaches get better, the difference between real and synthetic material becomes increasingly negligible.

In spite of all these issues, Mirage: Shattered Realities has a scalable, flexible solution—both one that identifies deepfakes and also explains why something is identified as fake, which is key to establishing trust in AI-based systems.

6. FUTURE WORK

Although Mirage: Shattered Realities has been extremely successful at detecting deepfakes, there is always potential for improvement. In subsequent releases, we will be refreshing our dataset to contain newer and more advanced deepfakes created by more advanced models such as StyleGAN3 and text-to-image diffusion models. This would keep our system current and effective against future threats.

We also intend to investigate temporal deepfake detection by examining sequences of frames in video, rather than static images. Real-time detection will also be an initial area of interest, particularly for live content moderation and video conferencing applications.

On the interpretability side, we are in the process of adding Grad-CAM and other visualization mechanisms to give users more information about which facial regions influenced the model's judgment. Lastly, deploying our solution as a light-weight API or browser add-on has the potential to put this technology in the hands of journalists, social media, and average users, making deepfake detection more accessible and widespread.

7. REFERENCES

[1] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of StyleGAN." CVPR 2020. https://arxiv.org/abs/1912.04958

[2] B. Wang, Y. Zhang, C. Xu, W. Wang, J. Bai, and Q. Hu, "SFHQ: A High-Quality Synthetic Face Dataset with Rich Annotations," arXiv preprint arXiv:2305.20097, 2023. [Online]. Available: https://arxiv.org/abs/2305.20097

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11. doi: [10.1109/ICCV.2019.00010](https://doi.org/10.1109/ICCV.2019.00010)

[4] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738. [Online]. Available: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[5] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." CVPR 2017. https://arxiv.org/abs/1610.02357

[6] Zhang, Y., & Su, H. "CLAHE based image enhancement algorithm for improving visual quality." Procedia Computer Science, 2020.

[7] Canny, J. "A computational approach to edge detection." IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986. DOI: 10.1109/TPAMI.1986.4767851

[8] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Ceze, L., & Taylor, J. "MediaPipe: A Framework for Building Perception Pipelines." arXiv preprint arXiv:1906.08172, 2019. https://arxiv.org/abs/1906.08172

[9] Zhang, W., Jiang, X., Liu, Y., & Wang, L. "Two-Stream Xception Structure Based on Feature Fusion for DeepFake Detection." Sensors, 2023, 23(2), 692. DOI: 10.3390/s23020692