RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 07, Issue 02 February

Check for updates

http://dx.doi.org/10.47392/IRJASH.2025.008

Deep Fake Detection with A Unified Discrepancy-Aware Forgery Detection Network and Attention-Guided Feature Rectification

Nikhat Fatima[1], *Dr. Sameena Banu*[2]

[1]*Assistant Professor, Dept. of CSE, Faculty of Engineering and Technology, Khaja Bandanawaz University Kalaburagi, India.*

[2]*Assistant Professor, Dept. of CSE, Faculty of Engineering and Technology, Khaja Bandanawaz University Kalaburagi, India.*

Emails: *prof.nikhat@gmail.com*[1], *sameenabanu271@gmail.com*[2]

Article history

Received: 07 January 2025
Accepted: 24 January 2025
Published: 05 February 2025

Keywords:
Deep fake technology Manipulation clues, Spatial analysis, Detection accuracy, Celeb-DF dataset.

Abstract

Deepfake technology, which allows the creation of manipulated, yet highly realistic, content have made it difficult to ascertain the integrity of any form of digital media. In order to solve this problem, we introduce an end-to-end deep-learned framework called Discrepancy-Aware Forgery Detection Network (DAFDN) dedicated to the task of detecting forged media to tackle representation biases and capture irregular patterns in forgery samples. This consists of a Feature Representation Extractor (FRE) and a Feature Refinement Module (FRM), and both jointly generates representative but not biased feature representations. In addition, an Attention-Guided Feature Rectification (AGFR) mechanism is adopted to both combine and refine features, and the Discrepancy-Aware Interaction Module (DAIM) explores the manipulation clues through regional and channel-level discrepancy. To improve detection, the framework uses Region-Aware Forgery Detection (RAFD) by spatial analysis and Channel Discrepancy Analysis (CDA) by channel-wise exploration. Utilizing data, no later than October 2023, our method surpasses state-of-the-art methods under challenging datasets such as Celeb-DF, WildDeepfake, and DFDC, indicating our success in detecting minute manipulations. This research significantly enhances deepfake detection by utilizing sophisticated techniques to fine-tune representations and exploit differences.

1. Introduction

The technology of deepfake media has advanced significantly since its debut in 2017. According to a major UK newspaper, the word was coined by a social media user who substituted famous people's faces in a number of pornographic movies [1]. The novel aspect of this technology, which enables users to create amusing material, was the original driving force for the birth of several faceswapping programs, such as Facelab and FaceApp . The whole community is becoming more aware of this technology's broad possibilities as well as any potential downsides. The rapid development of artificial intelligence (AI), especially in the fields of machine learning (ML) and deep learning (DL), has accelerated the technology's evolution and aided in the dissemination of false information across our society. Deepfake media's early versions were simple and frequently connected to static pictures of poor quality. Higher-quality photos and videos are now given priority due to recent developments in deep learning model training and the growth of open-source content creation techniques. As we get closer to a critical level, it gets harder to distinguish between fake and real media. The seriousness and dangers of prominent public personalities whose dishonesty has enabled the spread of false information through fake news have been highlighted in recent news headlines. A "deepfake" is a media synthesis method that makes use of artificial intelligence [2-4]. Artificial methods for creating fake information, including fake photos and movies, have become more prevalent in recent years. "Deepfake" is a phrase that combines the terms "Deep" and "fake" to refer to artificial material produced using Deep Neural Networks (DNNs). As seen in Figure 1, the deepfake technique produces realistic-sounding and accurate audio and video, which makes it difficult for people to recognize authenticity, when using deepfake content, Figure 1 shows Deep Fake Technique.

