# Attention-Guided Discrepancy Analysis for Robust Deepfake Detection

**Mrs. Nikhat Fatima , Dr. Sameena Banu**

Assistant professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Khaja Bandanawaz University

**Abstract**

The rise of deepfake technology has created significant challenges in digital media forensics, leading to concerns about misinformation, privacy breaches, and identity fraud. Existing deepfake detection methods often suffer from identity expression bias and fail to generalize across different manipulation techniques. To address these limitations, we propose Attention-Guided Discrepancy Analysis for Deepfake Detection (AGDA-DD), a novel framework that enhances deepfake detection by leveraging advanced feature analysis and discrepancy exploitation techniques. AGDA-DD follows a two-stage detection framework integrating an Adaptive Feature Representation Module (AFRM) and a Bias Reduction Mechanism (BRM) to extract unbiased identity features, improving robustness against identity-related inconsistencies. Additionally, the framework introduces Context-Aware Feature Refinement (CAFR), which dynamically enhances feature representations by focusing on critical identity distortions while reducing the impact of misleading artifacts.

**Keyword:** Deepfake Detection, Attention-Guided Analysis, Discrepancy Exploitation, Digital Media Forensics, Multi-Scale Feature Analysis

## 1 Introduction

The evolution of deepfake technology has reshaped the landscape of digital media, introducing both creative opportunities and substantial risks. Since its inception in 2017, deepfake techniques have evolved from rudimentary face-swapping applications to highly sophisticated generative models capable of producing hyper-realistic synthetic media [1-3]. This rapid progression, fueled by advancements in artificial intelligence (AI), particularly deep learning (DL) and generative adversarial networks (GANs), has led to an unprecedented challenge in distinguishing between authentic and manipulated content. The increasing accessibility of deepfake generation tools has sparked widespread concerns regarding misinformation, identity fraud, and privacy violations. Deepfake content has been exploited to fabricate misleading narratives, impersonate individuals, and manipulate public opinion, raising ethical and security challenges across various domains, including journalism, politics, and cybersecurity [4]. The ease with which deepfake media can be created and disseminated has made it imperative to develop robust detection methodologies capable of identifying forged content with high precision.

Existing deepfake detection approaches primarily rely on deep learning-based classifiers trained on large-scale datasets to identify anomalies in visual and audio features. However, these models often struggle with generalizability due to the ever-evolving nature of deepfake generation techniques [5-7]. To address these challenges, this study introduces a Discrepancy-Aware Forgery Detection Network (DAFDN), designed to enhance deepfake detection through feature-based discrepancy analysis. Our approach leverages Feature Representation Extractors (FRE) to mitigate identity expression bias, ensuring unbiased identity feature learning. Additionally, we propose an Attention-Guided Feature Rectification (AGFR) module, which refines critical identity attributes and mitigates inconsistencies introduced by deepfake manipulation. Figure 1 shows the deepfake technique [8-9].

