Beyond the Face: Illumination-Aware and Explainable Deepfake Detection

Anagha S Menon, Gayathri Venugopalan, Karthika Sreenivasan, Lavanya P V Malavika Ajith, Nithya Viju, Anusree K S
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
AM.EN.U4CSE22212, AM.EN.U4CSE22220, AM.EN.U4CSE22237, AM.EN.U4CSE22239,
AM.EN.U4CSE22242, AM.EN.U4CSE22243

***Abstract*—The extensive use of deepfake images for identity theft, misinformation, and digital manipulation present a serious problem for media authenticity. Despite the fact that many deepfake detection models exist, many models purpose are being black boxes, with no reasonings on how decisions were made. This paper introduces a hybrid Explainable AI (XAI) system that not only detects deepfakes but also highlights areas that led to model predictions. For generating the visual explanations of predictions, we used Grad-CAM on a ResNet-18. Also, we are proposing a novel module that detects inconsistencies in lighting, shadows, and reflections—artifacts which are usually ignored by traditional models. They are analysed using edge detection and contour-based techniques and we perform a comparative evaluation of other CNN models— VisionTransformers, EfficientNet, etc. The benchmark datasets we used for this are FaceForensics++ and Celeb-DF, demonstrating high accuracy and explainability. This hybrid approach increases reliability in deepfake detection by combining data-driven learning with physics-based visual reasoning.**

***Index Terms*—Deepfake detection, Explainable AI (XAI), Vision Transformers, Grad-CAM, digital forensics**

I. Introduction

Deepfake methodologies, as a result of the more recent technological advances in artificial intelligence (AI) and the field of deep learning (DL), have considerably impacted the landscape of digital media production. Look at generative adversarial networks (GANs), autoencoders, or vision transformers (ViTs) [12], [13] - these technological advancements have made the production of photorealistic images and videos a reality. In fact, these technological advancements are now a backbone to the creative industries for gaming, education, virtual reality, and the like. However, these advancements are turbulently impacting digital authenticity and public trust.

The rise of deepfake technology has made it terrifyingly easy for malevolent players to spawn verisimilitudes wherever visual media is present. These artifacts have been used for identity theft, political misinformation, reputation takedowns, and cyber deception. In fields like cybersecurity and journalism—where there is a high premium on trust in visual evidence—such deceptions can carry profound implications.

Although contemporary deepfake detectors have been able to attain high performance, most are black boxes and provide little indication of how and why they made a prediction. This lack of interpretability [13], [16] restricts their deployment in mission-critical domains such as legal forensics or news authentication. In addition, these models tend to ignore minor environmental anomalies—like unnatural illumination, missing shadows, or inconsistent reflections—that an average human would catch intuitively.

To overcome such constraints, we introduce a hybrid Explainable AI (XAI) system that not only detects deepfakes with superior accuracy but also offers visual explanations for every prediction. We adopt Gradient-weighted Class Activation Mapping (Grad-CAM) for CNN models and LIME citeb10 for transformer-based models. These mechanisms point out precise areas of an image that impacted the model response, enhancing user trust and clarity.

Moreover, we propose an environmental cue-based detection strategy that analyzes lighting direction, reflection consistency, and shadow coherence using techniques like histogram comparison, feature matching, and geometric validation. This human-like reasoning goes well with the model's predictions, making our model's detection more strong under various scenarios.

We benchmark on state-of-the-art datasets like FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge (DFDC), and show that the synergistic combination of interpretability and green awareness substantially boosts performance.

A. Key Contributions

- We propose a hybrid deepfake detection framework that combines deep learning with visual explanation methods (Grad-CAM, SHAP, and LIME) to improve model interpretability.
- We introduce a novel detection mechanism based on environmental cues—such as lighting inconsistencies, shadow alignment, and reflection symmetry—that are typically ignored by standard detection models.
- Our approach is evaluated on benchmark datasets (FaceForensics++, Celeb-DF, DFDC), showing improved accuracy and robustness in realistic and adversarial conditions.
- The framework enhances transparency and trust in AI-based detection, making it suitable for critical applications in digital forensics, journalism, and cybersecurity.