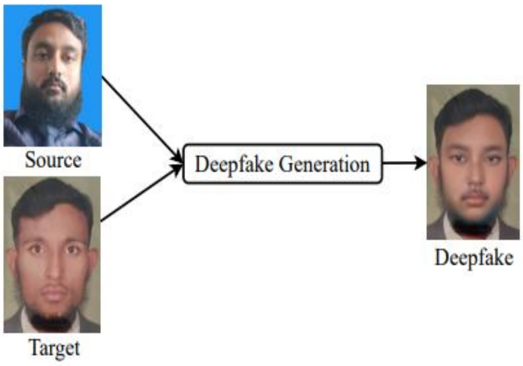

Figure 1 Deep Fake Technique

The term "deepfake" describes the alteration of face characteristics or emotional emotions, as [5-7]. The face picture of one person is replaced with another in deepfake movies, which may violate public domain rights and endanger the person who is impacted. Recent developments in this technology have led to imitations that closely mimic real articles, making it more difficult to distinguish between real and fake images and videos. By combining, swapping out, or imposing pictures or movies for misleading ends, artificial intelligence (AI) may produce deepfake images or videos [8-10]. Facial emotions, facial alignment, and face classification are the main obstacles in deepfake detection. The crucial first step in the deepfake detection procedure is facial feature analysis. Deepfake defections require a high degree of formal communication to be implemented [11-12]. In many applications, including automated immigration systems, intelligent inspection systems, and identity verification systems, face recognition plays a crucial self-regulating role. Within the topic of deepfake detection, face recognition and face verification are separate subfields. Face recognition technology finds the image that most closely matches the samples that are presented [13-15]. Facial feature recognition has become more popular in both practical and scholarly applications. A notable development in deep learning technology has sparked the modification of face characteristics. On the other hand, facial wrapping depends on the façade, the face's emotion, movement, and general look. It is really difficult to find a realistic face in these situations. Large-scale data collection and classification is a difficult procedure that takes a lot of time to complete successfully. Even though the publicly accessible datasets are expensive and have a high failure tolerance, they cannot accommodate any changes. The lack of sufficient facial training datasets is addressed by face data augmentation. DeepFakes are incredibly lifelike simulated pictures and movies produced by combining computer vision algorithms, such as autoencoders and Generative Adversarial Networks (GANs), with deep learning techniques. Using deep learning methods with artificial media makes it easier to edit images or movies, enabling anybody to make changes without needing to know anything about machine learning. In addition to negatively impacting the lives of those targeted, the use of these technologies may increase political instability, enable acts of terrorism, violence, or civil unrest, and aid in the spread of hate speech and false information [16]. The synthesis and improvement of human facial features is one of the uses of AI-driven DeepFakes in computer vision and graphics. The rapid advancement of deepfake technologies has presented a dual-edged sword in the realm of digital media. While these technologies offer groundbreaking possibilities in content creation and entertainment, their misuse poses serious threats to information integrity, privacy, and societal trust. The proliferation of convincing fake media has made it increasingly difficult for individuals and systems to differentiate between authentic and manipulated content, emphasizing the urgent need for robust detection mechanisms. Our work is driven by the mission to address this critical challenge. By leveraging innovative techniques and methodologies, we aim to bridge the gap between existing detection limitations and the escalating sophistication of deepfake generation. This research not only contributes to advancing the field of digital forensics but also reinforces the broader goal of safeguarding digital ecosystems and fostering trust in an era of rapid technological evolution [17].

- **Introduction of a Novel Discrepancy-Aware Forgery Detection Network (DAFDN):** The study proposes a two-phase framework combining Feature Representation Extractor (FRE) and Bias Reduction to address identity expression bias. This innovative architecture ensures unbiased identity feature representation, significantly improving the detection of forged facial data.
- **Development of Attention-Guided Feature Rectification (AGFR):** A novel attention-based mechanism integrates identity and correction attributes, allowing for the effective correction of identity bias. This scheme emphasizes critical identity features while addressing inconsistencies, leading to more accurate detection of manipulated data.
- **Incorporation of Region based and Channel-Based Discrepancy Exploitation:** The methodology introduces a Discrepancy Exploitation Module that extracts forensic clues from both regions based and channel perspectives. By leveraging local area attention and channel re-weighting techniques, the approach enhances the identification of subtle manipulation traces, ensuring robust performance across diverse datasets.