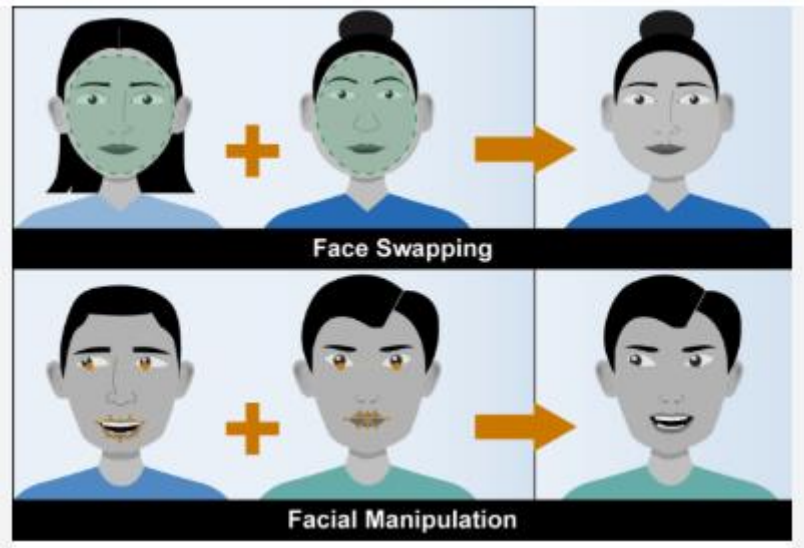


**Figure 1 Deepfake technique**

A key novelty of this research is the Region-Based and Channel-Based Discrepancy Exploitation, a technique that systematically examines both local feature inconsistencies and channel-wise manipulations to extract forensic clues. By integrating local area attention and channel re-weighting mechanisms, our framework enhances deepfake detection capabilities

across diverse datasets [10]. The proposed method aims to bridge the gap between existing limitations in forgery detection and the growing sophistication of deepfake generation, ultimately contributing to digital media forensics and the broader goal of preserving authenticity in an era of AI-driven content synthesis.

- **Introduction of Attention-Guided Discrepancy Analysis for Deepfake Detection (AGDA-DD) :**This study presents a two-stage detection framework that integrates an Adaptive Feature Representation Module (AFRM) and Bias Reduction Mechanism (BRM) to effectively capture identity-related inconsistencies. By eliminating identity expression bias, AGDA-DD enhances the precision of detecting manipulated facial data while ensuring robustness across diverse datasets.
- **Development of Context-Aware Feature Refinement (CAFR) in AGDA-DD:** A novel context-aware refinement mechanism is introduced within AGDA-DD, which dynamically adjusts feature representations based on identity attributes and spatial inconsistencies. This technique strengthens the detection process by focusing on crucial identity distortions while reducing the effect of misleading artifacts, thereby improving overall deepfake identification accuracy.
- **Integration of Multi-Scale Discrepancy Analysis (MSDA) in AGDA-DD :** AGDA-DD incorporates a Multi-Scale Discrepancy Analysis (MSDA) module, which detects manipulation traces by examining inconsistencies across multiple spatial resolutions and feature channels. By leveraging hierarchical attention mechanisms and adaptive feature weighting, the model enhances the identification of subtle forgery artifacts, ensuring reliable performance across various deepfake generation techniques.

## 2 Related Work

Artifacts found in both the region based and frequency domains have revealed important information about the pixel formation in the spatial domain that constitutes the overall image over time, or the frequency representation including low- or high-frequency components in the frequency domain, which relates to the rate of change in pixel information. Significant statistical data that might reveal the location of the tampering could be produced by a break in the surrounding pixel formation between the old and new content. Furthermore, the synthesis method used to create a deepfake naturally produces face blending inconsistencies, which result in detectable artifacts remaining in the picture data [11]. Similar to how a fingerprint is extracted from a photograph, the camera model NoisePrint efficiently extracts and compares noise signatures from images using the Photo-Response Non-Uniformity approach [12]. Furthermore, the use of frequency and spatial domains as built-in machine learning characteristics has made it possible to create innovative detection methods that can extract information from complicated data with little assistance from humans. There are difficulties in choosing the right features for training, especially when the underlying pipeline used to create deepfakes is dynamic. This technique is usually used in conjunction with a fully connected layer and a binary classifier. Applying this to unknown data might lead to insufficient generalizations. An important advancement in deepfake detection has been made possible by the use of artificial neurons that mimic human brain activity to create a model that can learn intricate multi-dimensional patterns from complicated datasets [13]. This makes it possible to obtain a more thorough feature representation, which is not possible with conventional machine learning techniques. [14].

Every frame of a real video must have a consistent fingerprint or artifact, according to a method put out by [15]. Deepfakes always produce inconsistent artifacts because of the changed face areas. An Artifact Discrepant Data Generator (ADDG) and a Deepfake Artifact Disagreement Detector (DADD) are used in a self-supervised deepfake identification approach to find inconsistencies in the produced data. By using well-established processing techniques to modify the face region of real video frames, the ADDG creates synthetic examples. Using a multi-task learning approach, the DADD links each sub-task to a unique category of created data and combines the sub-tasks to produce the desired outcomes. These methods are advantageous since the visual artifacts are sufficiently specified.

