International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 11 Issue 10 October 2022 || PP 69-77

Analysis of Visual Media Alteration to Prevent Frauds Using Deep Learning Methods

Nikhat Fatima[1], Dr. Sameena Banu[2]
[1,2]*Assistant professor, CSE Dep of Faculty of Engineering and Technology, Khaja Bandanawaz University.*

Abstract
Fake materials on social media, known as deepfake media, comprise images and videos altered with artificial intelligence tools. Made with techniques including face swapping, lip-sync, face synthesis, and attribute modification, these fakes could be visual, aural, or textual. With an expected 500,000 video and audio deepfakes broadcast on social media platforms globally by the end of 2023, deepfake has evolved into one of the top five identity fraud kinds in 2023. Deepfake technology can, however, also be applied in creative endeavours like upgrading multimedia, movies, instructional materials, digital communications, gaming and entertainment, social media, healthcare delivery, material science, and many commercial and content development sectors. It can compromise democracy, damage people and companies, and sour relations. Although DeepFake technologies have evolved, the creation of the Deepfake generation model poses difficulties for forensics professionals trying to counter these risks. Forensics tools should be able to identify situational hazards and real-world events that compromise test accuracy, therefore exposing weaknesses in present solutions and supporting research to identify a more strong resolution.
Keywords: *Deepfake, social media, Deep learning and Fake images.*

I. INTRODUCTION

Deepfake is a technical term for fake content on social platforms. This mainly includes fake images and videos. Fake images and videos are an old tradition. Since the advent of digital visual media, there has been a desire to manipulate them. Manipulation technologies have been widely used to forge images and videos for deception and entertainment. Using professional software like Adobe Photoshop to edit an image takes knowledge, time, and work. Instead of editing software like Adobe Shop, fake videos and images can be made by machines that don't require domain knowledge. In these new images and videos, an individual's face is transformed to mimic that of a target subject resulting in an amazingly realistic image or video of events that never occurred. For example, deepfake may modify a person's appearance while preserving their facial expression. Deepfakes, made up of images, audio, and videos, seem to be the most common type of fake media. The very frst "deepfake" video was released in 2017, in which a celebrity's face was replaced with that of a porn actor. Deepfakes received attention and began to become widespread when a Reddit user known as "Deepfake" demonstrated how a renowned person's face could be modifed to give them a featured part in a pornographic video clip (Güera and Delp 2018). Deepfake is among the top fve identity fraud types in 2023. According to DeepMedia, a startup developing tools to identify fake media, the number of video deepfakes of all types has tripled, and the number of speech deepfakes has increased eightfold in 2023 compared to the same period in 2022. They have estimated that about 500,000 video and audio deepfakes will be uploaded on social media sites worldwide by the end of 2023. Deepfake media can be of diferent types based on the content that has been manipulated. These manipulations include visual, audio, and textual modifcations . Figure 1 shows types of deepfake content. Among visual, text-based, and audio, visual deepfakes are most common. They mainly include fake images and videos. As we know, today is the era of social media. These fake images and videos are used on social media platforms to spread false information about events that have never happened (Zhou and Zafarani 2020). "Face swapping", involves replacing the target's face with that of the original image, is a common method for creating deepfake images. On the other hand,