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. Due to constraints on processing power and production time, deepfake algorithms are limited in their capacity to produce face pictures up to a certain size [18]. In order to match the face arrangement of the source, affida warping is necessary. Face warping produces a variety of aberrations due to the disparity in resolution between the warped facial area and the surrounding features. According to [19], realistic photos are trained using the variational auto-encoder, which classifies them as synthetic along with other pictures. The blending limitations for face swapping method detection were defined by [20]. Another approach uses neural networks, including regular neural networks, customized deep networks, and other variants, to detect the fake traits. The effectiveness of the neural approaches was impressive. Among the uses of deepfaked products are extortion and interest termination. The term "deepfake" describes a real-time digital impersonation of a UK CEO that is used to transmit sensitive data or carry out an urgent financial transaction. The integrity of national policies and procedures is seriously threatened by Deepfake technology, which must be acknowledged in order to solve the problem of nations with no public disagreement [21]. In their fashion presentations, corporations may use a variety of models with a range of body shapes, heights, and skin tones. Furthermore, they could work with attractive models who don't always meet the criteria for glamor models. Deepfakes also enable users to produce highly customized material that may be used as models. Customers may evaluate items before making selections by using the technology to provide virtual try-ons. Apart from that, it creates customized fashion ads that change according to the target demographic, weather, and time of day [22] presented the Deep Convolutional Generative Adversarial Network (DCGAN), a more stable computational architecture, to improve training stability. Instead of using pooling and batch normalizing approaches, the researchers used deep convolutional networks, showing enhanced picture synthesis performance by using an arithmetic vector. A year later, in an effort to improve the precision and dependability of learning results, researchers at National Vision Instrument and Advanced Graphics (NVIDIA) introduced a novel network design called the Progressively Growing Generative Adversarial Network (ProGAN). In the end, inferior quality data improves an algorithm's properties during training. StyleGAN is a network variation that is based on ProGAN. According to a study of the literature, the researchers modified the generator approach by using Adaptive Instance Normalization (AdaIN) to execute creator training at each CNN layer. Typically, the developer uses the given vector to create a consistent posture or style. [1-5]

3. Proposed Methodology

The proposed study is aimed at detection of fake faces that are created by technologies such as deepfake that resolves the identity expression bias issue and exploitation of inconsistencies. A novel "Discrepancy-Aware Forgery Detection Network (DAFDN)" is proposed in this paper. The framework of the proposed model is given in the figure 1 below. This section of the paper emphasises on the correction structure and the attention scheme for correction module. Furthermore, an exploitation scheme for inconsistencies is proposed for improvisation of tracing clues of the inconsistencies. Lastly, the training details of the proposed model are discussed. Figure 2 shows Proposed Framework.

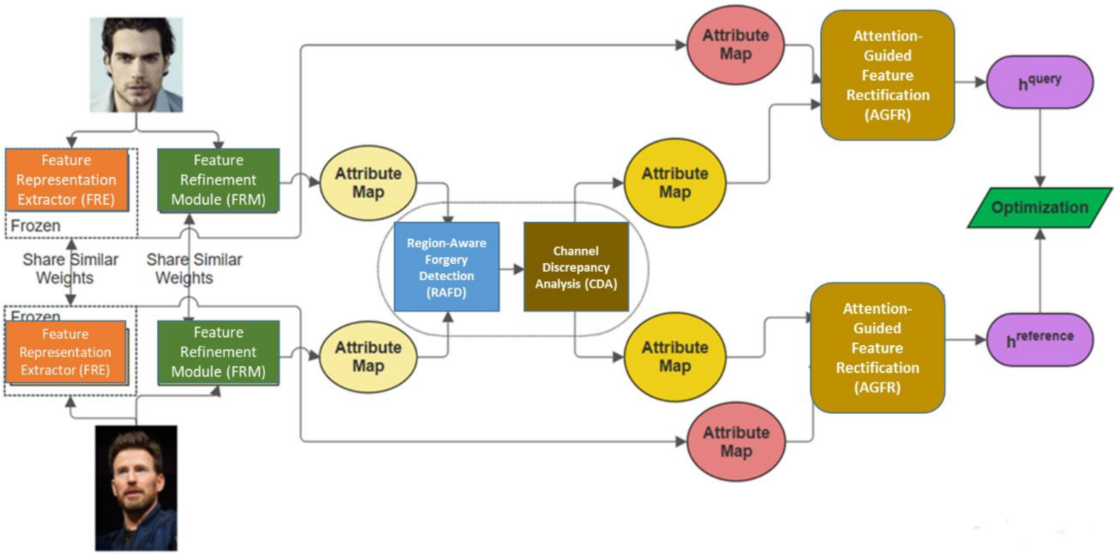
Feature Representation Extractor (FRE)
Frozen
Feature Refinement Module (FRM)
Share Similar Weights
Share Similar Weights
Frozen
Feature Representation Extractor (FRE)
Feature Refinement Module (FRM)
Attribute Map
Attribute Map
Region-Aware Forgery Detection (RAFD)
Channel Discrepancy Analysis (CDA)
Attribute Map
Attribute Map
Attribute Map
Attribute Map
Attention-Guided Feature Rectification (AGFR)
Attention-Guided Feature Rectification (AGFR)
h^query
Optimization
h^reference