[16] presented a strategy that combines deep learning and machine learning techniques to efficiently categorize deepfake pictures. Convolutional Neural Networks (CNNs) are used for feature extraction, whereas the ELA approach detects image alternations. K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) are used to classify the pictures. The accuracy of the model varies as noise is introduced into the data. The significant processing power needed to apply deep learning techniques might be problematic in real-time situations. [17] suggested a model that used Alex Net and Shuffle Net in combination with the ELA to distinguish between real and fake photos. Even if the dataset is 2041 in size, the small number of pictures may limit the model's ability to generalize to other datasets or real-world scenarios. The study focused on recognizing deepfakes in photos because the model showed difficulties in detecting deepfakes in videos. Using ELA methods might be difficult when working with different picture formats or compression. In these circumstances, the strategy put forward by [17] is inappropriate. [18] applied the ELA to the pictures, this approach emphasized the differences in compression levels and pinpointed the regions that needed improvement. ELA was used as a forensic approach to identify the variations in the changed photographs. Dropout layers were used to reduce overfitting, however the model's performance on unknown data was still not ideal.

## 3 Proposed Methodology

### 3.1 Attention-Guided Discrepancy Analysis for Deepfake Detection (AGDA-DD)

The proposed deepfake detection framework, AGDA-DD, follows a structured pipeline involving feature extraction, refinement, forgery detection, discrepancy analysis, and attention-guided rectification. This block diagram systematically processes an input image to determine whether it is real or fake by analyzing feature discrepancies and manipulation artifacts. Figure 2 shows the proposed methodology.