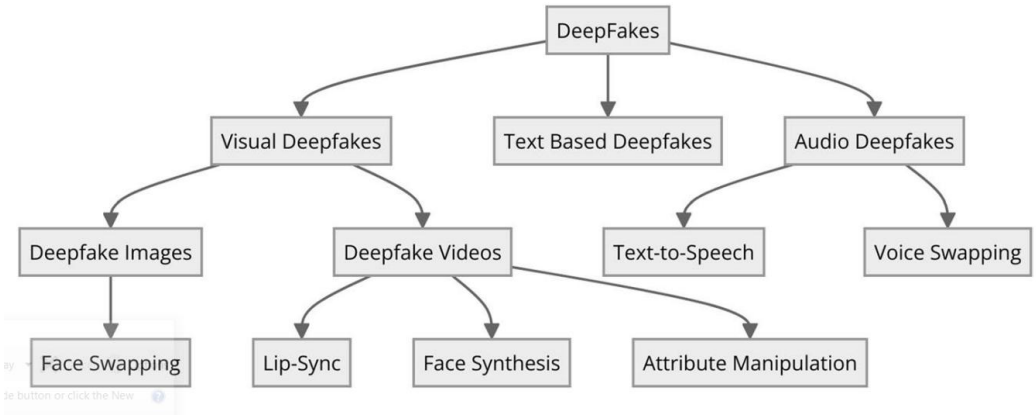

Figure I-1 *Types of Deep-fake mechanisms*

The deepfake videos may be created using three techniques: lip-sync, face synthesis, and attribute manipulation. The second type of deepfake is text-based deepfake. These textual deepfakes are mostly used on social media for fake comments and reviews on e-commerce websites. The third kind of deepfake is f as an audio deepfake. Such deepfakes involve using AI to create synthetic, realistic-sounding human speech. These deepfakes can be created using text-to-speech or voice-swapping methods. Although deepfake technology is seen from a detrimental perspective, it can also be used in some productive projects. Deepfake can potentially improve multimedia, movies, educational media, digital communications, gaming and entertainment, social media, healthcare delivery, material science, and many commercial and content development industries [4].

1.1 Malicious use of Technology

The real danger of this technology lies in the different ways it can be misused and the largescale impact it can have, courtesy of being misused. Here are the some of the threats that it may create: Threat to individual/organization: Deepfake holds great potential for inflicting tangible harm, psychological stress, physical pain and sabotaging the reputation of an individual and organization. For inflicting harm, a fraudster may use deepfake to extract something of value. To prevent the release of such deepfake, the victim provides money, personal banking details and business secrets [5]. The most common form of exploitation is in the form of deepfake pornographic videos. One can victimize the individual to any form of violent or humiliating act to gratify their wants. Threat to society: Deepfake can have a huge societal impact considering its realism and fast propagation through different social media networks. Prejudices in society are prevalent and are further aggravated by this technology when the lies are shared through different channels. Societies that are already divided based on caste, creed, religion, color and language, deepfake can further add fuel to the existing fire [6]. Threat to democracy: Deepfake can affect national and international relations; it can sour bilateral ties whose impact may last up to generations. Deepfake can prove to be very lethal, as it gives the option to external entities to influence the democratic process of a nation . Deepfake can sway the results of an election, when a fake video about a political candidate is circulated just on time, such that it has enough time to spread but narrow time to prove it faked and reverse its effect(e.g. on the eve of an election) [7]. Threat to the Business: People are losing money every year in businesses, be it the stock market or business deals; because of the disinformation. Deepfake technology allows anyone to impersonate voices of different identities like the Business leader and CEO to incur fraud. A corporate workplace that is so strict about harassment, sexual abuse, molestation, racist remark, gender discrimination, where evidence in the form of audio or video is hardly questioned. In such an environment, deepfake audio or video can be a lethal weapon that can ruin someone's 6 career and future aspirations. When deepfake media back a rumor, then such rumors can manipulate the market in such a short time and someone's may lose or make a huge profit [8]

1.2 CHALLENGES FOR DEEPFAKE CREATION AND DETECTION

In recent years, many DeepFake tools have become available that have highly realistic performance levels, and many more are in development. In contrast, the development of the DeepFake generation model is creating large challenges for forensics experts in terms of combatting them. DeepFakes are AI-generated hyperrealistic images or videos that have been digitally edited using techniques such as face swapping, changing the attributes and representing individuals speaking and doing things that never happened. GANs, which are popular artificial intelligence (AI) techniques, consist of two discriminative and generative models that compete against each other to improve their performance to generate believable fakes. These impersonations of real persons are frequently highly viral and spread swiftly across social media platforms, thereby making them an effective tool for propaganda. In digital forensics, as in other security related disciplines, it is necessary to

account for the presence of an adversary who is actively attempting to fool investigators. In reality, a knowledgeable attacker who understands the concepts on which the forensic tools are based may take a variety of counterforensic steps to avoid detection [9]. Forensics tools should be able to detect such situational threats, as well as any real-world situations that tend to degrade test accuracy. Therefore, the numerous counter forensics approaches intended to confuse current detectors are a valuable aid in the development of multimedia forensics, as they expose the flaws in current solutions and encourage research to find a more robust resolution. To date, many models are available to create or detect fakes, but they still have weaknesses. In the following sub section, we will discuss the main challenges, point by point, in creating or detecting DeepFakes.

1.3 Problem statement

The rapid advancement of **deepfake technology** has led to an increase in **manipulated digital content**, making it difficult to distinguish real from fake media. Deepfake techniques exploit **AI-driven face swapping, voice cloning, and image synthesis**, posing significant risks to **privacy, security, and misinformation spread** across social platforms. Existing **deepfake detection models** struggle with evolving forgery techniques, requiring **more robust and accurate detection mechanisms**. This research aims to develop a **Discrepancy-Aware Forgery Detection Network (DAFDN)** that effectively mitigates **representation bias, enhances feature refinement, and exploits discrepancies** for improved **deepfake detection and media authenticity verification**.