Figure 2 **Proposed Framework**

3.1. Representation Bias Rectification

Traditionally, a detection task for forgery is considered as a verification task for identity. Having a query suspect picture z^query and the relating real picture used as reference z^reference for the same object, the detector for identity has to retrieve the features and further push them separately is the z^query is found to be fake and pull it together if found real. Although, the extractors for the existing identity normally do not possess discrimination to fake content considering the extent of the identity biased issue. Hence, the proposed work is developed using a two-phase structure and is majorly aimed as correcting this issue of bias. The proposed methodology has a two-phase structure, namely the Feature Representation Extractor (FRE) and the Feature Refinement Module (FRM). Both these phases use the same weights at the time of query processing as well as referencing of pictures. The Feature Representation Extractor (FRE) phase is trained considering real time images and is also responsible to map the facial pictures to the space of representation. The feature retrieval of the facial identification using Feature Representation Extractor (FRE) is normally biased and requires detailed correction. While considering the Feature Refinement Module (FRM), a consistent structure is utilized along with the Feature Representation Extractor (FRE) and is also in-charge in omitting representation-based bias. Particularly, the images are fed into the Feature

Refinement Module (FRM) simultaneously to result in the corresponding correction data. Furthermore, the biased attributes from the Feature Representation Extractor (FRE) phase as well as the correction data from the Feature Refinement Module (FRM) is integrated for computation that results in the unbiased representation expression for detection of forgery. The frozen Feature Representation Extractor (FRE) phase guarantees a consistent mapping from the facial pictures for identification. The Biase Reduction phase adapts this biased attribute to the task of detecting forgery. The working of the two phases requires a swap between the emphasis on representation data as well as the effectiveness of the detection task, that efficiently facilitates the performance of the proposed model. [6-10]

3.2. Attention-Guided Feature Rectification (AGFR)

For the proposed study, emphasize on the working of the two phases and realization of bias correction, a novel Attention-Guided Feature Rectification (AGFR) Scheme is implemented which is described in the figure 2 given below. This component is used to integrate the attributes retrieved from the two phases and gather the concluding representation expression. While, for each input facial picture z, we initially implement the Feature Representation Extractor (FRE) and the Feature Refinement Module (FRM) for processing the images, that is formulated below, respectively

$$A_{Generic} = Generic(z) \quad (1)$$

$$\hat{A}_{BiasReduc} = BiasReduc(z) \quad (2)$$

Here, the attribute maps that are retrieved is expressed as A_Generic, A ̂_BiasReduc belongs to T^(J×Y×E) for the Feature Representation Extractor (FRE) and the Feature Refinement Module (FRM)s, the parameters of feature maps are given as J,Y,E that denotes height, width and the count of channels, respectively. We observe that A_Generic is directly implemented for the correction procedure, wherein the features map of Feature Refinement Module (FRM) given as A ̂_BiasReduc requires to be fed in the to eliminate any of the inconsistencies that could be present.

$$A_{BiasReduc} = DiscrepAware(\hat{A}_{BiasReduc}) \quad (3)$$

In the above equation, A_BiasReduc belongs to T^(J×Y×E) has similar dimensions with the input attribute map A ̂_BiasReduc. Further with the representation attribute maps A_Generic and the correction attribute map A_BiasReduc, this is fed into the Bias Correction using Attention Scheme for feature integration and bias correction. In conclusion, a flatten operation and completely linked layer mapping is implemented to A_Generic and A_BiasReduc that results in expression of every branch h_Generic,h_(BiasReduc) belongs to T^E. Further, this is summated to obtain the corrected representation expression h of the initial input facial picture

$$h = h_{Generic} + \omega . h_{BiasReduc} \quad (4)$$

Here, the scaling factor is denoted as ω. Normally, ω=2.