This work emphasizes not only the technical efficiency of deepfake detection models but also their interpretability and reliability in sensitive domains.

II. Related Works

Several studies have been examined to highlight the growing threat of deepfake technology, especially in the field of cybersecurity and media integrity.

Awodiji and Owoyemi [1] presented a in depth comprehensive review of advanced deepfake detection and prevention methods. The paper focused on the role played by the AI models to increase the trust worthiness of visual information and prevent any malicious manipulation. The research of their paper focus on real-time detection systems in cybersecurity.

Priyanka Kapoor [2] highlighted the social effects of deepfake technology, focusing on how it affects the political, entertainment, and individual privacy areas.This paper examined the ethical concerns and demands strong regulatory and technical measures.

Wang et al. [3] provided a survey of finding technologies and explored the technical details of the methods used in deepfake creation. A breakdown of deepfake methods and a comparison of the accuracy currently available detection algorithms are included.

George and George [4] explained the history of deepfake technology, the invention of generative models has made it possible to generate hyper-realistic content. They discussed about the risks associated with this kind of realism and how urgent it was to create preventative measures.

Thing [5] did a comparative study of transformer-based models and convolutional neural networks (CNNs) for identification of deepfakes. In spite of transformers' ability for learning complex patterns, CNNs stay competitive because of their computational efficiency, indicated in the study.

Nirkin et al. [6] presented a new method for identifying deepfakes that depends on identifying inconsistencies between the face and its surrounding area. On benchmark datasets, their approach is very precise and makes use of environmental knowledge and face boundary traits.

Masi et al. [7] used dynamic feature combination and a two-branch recurrent network to extract deepfake patterns from video. Their method takes into consideration movement and appearance adjustments, and increases detection reliability.

Verdoliva [8], had a complete examination of the media forensics technique used in deepfake detection. The work analyzes the methods for detection and shows the need for forensic reliability as an important requirement for preventing visual misinformation.

Selvaraju et al. [9] proposed Grad-CAM, an gradient-based class activation mapping method to explain the predictions of CNN-based models, . The approach increases the understanding of AI in the eXplainable AI (XAI) field by making deep fake detection models understandable.

III. Methodology

The deepfake detection method we proposed integrates a CNN-based classification model with explainability and environmental cue analysis. The stages includes data preparation, model training, Grad-CAM visualization, and shadow or reflection anomaly detection, as shown in the Figure 1.

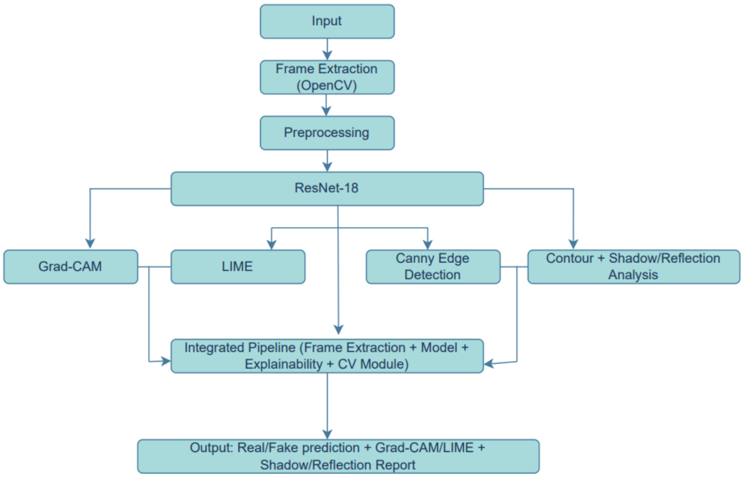
Input
Frame Extraction (OpenCV)
Preprocessing
ResNet-18
Grad-CAM
LIME
Canny Edge Detection
Contour + Shadow/Reflection Analysis
Integrated Pipeline (Frame Extraction + Model + Explainability + CV Module)
Output: Real/Fake prediction + Grad-CAM/LIME + Shadow/Reflection Report

Fig. 1: Architecture

A. Data Acquisition and Preprocessing

The videos used for our model are taken from FaceForensics++ and Celeb-DF datasets. Around 15 frames are taken from 15 different videos of the dataset which ensures that if any frame in the video is made using deepfake it is mostly likely to be a fake video . Frames are resized to 224×224 and is normalized using ImageNet statistics. Data augmentation techniques such as random flipping, rotation, and brightness adjustments are applied to improve the generalization.