1.4 Research contribution made

- **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.

II. Related work

The authors have developed various deep learning models to detect fake content evidence in videos. They used AlexNet and VGG16 to extract features from faces, resulting in an accuracy of 94.01% on the CASIA dataset. They also used modified CNN models such as ResNext, Xception, and Ensemble to detect deep counterfeit images and videos. Other methods include a biometric-based forensic technique for deepfake detection, a quantum-inspired evolutionary-based feature selection method, and a deep CNN architecture for face forgery detection. The authors also introduced Mesonet, an efficient network designed to detect deepfake and Face2Face-tampered videos, and the Fakecatcher, a deepfake detection network that employs biological signals as an implicit descriptor of authenticity. They also compared three distinct 3D-CNN models for deepfake detection in videos and tested them on Face2Face and DeepFake first-order motion datasets. The authors also proposed a five-layer CNN model for DeepFake detection and classification, with an average prediction rate of 98% for DeepFake videos and 95% for Face2Face videos in real network diffusion cases.

Wang, Li, and Zhao proposed a combined approach using CNN for image feature extraction and SVM for deepfake prediction in video frames. The method achieved the highest AUC values and the best performance using the feature vector of edge details with SVM. El Rai et al. also developed a deepfake detection approach using residual noise and

Table 1

Ref num	Method	Advantages	Disadvantages	Research Gap
[1]	AlexNet & VGG16 for feature extraction	High accuracy (94.01%) on CASIA dataset	Limited to CASIA dataset, may not generalize	Limited dataset generalization
[2]	CNN architectures on YouTube deepfakes	Assesses generalizability on unseen data	Struggles with novel deepfake creation methods	Struggles with evolving deepfake techniques
[3]	Modified CNNs	Ensemble method improves	Computationally expensive	Computational efficiency

	(ResNext, Xception, Ensemble)	detection accuracy	for real-time detection	issues
[4]	Optical flow vectors for video forgery detection	Effective in tracking manipulation inconsistencies	Not robust against adversarial attacks	Vulnerability to adversarial attacks
[5]	Deep learning & super-resolution algorithms	Improves deepfake exposure by detecting anomalies	Limited effectiveness on highly compressed videos	Low performance on low-quality images
[6]	Face warping artifact detection	Targets artifacts without deepfake datasets	Cannot handle real-time deepfake analysis	Lack of real-time analysis
[7]	Biometric-based forensic technique	Uses facial biometrics for better detection	Limited adaptability to evolving deepfake methods	Challenges in adapting to new deepfake techniques
[8]	Quantum-inspired evolutionary feature selection	Enhances feature selection for classification	May not work well on diverse datasets	Inconsistency across different datasets
[9]	Comparison of CNN architectures for forgery detection	Comprehensive comparison of CNN models	Some models fail to capture fine-grained deepfake details	Insufficient benchmark comparison
[10]	MesoNet for detecting Face2Face & deepfakes	High success rate for Face2Face & deepfakes	Performance drops on low-resolution videos	Difficulty handling highly compressed videos
[11]	Attention-based CNN for deepfake detection	Attention mechanisms improve detection accuracy	Requires high-quality input for effective detection	Limited explainability of detection decisions
[12]	FakeCatcher using biological signals	Uses biological signals for deepfake detection	Limited dataset generalization	Lack of robustness across multiple datasets
[13]	3D CNNs (I3D, ResNet 3D, ResNeXt 3D)	3D CNNs leverage temporal relationships	Struggles with high compression & low-quality data	Struggles with real-time deepfake detection
[14]	CNN-ViT hybrid for deepfake detection	Combines CNN with ViT for better performance	Requires extensive computational resources	Computational inefficiency and scalability issues
[15]	Five-layer CNNs for deepfake classification	High prediction accuracy for real-world deepfakes	High training complexity and dataset dependency	Need for better generalization across datasets

a. Research gap

- **Limited Generalization Across Datasets**

Most existing deepfake detection models are trained on specific datasets and fail to generalize well to unseen deepfakes created with different generative techniques.

- **Robustness Against Adversarial Attacks**

Deepfake detection models are vulnerable to adversarial manipulations, where slight modifications to fake media can bypass detection systems.

- **Lack of Real-Time Detection Capabilities**

Many detection models are computationally expensive and unsuitable for real-time applications, limiting their practical deployment in social media and security systems.

- **Inconsistent Performance on Low-Quality Media**

Deepfake videos with heavy compression, blurring, or noise degrade detection accuracy, as most models rely on high-resolution artifacts for classification.

- **Explainability and Interpretability Issues**

Current deepfake detection methods function as black-box models, providing little interpretability or reasoning behind their classification decisions.