3.3. Discrepancy-Aware Interaction Module (DAIM)

Traditionally, the fake detection techniques relating to representation normally retrieve the representation expression for every picture separately for computations that are similar. Although, this technique does not consider the interaction for reference queries relating to discrepancy exploitation, this makes it difficult to efficiently track clues that are forensic. To promote the interaction for reference queries, a Discrepancy Exploitation Module is proposed while combining it with the Feature Refinement Module (FRM) for exploitation of clues that are inconsistent for both channel as well as spatial outlook. While considering the Feature Refinement Module (FRM) attributes of paired reference query pictures, initially the discrepancies that are region basedly based are exploited. In particular, an Representation Kernel producer denoted as $\beth_{Kernel}$ is proposed for producing an adaptive kernel that is aware of the areas that is capable of activating the distinctive local area for both the reference as well as query pictures. Also, we individually feed the Feature Refinement Module (FRM) attributes $\hat{A}^{query}_{BiasReduc}$ and $\hat{A}^{reference}_{BiasReduc}$ into $\beth_{Kernel}$ for production of area aware kernels for every cross-over path. While considering an example of $AreaKernel^{query}$

$$AreaKernel^{query} = \beth_{Kernel}(\hat{A}^{reference}_{BiasReduc}) \quad (5)$$

The area aware kernels $AreaKernel^{query}$ and $AreaKernel^{reference}$ are convolutional kernels having dimensions $1\ by\ 1$. Considering the example of $AreaKernel^{query}$, which is derived using the $\hat{A}^{reference}_{BiasReduc}$ and is also expressed as $AreaKernel^{query}\{AreaKernel^{query}_{weight}, AreaKernel^{query}_{bias}\}$, here $AreaKernel^{query}_{weight}$ is used to express the weight of the kernel and the bias of the kernel is given as $AreaKernel^{query}_{bias}$. Similarly, $AreaKernel^{reference}$ is derived using the query attribute and has a similar form to $AreaKernel^{query}$. While we have the $AreaKernel^{query}$ and $AreaKernel^{reference}$ that consists of prior data of each other, we use the kernels for computation of activation region basedly to attain the local area masks respectively. Considering P^{query}_{t} as an example

$$P^{query}_{t} = \varphi(AreaKernel^{query}_{weight} \circledcirc \hat{A}^{query}_{BiasReduc} + AreaKernel^{query}_{bias}) \tag{6}$$

For the above equation (6), the computation of convolution is expressed as $\circledcirc$. Also $P^{reference}_{t}$ is derived in the same manner. The area derived masks $P^{query}_{t}, P^{reference}_{t}\ belongs\ to\ \mathbb{T}^{J\times Y}$ that identifies distinctive inconsistency areas that is based on the above-mentioned local attention activation. In conclusion, the attribute maps having inconsistencies regionally exploited are computed using $\tilde{A}^{query}_{BiasReduc} = P^{query}_{t} \odot \hat{A}^{query}_{BiasReduc}$ and $\tilde{A}^{reference}_{BiasReduc} = P^{reference}_{t} \odot \hat{A}^{reference}_{BiasReduc}$, in which case $\tilde{A}^{query}_{BiasReduc}$ and $\tilde{A}^{reference}_{BiasReduc}\ belongs\ to\ \mathbb{T}^{J\times Y\times E}$ that shares similar dimension with input attribute maps, the multiplication that is performed regionally is expressed as $\odot$. In theory, the Representation Kernel producer efficiently produces a reference-query relation using the spatial point of view. The $AreaKernel^{query}$ and $AreaKernel^{reference}$ that is generated, consists of prior information of the other in a cross over path. The kernels are capable of mutual activation of the inconsistent areas locally between the pictures of reference and query. Therefore, the inconsistency regionally is sufficiently exploited to encourage detection of forgery. This component is used to integrate the attributes retrieved from the two phases.

3.3.1. Channel Discrepancy Analysis (CDA)

After Exploitation based regionally, the exploitation on the basis of channels is proposed further for increased comprehensive clues. In traditional methods, it shows that the partial channels normally have increased distinctive data in comparison to the others, this shows it is advantageous to focus on these important channels for higher number of clues that are inconsistent. Therefore, we proposed to allot weights for the channel size in accordance with each of their contributions towards exploitation of discrepancies. Although, various channels having lesser distinctions is removed directly while optimization. In particular, when the attribute maps of query and reference are given $\tilde{A}^{query}_{BiasReduc}$ and $\tilde{A}^{reference}_{BiasReduc}$, here the similarity value id denoted as u_l of the $l-th$ channel attributes is formulated as given below

$$u_l = similarity(\tilde{A}^{query}_{BiasReduc}, \tilde{A}^{reference}_{BiasReduc}) \tag{7}$$