B. Model Architecture

We use ResNet-18, a convolutional neural network pre-trained on ImageNet, with the ending fully connected layer modified for binary classification (in our case fake or real). The model uses residual learning to extract robust dimensional features from facial regions.

C. Training Procedure

We trained the model for 10 epochs using the Adam optimizer with a default learning rate of 0.001. Cross-entropy loss and learning rate scheduling are applied. The dataset split is 80% for training and 20% for testing. Accuracy and loss for each epochs are recorded.

D. Explainability using Grad-CAM

We uses Gradient-weighted Class Activation Mapping (Grad-CAM) to interpret the models prediction. The regions of the image which influenced the model's decision is visualized by generating heatmap from last convolutional layer by this technique.

E. Environmental Cues Analysis

We introduced a post-processing module that analyzes physical inconsistencies which are beyond the neural prediction. The system detects unnatural shadows, lighting differences, and reflection inconsistencies with Canny edge detection and contour analysis. These cues are very helpful as they often betray manipulated content, offering our model an additional forensic layer.

F. Model Comparison Strategy

For identifying the most suitable architecture for our deepfake detection model, we conducted a comparison analysis of several convolutional neural networks models. The models used included ResNet-18, Vision Transformer, EfficientNet, and a custom three-layer CNN designed for baseline comparison. Each model was fine-tuned on the same training set of extracted frames from the datasets and evaluated on similar test sets to ensure the consistency. The preprocessing pipeline and hyper parameters were kept uniform across all models to allow for a proper comparison. Performance metrics such as accuracy, precision, recall, and F1-score were computed for each model. Based on the balance of interpretability, performance, and computational efficiency, we selected ResNet-18 as the our architecture for the final explainable and illumination-aware pipeline.

IV. Implementation

This proposed deepfake detection system was implemented fully in Python, leveraging the PyTorch deep learning framework due to its modularity, GPU acceleration, and dynamic computation graph capabilities. The pipeline include five main components that includes video frame extraction, data preprocessing and augmentation, model architecture and training, explainability via Grad-CAM, and post-processing for physical cue analysis.

A. Frame Extraction and Dataset Preparation

The videos were obtained from reference datasets like FaceForensics++ and Celeb-DF. Using OpenCV, we implemented a custom frame extractor that downsampled 10 equally spaced frames per video, ensuring uniform coverage across each video's timeline. This cuts down on redundancy while enabling diverse visual cues. The retrieved frames were stored as RGB image and labeled as real or fake. For consistency and compatibility with ResNet architecture, we resized the frames to a uniform dimension of 224×224 pixels to normalize the input For consistency in lighting and colour distribution. We normalized the images with ImageNet statistics: mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]. Additionally, data augmentation was used with torchvision. It transform augmented data with random horizontal flipping, random brightness changes, and small rotation, enhancing generalization across different capturing conditions.

B. Model Architecture and Training

The model we used is ResNet-18, a CNN, pretrained on ImageNet dataset. The final fully connected (FC) layers was modified and replaced with a linear output layer containing 2 neurons for binary classification. This allowed the model to keep the visual attributes from pretraining while learning deepfake-specific cues during finetuning at the same time. In this model, the Adam optimizer was used for optimization with the learning rate of 0.001. The function which was used as the loss function was the Cross-entropy loss. The training of the model was done on 10 epochs with batch size of 32. The dataset was split as 80-20. The evaluation metrics used for the model performances were accuracy, precision, recall and F1 score.