- **Limited Multimodal Deepfake Detection**

Most detection techniques focus on visual cues but fail to integrate multimodal analysis, such as combining facial, audio, and textual inconsistencies for improved detection.

- **Evolving Deepfake Generation Techniques**

Rapid advancements in AI-generated media, such as **GANs and transformers**, consistently outpace detection capabilities, making existing models obsolete quickly.

- **Lack of Benchmarking and Standardized Evaluation**

There is no universal benchmark for evaluating deepfake detection models, making it difficult to compare methodologies and establish best practices in the field.

III. 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.

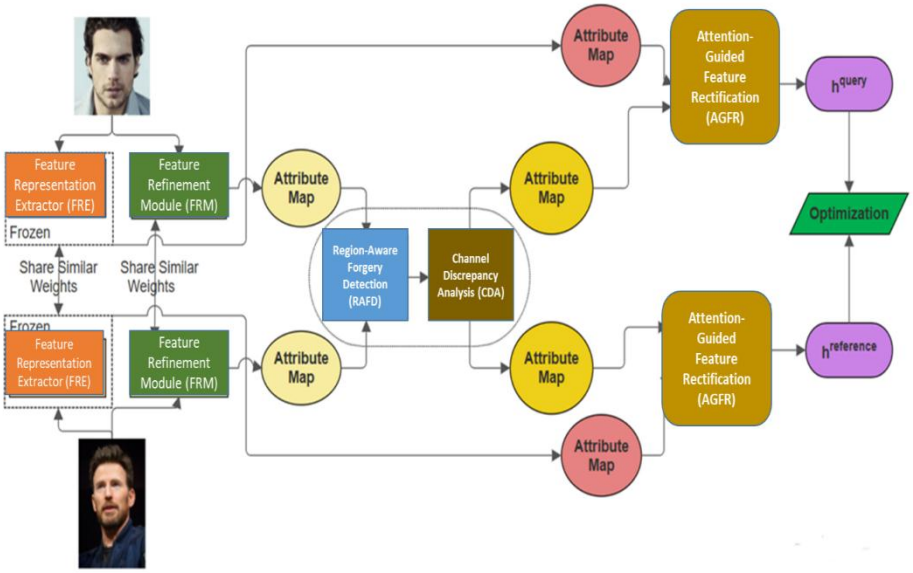

Figure 2 *Proposed Framework*

3.1 Representation Bias Rectification

The proposed methodology focuses on detecting forgery using a two-phase structure: Feature Representation Extractor (FRE) and Feature Refinement Module (FRM). The FRE phase maps facial images to representation space, while the FRM uses a consistent structure to omit representation-based bias. This approach ensures consistent mapping and bias reduction, enhancing the model's performance while balancing representation data and detection task effectiveness.

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.

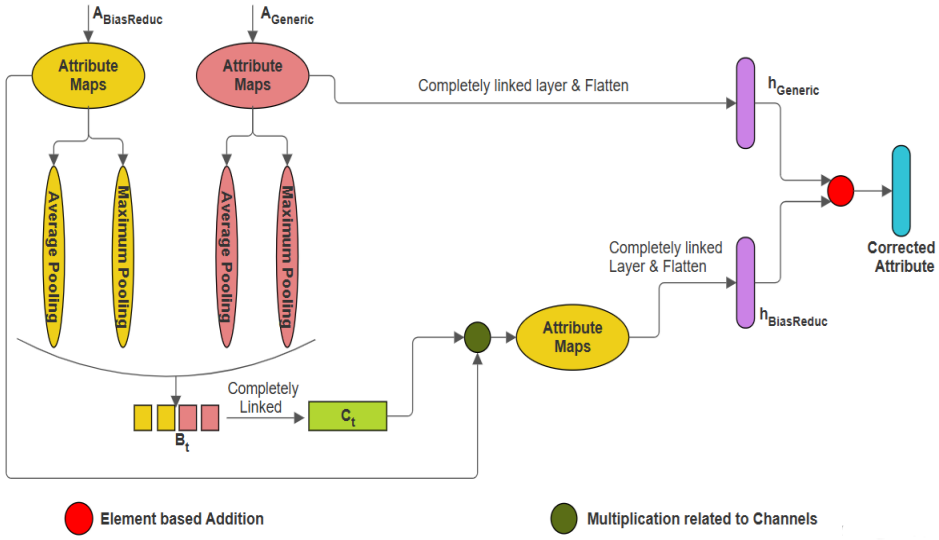

Figure 3 3.2 *Attention-Guided Feature Rectification (AGFR)*

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

$A_{Generic} = Generic(z)$	(1)
$\hat{A}_{BiasReduc} = BiasReduc(z)$	(2)

Here, the attribute maps that are retrieved is expressed as $A_{Generic}, \hat{A}_{BiasReduc}\ belongs\ to\ \mathbb{T}^{J\times Y\times E}$ 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 $\hat{A}_{BiasReduc}$ requires to be fed in the to eliminate any of the inconsistencies that could be present.