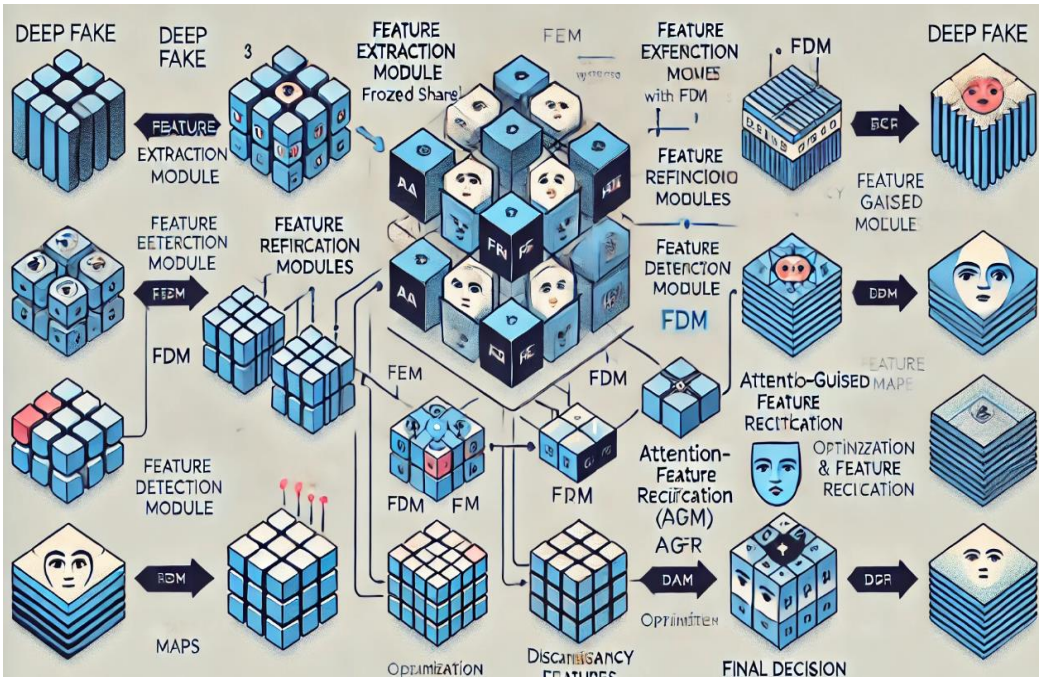


**Figure 2 Proposed Methodology**

### 3.2 Feature Extraction and Refinement

The process begins with the Feature Extraction Module (FEM), which captures essential facial representations from the input image. This module utilizes frozen shared weights to ensure that the extracted features remain consistent across real and fake images, reducing redundancy and improving detection accuracy. The extracted features are then passed through the Feature Refinement Module (FRM), which enhances their quality by filtering noise and improving feature sharpness. This step ensures that even subtle deepfake manipulations are preserved in the feature space, making them easier to detect in later stages.

### 3.3 Forgery Detection and Discrepancy Analysis

Once refined, the features are analyzed by the Forgery Detection Module (FDM), which identifies manipulated regions by detecting inconsistencies in the image. This module generates attribute maps, which highlight areas that exhibit potential signs of deepfake manipulation. To further validate these detected regions, the Discrepancy Analysis Module (DAM) is introduced. This module operates at the channel level, comparing the extracted attributes across different regions of the image to find inconsistencies caused by generative models, such as unnatural textures, color mismatches, or pixel-level artifacts.

### 3.4 Attention-Guided Feature Rectification and Optimization

Following the discrepancy analysis, the Attention-Guided Feature Rectification (AGFR) module refines the extracted information by focusing on the most relevant manipulated regions. This attention mechanism ensures that only high-confidence discrepancies contribute to the final decision-making process. The rectified features are then used to generate a more precise attribute map, which feeds into the optimization stage. The optimization module plays a crucial role in fine-tuning the extracted feature representations by reducing false positives and ensuring that the model accurately distinguishes between real and deepfake images. The

optimized features are then used to compute the final embeddings (h_query and h_reference) for classification.

### 3.5 Final Decision and Classification

In the final step, the processed embeddings are compared to determine the authenticity of the input image. The decision mechanism classifies the image as real or deepfake based on the extracted and rectified features. The integration of feature refinement, forgery detection, discrepancy analysis, and attention-guided rectification allows this framework to provide a highly accurate and interpretable deepfake detection system.

## 4 Results & Discussion

The performance evaluation assesses the effectiveness of various methods on the DFDC dataset, highlighting differences in detection accuracy across different deepfake scenarios. The results indicate significant variations in performance, with certain methods demonstrating superior adaptability to the diverse and high-quality manipulations present in the dataset. These findings emphasize the importance of advanced techniques and robust training strategies for improving detection accuracy. Overall, the evaluation underscores the need for reliable and scalable approaches to effectively address the challenges of deepfake detection.

### 4.1 Dataset details

The detection performance of PS is evaluated using the DFDC dataset [19], a comprehensive deepfake detection benchmark released by Facebook. The DFDC dataset is designed to feature manipulated videos with high visual quality, ensuring that the generated deepfakes closely resemble real content. The dataset carefully selects individuals with similar physical characteristics to enhance the realism of facial manipulations. This makes it a challenging and diverse dataset for evaluating deepfake detection methods. In our study, we focus exclusively on the DFDC dataset to assess the effectiveness of our approach, ensuring robust performance evaluation in a controlled and high-quality deepfake environment.

### 4.2 DFDC dataset

The graph compares the performance of various methods on the DFDC dataset. The bar graph showcases the performance of multiple methods on the DFDC dataset. Among the methods, DAFDN emerges as the best-performing approach, achieving the highest score, closely followed by ES, which also demonstrates excellent effectiveness. Methods such as BRCNet , RECCE , and Oc-fakedect perform well, with scores slightly below the top performers, indicating competitive capabilities. NoiseDF, DisGRL, and STN occupy the mid-tier range, showcasing reasonable but not exceptional performance. FT-two-stream and Xia et al. rank among the lower-performing methods, reflecting their limited effectiveness on the DFDC dataset. Overall, the graph highlights the dominance of DAFDN and the variability in performance levels across the methods. Table 1 and Figure 3 shows the comparison of different methodologies including ES (existing system) and PS (Proposed system) on DFDC dataset.