C. Shadow and Reflection Detection via Physical Cues

We proposed a module from computer vision to detect physical inconsistencies in the fake images and this module use OpenCV's Canny edge detector to identify prominent edges in the facial regions. The resulting edge maps was further processed using contour analysis to localize specific areas such as the neck, hairline, and cheeks. Deepfakes finds it hard to maintain consistent lighting and realistic shadow behavior. The edge-based module detect missing or malformed shadows, irregular specular highlights, and blurred reflection zones. To confirm or question the model's predictions, these findings were cross-verified with Grad-CAM heatmaps and by doing so, this add an extra layer of transparency and strength to the system. All the modules— frame extraction, model training, physical anomaly detection and Grad-CAM visualization —were integrated into this framework.

D. Explainability via Grad-CAM and LIME

The proposed framework was incorporated with Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) to improve model interpretability. These techniques provide visual justifications that helps explain why the model predicted a frame as real or fake, thereby increasing transparency and trust. Grad-CAM highlight spatial areas in the image that most influenced the prediction by using gradients from the final convolutional layer of the ResNet-18 model. The resulting heatmap emphasize facial areas such as the eyes, jawline, and mouth regions where deepfakes often introduce subtle artifacts. LIME [10] complement Grad-CAM by fragmenting the image into superpixels and identifying which regions most impact the classification.

V. Results and Analysis

We checked the performances of 4 deep learning models: ResNet-18, EfficientNet, Vision Transformer (ViT), and a lightweight custom CNN. To verify fairness, these models were trained and tested on the same dataset. Table I shows the accuracy, precision, recall, and F1-score of these four models.

A. Model Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
ResNet-18	81%	0.82	0.81	0.81
EfficientNet-B0	80%	0.80	0.81	0.80
Custom CNN	75%	0.76	0.75	0.75
Vision Transformer	54%	0.56	0.54	0.53

TABLE I

Although EfficientNet showed a marginal improvement in recall, we selected ResNet-18 for deployment because of balance of performance and affinity with explainability frameworks like Grad-CAM. The model that underperformed was Vision Transformer, most likely due to its overfitting on less number of frames.

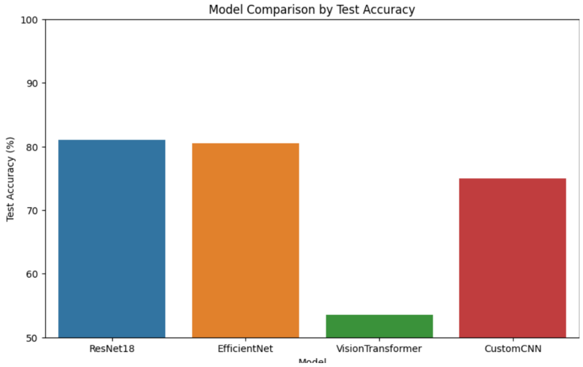

Fig. 2

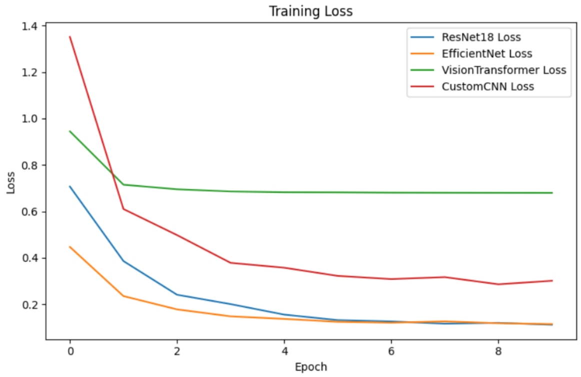

Fig. 3

B. ResNet-18 Evaluation and Analysis

Table II is a detailed explanation on the performance of ResNet-18 model. This model accomplished a total accuracy of 81%, a precision of 0.88 for real images and 0.74 for fake images. The recall for fake images was 0.81, indicating strong detection capability for manipulated frames.