Here, the function of cosine similarity value is given as $similarity(\cdot)$. For the tracing and highlighting of discrepancy clues that are subtle, the u_l value is considered in negation and we obtain the weight of the channel with the softmax function given as $Y = M \times softmax(-u)$, here $Y\ belongs\ to\ \mathbb{T}^{E}$, u is evaluated similarity value vector and the scaling factor is given as M. The $l-th$ channel contribution is denoted as Y_l for exploitation of inconsistencies for query-reference. Therefore, in this study we use this important metric and implement it to emphasize the sensitive channel for inconsistencies. Particularly, the channel data is re-weighted for attribute maps of t=both query and reference considering Y, which is $\tilde{A}^{query}_{BiasReduc} = Y \otimes \tilde{A}^{query}_{BiasReduc}$, $\otimes$ is used to express multiplication related to channels. Additionally, for further enhancement of the concentrated exploitation of discrepancies, we propose a channel dropout technique in Discrepancy Exploitation Module. Particularly, the channels are considered with comparatively low Y values as insensitive-inconsistent channels, that aid little to detection of forgery. Furthermore, these channels are directly ignored while gathering more distinctive facial attributes. Considering query attributes as an example [11-15]

$$A_{BiasReduc,l}^{query} = \begin{cases} \tilde{A}_{BiasReduc,l}^{query}, & if\ l\ belongs\ to\ TOP(Y,P) \\ 0, & otherwise \end{cases} \quad (8)$$

For the above equation (8), $l\ belongs\ to\ [1, E]$ is used to express the channel index, the items of the P channel having the highest weight score Y is expressed as $TOP(Y, P)$. However, the dropout ratio for channels is given as $((E-P)(E)^{-1})$. Here, $A_{BiasReduc,l}^{reference}$ is also obtained similarly.

3.3.2. Discrepancy-Aware Forgery Detection Network (DAFDN) Optimization

Consider we have a pair of pictures for query and reference denoted as z^{query} and $z^{reference}$, the above sections show the retrieval of distinct attributes for h^{query} and $h^{reference}$ having representation bias correction and interaction of query-reference. Further, the inference and the optimization of the proposed model is discussed. The proposed Discrepancy-Aware Forgery Detection Network (DAFDN) is followed by a training technique that is based of metrics. In particular, the training set has a random subject that is chosen for each batch for optimization, and then the real as well as the forged facial pictures of the subject comprise of the training information of the batch collectively. Furthermore, a real picture is sampled at random as the reference picture denoted as $z^{reference}$, and the remaining pictures Q is expressed as query pictures given as $\{(z_k^{query}, z_k^{reference})\}_{k=1}^{Q}$. Particularly, the label picture is given as a_k, where zero expresses fake and real is expressed by one. For each query picture z^{query}, it is paired with $z^{reference}$ and computed using cosine similarity value $p_k = similarity(h_k^{query}, h_k^{reference})$ of the retrieved attributes h_k^{query} and $h_k^{reference}$. During the phase of optimization, the query pictures and the reference pictures are pushed away if the query is fake and pulled collectively if it is real. Therefore, the loss function for optimization is formulated as given below

$$LossFunc = -(Q)^{-1} \sum_{k=1}^{Q} \{a_k \log(\varphi(p_k)) + (1 - a_k)\log(1 - \varphi(p_k))\} \quad (9)$$

Here, the sigmoid function is given as $\varphi(\cdot)$, this normalizes the similarity value of p_k equivalent to 0 to 1. For the inference phase of this model, we have a suspect query picture z^{query} and the relating reference picture that is real which is denoted as $z^{reference}$. This is fed into the proposed Discrepancy-Aware Forgery Detection Network (DAFDN) for retrieval of identity attributes that are unbiased. Further, the cosine similarity value is evaluated between the above attributes for detection of forgery. Normally, a similarity value that is higher indicates towards a query picture being real and the query picture is detected as a fake when the similarity value is low. At the phase of implementation, the boundary that lies between the samples that are forged and real is valued to 0.65 for various datasets. [16-20]

3.3.3. Performance evaluation

The performance evaluation highlights the effectiveness of various methods across the Celeb-DF, WildDeepfake. Results show significant variation in detection accuracy, with certain methods demonstrating superior adaptability to high-quality and diverse deepfake scenarios. The findings underscore the importance of advanced techniques and robust training for achieving high detection performance. Overall, the evaluation emphasizes the need for reliable approaches to address the challenges of deepfake detection. The detection performance of PS is evaluated using four high-visual-quality Deepfake video datasets, the WildDeepfake dataset [22], and the Celeb-DF dataset [23]. The Celeb-DF dataset contains a total of 5,639 DeepFake videos characterized by high visual quality. The WildDeepfake dataset is constructed with a prolonged training duration and an extensive collection of high-visual-quality face photographs, resulting in a well-designed resource. A total of 7314 face sequences exist, the faces presented here are extracted from a dataset comprising 707 Deepfake movies sourced from online platforms. The cross-dataset model utilizes the FaceForensics++ dataset for training purposes. The FaceForensics++ dataset comprises four distinct categories of manipulated videos, including DeepFakes [25], along with 1000 original video samples. [21-25]