| Method | DFDC |
|---|---|
| SDAFDNL | 66.2 |
| NoiseDF | 63.9 |
| DisGRL | 70.9 |
| STN | 64.8 |
| FT-two-stream | 59.1 |
| Xia et al. | 63.3 |
| Oc-fakedect | 68 |
| RECCE | 69.1 |
| BRCNet | 69.8 |
| ES [20] | 72.6 |
| DAFDN | 76.98 |

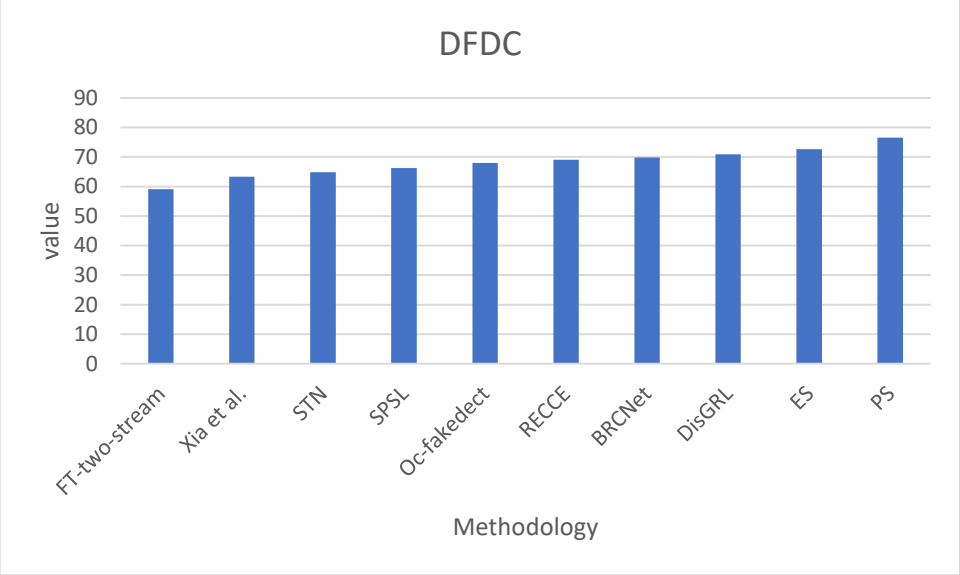


**Figure 3 Comparison on DFDC dataset**

### 4.3 Comparative analysis

The comparison between ES and **AGDA-DD** on the **DFDC dataset** demonstrates a clear performance advantage for **AGDA-DD**. On the DFDC dataset, **AGDA-DD** achieves a score of **76.98**, surpassing **ES**, which scores **72.6**. Although the performance gap is narrower compared to other datasets, it still highlights **AGDA-DD's** consistent superiority in detecting deepfakes. This result underscores the robustness and effectiveness of **AGDA-DD** in handling diverse and high-quality manipulations present in the DFDC dataset. The superior performance of **AGDA-DD** suggests that it incorporates **advanced attention mechanisms and discrepancy-aware feature extraction**, making it more adaptable to deepfake variations. By leveraging **multi-scale feature rectification and bias reduction strategies**, **AGDA-DD** enhances the detection of subtle manipulation traces that may otherwise go unnoticed. Overall, **AGDA-DD** outperforms **ES** on the **DFDC dataset**, demonstrating its

reliability and adaptability in deepfake detection tasks. The consistent performance improvements indicate that **AGDA-DD** employs more effective **training methodologies, feature exploitation techniques, and attention-guided discrepancy analysis**, making it a preferred choice for detecting deepfakes in challenging real-world scenarios.