Class	**Precision**	**Recall**	**F1-score**
Real	0.88	0.77	0.82
Fake	0.74	0.86	0.80
Overall Accuracy	81%		

TABLE II

The confusion matrix shown here (Figure 4) offers a deep understanding of classification errors. Most misclassifications happened in real images were marked as fake.

Given the strong performance of the ResNet-18 model, we selected this model as the foundation of our deepfake detection system.

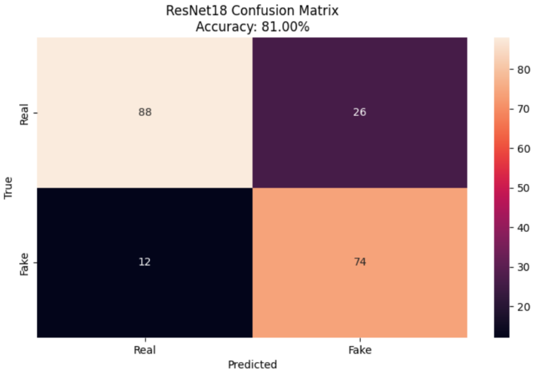

Fig. 4

C. Interpretability Insights

To give a transparent and explainable deepfake detection system, we used both model-centric and physics-aware visual explanation techniques. This hybrid interpretability strategy help verify the model's reasoning and uncover the visual inconsistencies often overlooked in manual inspection.

The original input frame that is given to the system (Figure 5), appear visually realistic at first look. But there are many mismatches that may not be obvious to the human eye like unnatural blending, missing shadows, and lighting. This is why interpretability is important.

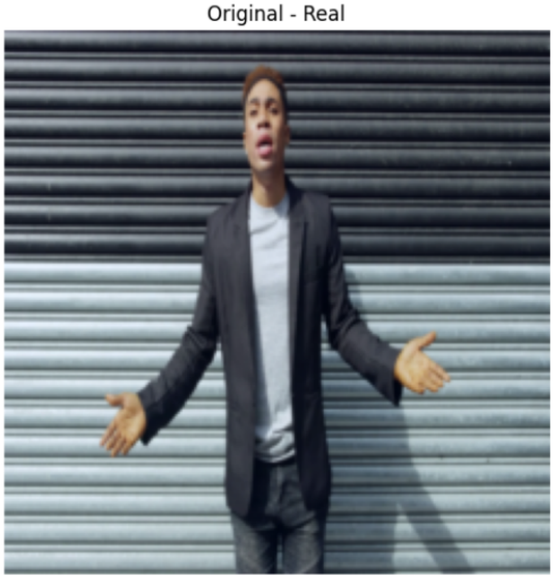

Fig. 5

Grad-CAM (Figure 6) reveal spatial areas that are most affected by the decision of the ResNet-18 model. Usually, attention is intense around the jawline, cheekbones, and eyes, areas where blending artifacts frequently appear.

LIME, shown in Figure 7, finds out the most influential areas and converts the image into superpixels. Intersection between Grad-CAM explanations and LIME strengthen model interpretability.