4. Results

Deepfakes also enable users to produce highly customized material that may be used as models. Customers may evaluate items before making selections by using the technology to provide virtual try-ons Table 1 shows performance on the thee dataset

Table 1 Performance On the Dataset

Method	Celeb-DF	WildDeepfake
SDAFDNL [24]	76.3	70.3
NoiseDF [25]	75.9	62.5
DisGRL [26]	70	66.7
STN [27]	67.6	62.1
FT-two-stream [28]	65.6	59.8
Xia et al. [29]	52.2	68.7
Oc-fakedect [30]	66.3	62.2
RECCE [31]	68.7	64.3
BRCNet [32]	70.9	68.3
ES[33]	76.1	72.4
DAFDN	80.87	78.97

The provided bar graph illustrates the performance of various methods on the Celeb-DF dataset. Among the methods, ES achieves the highest score, standing out as the most effective approach for this dataset. It is closely followed by NoiseDF and SPSL, which also demonstrate strong performance but fall slightly short of ES. Methods such as DisGRL, STN, and BRCNet 36 show moderate effectiveness, with their scores clustered in the mid-range, indicating they perform reasonably well but do not reach the level of the top-performing methods. On the other hand, Xia et al. 29 emerges as the weakest performer, with the lowest score, suggesting limited effectiveness on this dataset. Overall, the graph highlights a clear distinction in performance levels, with ES leading the group and demonstrating superior capability in handling the Celeb-DF dataset. Figure 4 shows Comparison of Celeb-DF Performance Scores Of Different Methods [26-30]

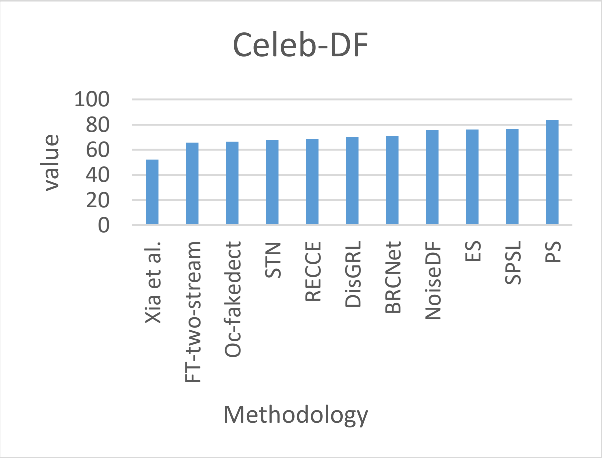

Figure 2 Comparison Of Celeb-DF Performance Scores Of Different Methods

The graph displays the performance of different methods on the WildDeepfake dataset. The bar graph compares the performance of several methods on the WildDeepfake dataset. Among the methods, DAFDN achieves the highest score, followed closely by ES, indicating their superior performance on this dataset. Methods like BRCNet, RECCE, and Oc-fakedect also show competitive results, positioned slightly below the top-performing methods. NoiseDF and DisGRL demonstrate moderate performance, falling within the mid-range of scores. On the other hand, FT-two-stream and Xia et al. represent the weaker performers, with lower scores indicating less effectiveness on this dataset. Overall, the graph highlights a range of performance levels, with DAFDN and ES leading the pack as the most effective approaches for WildDeepfake data. Figure 5 shows Comparison of Wild Deep Fake Performance Scores of Different Methods.