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

In the above equation, $A_{BiasReduc}\ belongs\ to\ \mathbb{T}^{J\times Y\times E}$ has similar dimensions with the input attribute map $\hat{A}_{BiasReduc}$. 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 this study, the features of these two phases are embedded into a attention correction vector. We have the input attribute maps $A_{Generic}$ and $A_{BiasReduc}$, initially we implement both global pooling average and global pooling maximum to these maps along the channel size for comprehensive data retention. Furthermore, we combine the four vectors and obtain $B_t\ belongs\ to\ \mathbb{T}^{4E}$. Additionally, a completely linked layer denoted as $\mathfrak{Z}_{CL}$ along with a sigmoid function that is used to attain an attention vector given as $C_t = \varphi(\mathfrak{Z}_{CL}(B_t))$, here $\varphi(\cdot)$ is used to express the sigmoid function. Further, the activation attention correction attribute map $A_{BiasReduc}$ is updated using $A_{BiasReduc} = C_t \otimes A_{BiasReduc}$, here $\otimes$ is used to express multiplication related to channels. Considering the attention vector given as C_t, $A_{BiasReduc}$ is derived that is capable of exploiting additional data and encourage 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 $h_{Generic}, h_{BiasReduc}\ belongs\ to\ \mathbb{T}^{E}$. Further, this is summated to obtain the corrected representation expression h of the initial input facial picture

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

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

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.

3.3.1 Region-Aware Forgery Detection (RAFD)

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 $\mathfrak{Z}_{Kernel}$ 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 $\mathfrak{Z}_{Kernel}$ for production of area aware kernels for every cross-over path. While considering an example of $AreaKernel^{query}$

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

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

3.4 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

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

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

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, $A^{reference}_{BiasReduc,l}$ is also obtained similarly.

The proposed model produces corelation of query and reference from both the channel as well as spatial point of views. This important interaction efficiently encourages the exploitation of inconsistencies and emphasises the clues for the representation based on detection of forgery.

3.5 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.

IV. Conclusion

Deepfake detection research has evolved significantly, leveraging **machine learning, deep learning, and hybrid approaches** to identify manipulated content. Various studies employ **CNN-based architectures, biometric-based forensic techniques, and attention mechanisms** to enhance detection accuracy. While models such as **MesoNet, 3D CNNs, and ViT hybrids** improve classification performance, they struggle with **dataset generalization, adversarial robustness, and real-time detection**. The primary challenges include **handling low-quality, compressed videos, integrating multimodal analysis, and countering evolving deepfake generation techniques**. The research highlights the need for **more scalable, interpretable, and computationally efficient solutions** to improve the reliability of deepfake detection in diverse applications.

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]. Z. Akhtar, "Deepfakes generation and detection: a short survey," Journal of Imaging, vol. 9, no. 1, p. 18, 2023.

[8]. "Deepfake:real threat." https://kpmg.com/kpmg-us/content/dam/kpmg/ pdf/2023/deepfakes-real-threat.pdf, 2024.
[9]. "Defense advanced research projects agency." https://www.darpa.mil/ news-events/2024-03-14, 2024.
[10]. 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.
[11]. 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.
[12]. K. Patil, S. Kale, J. Dhokey, and A. Gulhane, "Deepfake detection using biological features: a survey," arXiv preprint arXiv:2301.05819, 2023.
[13]. 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.
[14]. 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.
[15]. 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.
[16]. 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.
[17]. 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.
[18]. 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.
[19]. Prasadu Peddi, & Dr. Akash Saxena. (2016). STUDYING DATA MINING TOOLS AND TECHNIQUES FOR PREDICTING STUDENT PERFORMANCE. International Journal Of Advance Research And Innovative Ideas In Education, 2(2), 1959-1967.
[20]. 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.
[21]. 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.
[22]. 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.
[23]. 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.
[24]. Z. Shi, H. Chen, L. Chen, and D. Zhang, "Discrepancy-guided reconstruction learning for image forgery detection," in Proc. of IJCAI, 2023.
[25]. K. Lin, W. Han, S. Li, Z. Gu, H. Zhao, and Y. Mei, "Detecting deepfake videos using spatiotemporal trident network," ACM TMCCA, 2023.
[26]. 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.