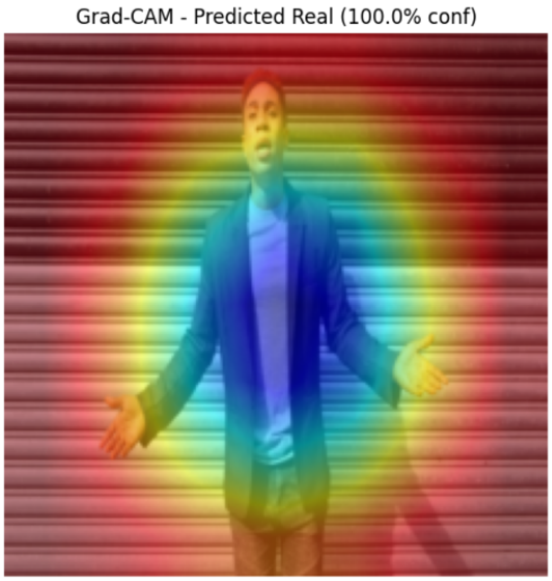

Fig. 6

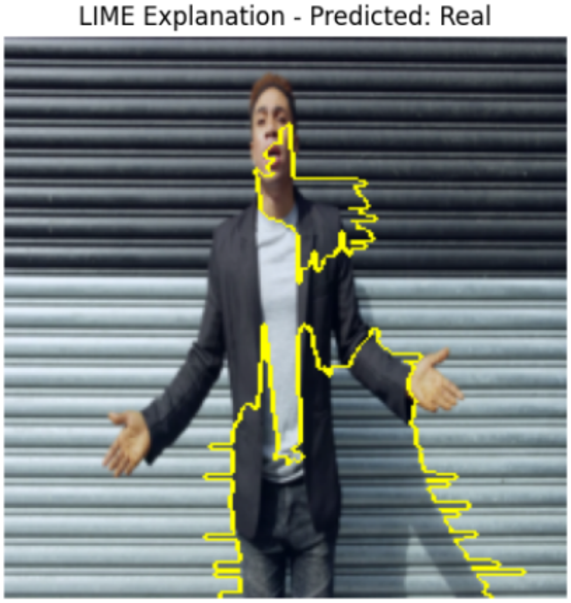

Fig. 7

Figure 8 illustrate the use of Canny edge detection and contour extraction to detect unnatural lighting. These shadow and reflection inconsistencies provides strong physical evidence for manipulation.

These explanations make the deepfake detection system more interpretable and trustworthy — important for real life applications such as forensics, journalism and digital integrity.

VI. Conclusion

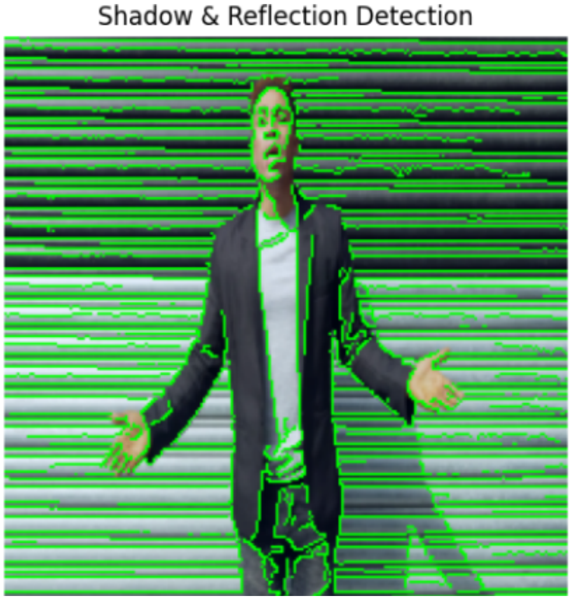

Fig. 8

Our work acknowledge the increasing challenges of deepfake detection by proposing a novel approach which combine computer vision techniques with deep learning and explainable AI (XAI). The proposed system combine the strengths of physical cue detection by focusing on the inconsistencies in shadows and reflections with a CNN classification model built on a modified ResNet18 architecture. The use of Grad-CAM provide an insight into which areas of the image influenced the model's predictions, adding transparency and interpretability to the decision-making process. This pipeline not only predict whether a frame is real or fake but also justify its decision through visual explanations. This makes it flexible for combining with more advanced models or real-time systems. Overall, this work contribute to the development of strong, explainable, and physically grounded deepfake detection systems that are essential for maintaining trust in digital media.

VII. Future Works

Even though, the proposed framework demonstrate the potential of combining physical cue analysis with deep learning and explainable AI for deepfake detection, several improvements can be made in the future.

- The anomaly detection module can be improved using more advanced computer vision techniques or integrating depth and infrared modalities to better capture inconsistencies.
- The CNN model can undergo fine-tuning with more diverse dataset and extended to handle full video sequences instead of single frames. [15].
- Incorporating transformer-based models helps capture long-range dependencies and subtle manipulation patterns.
- Real-time deployment on edge devices or browsers would also be valuable, especially in social media content verification.
- Frontend could be incorporated such that it encourages non technical users to utilize this framework.