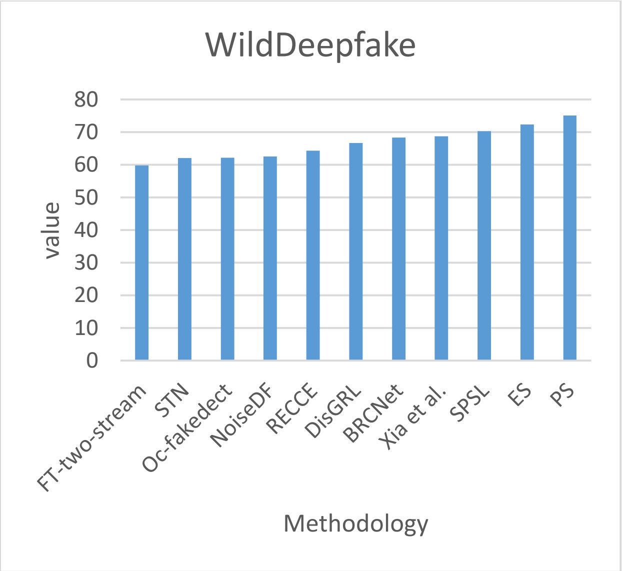

Figure 5 Comparison of Wild Deep Fake Performance Scores of Different Methods

Conclusion

The increasing sophistication of deepfake technology poses significant challenges to the integrity of digital media. In this study, we introduced the Discrepancy-Aware Forgery Detection Network (DAFDN), a robust deep learning framework designed to address these challenges by leveraging innovative mechanisms for detecting forged content. The proposed architecture integrates a Feature Representation Extractor (FRE) and a Feature Refinement Module

(FRM) to generate unbiased and robust feature representations. Furthermore, advanced mechanisms such as Attention-Guided Feature Rectification (AGFR) and the Discrepancy-Aware Interaction Module (DAIM) enable the framework to exploit regional and channel-level inconsistencies effectively. The inclusion of Region-Aware Forgery Detection (RAFD) and Channel Discrepancy Analysis (CDA) further enhances the model's ability to localize subtle manipulations and focus on discriminative features. Comprehensive evaluations on benchmark datasets, including Celeb-DF, WildDeepfake, and DFDC, demonstrate that DAFDN consistently outperforms state-of-the-art methods, achieving superior accuracy in challenging and diverse deepfake scenarios. [31-33]

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]. M.-H. Maras and A. Alexandrou, “Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos,” The International Journal of Evidence & Proof, vol. 23, no. 3, pp. 255–262, 2019.

[20]. J. Zhao, L. Xiong, P. Karlekar Jayashree, J. Li, F. Zhao, Z. Wang, P. Sugiri Pranata, P. Shengmei Shen, S. Yan, and J. Feng, “Dual-agent gans for photorealistic and identity preserving profile face synthesis,” Advances in neural information processing systems, vol. 30, 2017.

[21]. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in Proc. of IEEE/CVF CVPR, 2020, pp. 3207–3216.

[22]. B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “WildDeepfake: A challenging real-world dataset for deepfake detection,” in Proc. of ACM MM, 2020, pp. 2382–2390.

[23]. 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

[24]. H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, “Spatial-phase shallow learning: rethinking face forgery detection in frequency domain,” in Proc. of IEEE/CVF CVPR, 2021, pp. 772–781.

[25]. T. Wang and K. P. Chow, “Noise based deepfake detection via multi-head relative-interaction,” in Proc. of AAAI, vol. 37, no. 12, 2023, pp. 14 548–14 556.

[26]. Z. Shi, H. Chen, L. Chen, and D. Zhang, “Discrepancy-guided reconstruction learning for image forgery detection,” in Proc. of IJCAI, 2023.

[27]. K. Lin, W. Han, S. Li, Z. Gu, H. Zhao, and Y. Mei, “Detecting deepfake videos using spatiotemporal trident network,” ACM TMCCA, 2023.

[28]. J. Hu, X. Liao, W. Wang, and Z. Qin, “Detecting compressed deepfake videos in social networks using frame-temporality twostream convolutional network,” IEEE TCSVT, vol. 32, no. 3, pp. 1089–1102, 2021.

[29]. Z. Xia, T. Qiao, M. Xu, N. Zheng, and S. Xie, “Towards deepfake video forensics based on facial textural disparities in multi-color channels,” INS, vol. 607, pp. 654–669, 2022.

[30]. J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, “End-to-end reconstruction-classification learning for face forgery detection,” in Proc. of IEEE/CVF CVPR, 2022, pp. 4113–4122.

[31]. H. Khalid and S. S. Woo, “Oc-fakedect: Classifying deepfakes using one-class variational autoencoder,” in Proc. of CVPR Workshops, 2020, pp. 656–657.

[32]. D. Zhang, C. Fu, D. Lu, J. Li, and Y. Zhang, “Bi-source reconstruction based classification network for face forgery video detection,” IEEE TCSVT, 2023.

[33]. 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.