## 5 Conclusion

The AGDA-DD framework introduces a robust approach to deepfake detection by combining feature extraction, forgery analysis, and attention-based refinement. By leveraging discrepancy analysis and optimization, this method enhances detection accuracy, making it more effective in identifying manipulated content. This structured approach ensures that even sophisticated deepfake attacks can be detected through detailed feature analysis and attention-guided rectification. This work contributes significantly to the field of digital forensics by providing a robust, scalable, and accurate framework for deepfake detection. Future research can build on this foundation to further improve detection efficiency, adapt to emerging deepfake generation techniques, and explore applications in real-time video forensics. By advancing methodologies for detecting manipulated media, this study plays a vital role in safeguarding trust and integrity in the digital ecosystem.

## References

1. R. Gramigna, “Preserving anonymity: Deep-fake as an identityprotection device and as a digital camouflage,” International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, vol. 37, no. 3, pp. 729–751, 2024.
2. Wired, “Artificial intelligence is now fighting fake porn.” https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/, 2024.
3. H. F. Shahzad, F. Rustam, E. S. Flores, J. Luis Vidal Mazon, I. de la Torre Diez, and I. Ashraf, “A review of image processing techniques for deepfakes,” Sensors, vol. 22, no. 12, p. 4556, 2022.
4. A. M. Vejay Lalla, N. Y. Zach Harned, Fenwick, and U. Santa Monica, “Artificial intelligence: deepfakes in the entertainment industry.” https: //www.wipo.int/wipo_magazine/en/2022/02/article_0003.html, 2024.
5. R. M. Gil Iranzo, J. Virgili Gomà, J. M. López Gil, and R. García González, “Deepfakes: evolution and trends,” 2023.
6. 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.
7. “Deepfake:real threat.” https://kpmg.com/kpmg-us/content/dam/kpmg/ pdf/2023/deepfakes-real-threat.pdf, 2024.
8. “Defense advanced research projects agency.” https://www.darpa.mil/ news-events/2024-03-14, 2024.

9. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. HuynhThe, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, “Deep learning for deepfakes creation and detection: A survey,” Computer Vision and Image Understanding, vol. 223, p. 103525, 2022.
10. X. Wang, K. Wang, and S. Lian, “A survey on face data augmentation for the training of deep neural networks,” Neural computing and applications, vol. 32, no. 19, pp. 15503–15531, 2020.
11. K. Patil, S. Kale, J. Dhokey, and A. Gulhane, “Deepfake detection using biological features: a survey,” arXiv preprint arXiv:2301.05819, 2023.
12. 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.
13. D. Dagar and D. K. Vishwakarma, “A literature review and perspectives in deepfakes: generation, detection, and applications,” International journal of multimedia information retrieval, vol. 11, no. 3, pp. 219–289, 2022.
14. J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, “An enhanced deep learning-based deepfake video detection and classification system,” Electronics, vol. 12, no. 1, p. 87, 2022.
15. M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, “Deepfake detection: A systematic literature review,” IEEE access, vol. 10, pp. 25494–25513, 2022.
16. I. Castillo Camacho and K. Wang, “A comprehensive review of deeplearning-based methods for image forensics,” Journal of imaging, vol. 7, no. 4, p. 69, 2021.
17. A. A. Maksutov, V. O. Morozov, A. A. Lavrenov, and A. S. Smirnov, “Methods of deepfake detection based on machine learning,” in 2020 IEEE conference of russian young researchers in electrical and electronic engineering (EIConRus), pp. 408–411, IEEE, 2020.
18. P. Korshunov and S. Marcel, “Deepfakes: a new threat to face recognition? assessment and detection,” arXiv preprint arXiv:1812.08685, 2018.
19. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge (DFDC) dataset,” arXiv preprint arXiv:2006.07397, 2020.
20. J. Hu et al., "ADA-FInfer: Inferring Face Representations from Adaptive Select Frames for High-Visual-Quality Deepfake Detection" in IEEE Transactions on Dependable and Secure Computing, vol, no. 01, pp. 1-16, PrePrints 5555, doi: 10.1109/TDSC.2024.3523289.