References

[1] T. O. Awodiji and J. Owoyemi, “Advanced detection and mitigation techniques for deepfake video: Leveraging AI to safeguard visual media integrity in cybersecurity,” *Int. Adv. Res. J. Sci. Eng. Technol.*, vol. 11, no. 9, pp. 76–85, September 2024.

[2] Priyanka Kapoor, “Study on the impact of artificial intelligence enabled deepfake technology,” *Int. J. Creat. Res. Thoughts (IJCRT)*, vol. 12, no. 5, May 2024.

[3] R. Wang, Z. Huang, Z. Chen, L. Liu, J. Chen, and L. Wang, “Study on the impact of artificial intelligence enabled deepfake technology,” *arXiv preprint arXiv:2206.00477*, June 2022.

[4] A. Shaji George and A. S. Hovan George, “Deepfakes: The evolution of hyper realistic media manipulation,” *PUIRP*, vol. 1, no. 2, pp. 58, Nov.–Dec. 2023.

[5] V. L. L. Thing, “Deepfake detection with deep learning: Convolutional neural networks versus transformers,” *IEEE*, in press, doi: 10.1109/CSR57506.2023.10225004.

[6] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, “Deepfake detection based on discrepancies between faces and their context,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6111–6121, Oct. 2022.

[7] I. Masi, S. Killekar, A. Tula, C. Spampinato, and P. Natarajan, “Two-branch recurrent network for isolating deepfakes in videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 101–110.

[8] L. Verdoliva, “Media forensics and deepfakes: An overview,” *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020.

[9] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[11] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[12] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[13] A. Saha, S. Maity, D. Das, A. Ray, and S. K. Naskar, “Deepfake detection using vision transformers and CNN features fusion,” *arXiv preprint arXiv:2202.03729*, 2022.

[14] Y. Li and S. Lyu, “Face X-ray for more general face forgery detection,” *arXiv preprint arXiv:1912.13458*, 2020.

[15] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” *arXiv preprint arXiv:1905.00582*, 2019.

[16] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Learning rich features for image manipulation detection,” *arXiv preprint arXiv:1805.04096*, 2018.

[17] A. R. Anagha, A. Arya, V. Hari Narayan, S. Abhishek, and T. Anjali, “Audio deepfake detection using deep learning,” in *Proc. 2023 Int. Conf. on System Modeling and Advancement in Research Trends (SMART)*, IEEE, 2023, pp. 176–181.

[18] G. S. Prabith and T. Anjali, “Hierarchical ST-CEN with dynamic attention mechanisms for enhanced deepfake image detection,” in *Proc. 2023 Int. Conf. on Intelligent Computing, Communication and Convergence (ICI3C)*, IEEE, 2023, pp. 377–383.

[19] A. A. Thampi, R. Megha, V. B. C. Varma, S. Abhishek, and T. Anjali, “Deception detection in the digital age: Pioneering linguistics and vision,” in *Proc. 2023 Int. Conf. on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, 2023, pp. 1–6.

[20] T. Bikku, K. Bhargavi, J. Bhavitha, Y. Lalithya, and T. Vineetha, “Deep residual learning for unmasking DeepFake,” in *Proc. 2023 Int. Conf. on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS)*, IEEE, 2023.

[21] A. Siva Sankar, A. S. Akshay, K. M. Ananthakrishnan, A. G. Menon, and S. Siji Rani, “Guardians of truth: Advancing deepfake detection with transparency and insight,” in *Proc. Int. Conf. on Data Science and Applications (ICDSA)*, Springer, 2024, LNNS vol. 1237, pp. 193–205.

[22] S. Babitha, V. Sundaram, and S. Vekkot, “Enhancing deepfake detection: Leveraging deep models for video authentication,” in *Proc. 2024 Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2024.