Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues

Tanzila Nasrin Tazin
ID: 200488998
Under the supervision of Dr. Jingtao Yao

Abstract. The rapid advancement of realistic facial manipulation technologies has raised social concerns about potential malicious misuse. In response, this research addresses the emerging topic of face forgery detection, which has become challenging due to recent technological advancements that can forge faces beyond human perception, particularly in compressed images and videos. The proposed solution, the Frequency in Face Forgery Network (F3-Net), leverages frequency-aware clues to effectively detect face forgery patterns. F3-Net employs Discrete Cosine Transform (DCT) and a collaborative learning framework with MixBlock for integration. The Frequency-aware Decomposition (FAD) module dynamically dissects images in the frequency domain, emphasizing forgery patterns in higher-frequency components, while the Local Frequency Statistics (LFS) module analyzes frequency statistics to detect anomalies in forgery images. Evaluation of the Celeb-df-v2 dataset, including 590 authentic videos and 5,639 DeepFake videos, demonstrates F3-Net's superior performance across various compression qualities. The study establishes F3-Net as an effective solution for addressing the challenges associated with advanced face manipulation techniques, showcasing its potential for real-world applications, notably in detecting manipulated faces, especially within the context of DeepFake videos.

Keywords: Face forgery detection; DeepFake; F3-Net; Frequency

1. Introduction

In recent years, the swift progress in realistic facial manipulation technologies has sparked concerns about potential malicious uses. Consequently, there is a significant research focus on detecting manipulated faces, as these technologies have advanced to a point where they can create forgeries that surpass human perceptual abilities, especially in compressed images and videos. To address this challenge, the study introduces the Frequency in Face Forgery Network (F3-Net), a novel approach that utilizes frequency analysis to identify subtle patterns indicative of forgery. Despite the impressive sophistication of recent face manipulation techniques like DeepFake, FaceSwap, Face2Face, and NeuralTextures, which excel at concealing forgery artifacts, particularly in low-quality media, the proposed F3-Net aims to outperform existing methods, especially in the demanding area of low-quality forgery detection.

In the study, difficulties posed by compressed images and videos are acknowledged, and a solution rooted in the awareness of frequency patterns is proposed. It is found that mining forgery patterns with a focus on frequency offers a complementary perspective, capable of describing both subtle forgery artifacts and compression errors. The proposed F3-Net capitalizes on two distinct yet complementary frequency-aware clues: 1) frequency-aware decomposed image components, and 2) local frequency statistics. The application of Discrete Cosine Transform (DCT) as the frequency-domain transformation facilitates the integration of frequency information into the forgery detection process.

The research builds on the idea that decomposing an image into its frequency signals allows for the identification of subtle forgery artifacts, particularly in higher-frequency components. Additionally, local frequency statistics are introduced, which describe the frequency-aware patterns in the spatial domain and are compatible with convolutional neural networks (CNNs). The combination of these two frequency-aware clues is achieved through a two-stream collaborative learning framework, utilizing a cross-attention module named MixBlock.

- Frequency-aware Decomposition (FAD): FAD is designed with the goal of understanding and identifying forgery patterns within images by focusing on their frequency characteristics. The FAD module employs a dynamic approach to divide input images in the frequency domain, utilizing learnable frequency bands. The result is a representation of the image as a set of frequency-aware components, which can reveal patterns and anomalies associated with forgery.

- Local Frequency Statistics (LFS): LFS takes a localized approach to analyze frequency statistics within images. By extracting statistical information specific to different frequency bands, LFS aims to highlight discrepancies between genuine and forged faces. This method enables the identification of unusual statistical patterns in forgery images across various frequency bands while maintaining an understanding of the underlying structure found in natural images. The use of Convolutional Neural Network (CNN) enhances the effectiveness of mining these localized frequency statistics.

- Collaborative Learning: Collaborative Learning with MixBlock introduces a framework that fosters cooperation between the FAD and LFS modules. This collaborative learning approach is facilitated by MixBlock, a cross-attention module. MixBlock enables robust interactions between the FAD and LFS branches, allowing them to collectively learn and leverage frequency-aware clues for improved forgery detection. The integration of these modules provides a comprehensive and adaptive strategy for identifying forged images based on both global and localized frequency characteristics.

While the paper primarily focuses on the FaceForensics++ dataset, the evaluation extends to include the Celeb-df-v2 dataset, which comprises 590 authentic videos and 5,639 DeepFake videos. This dataset serves as a real-world benchmark to validate the efficacy of the proposed F3-Net in detecting manipulated faces, especially in the context of DeepFake videos. Through comprehensive studies and comparisons with state-of-the-art methods, the superior performance of F3-Net across various compression qualities is demonstrated, establishing its effectiveness in addressing the challenges posed by advanced face manipulation techniques.

2. Related Works

The increasing advancements in computer graphics and neural networks, particularly the emergence of generative adversarial networks (GANs), have led to a growing focus on detecting face forgery in our society. To tackle this issue, numerous supplementary pieces of information are employed to improve detection performance.

2.1 *Face Forgery Detection*

Detecting face forgery is a classical challenge in computer vision and graphics. Previous research focused on manual features like eye blinking, inconsistent head poses, and visual artifacts. However, with the rise of deep learning, convolutional neural networks (CNNs) have become widely employed, yielding improved performance in face forgery detection tasks. For instance, some studies used attention mechanisms to identify manipulated regions, while others explored frequency domain differences between real and forged faces. Subsequent works integrated frequency clues as a supplement to RGB information.

While these methods excel in intra-domain scenarios where training and test data distributions are similar, their performance declines significantly in unseen domains. Recent efforts have focused on general face forgery detection. One method is supervised by forged boundaries in blending operations, another employs sample weighting and gradient regularization via meta-learning, and yet another mitigates texture bias through SRM operations to prevent overfitting. However, these methods, derived from image classification models, emphasize category-level differences rather than the fundamental distinctions between real and fake images. To address these issues, a dual-granularity contrastive learning framework is introduced to manage intra-class variance and preserve transferability.

2.2 *RGB-Frequency Fusion for Manipulated Face Detection*

The manipulated face detection approach integrates RGB and frequency domain information through three key modules: Frequency-aware Cue, RGB-Frequency Attention Module (RFAM), and Multi-scale Patch Similarity Module (MPSM).

In the Frequency-aware Cue module, RGB images undergo Discrete Cosine Transform for frequency domain representation. Low-frequency information is filtered out to enhance subtle artifacts, and the processed data is inverted back to RGB color space. The RGB-Frequency Attention Module (RFAM) utilizes a two-stream network, merging RGB and frequency data. RFAM fuses these streams at different semantic layers and employs attention maps to highlight regions of interest in both domains. [3]

While these methods excel in intra-domain scenarios where training and test data distributions are similar, their performance declines significantly in unseen domains. Recent efforts have focused on general face forgery detection. One method is supervised by forged boundaries in blending operations, another employs sample weighting and gradient regularization via meta-learning, and yet another mitigates texture bias through SRM operations to prevent overfitting. However, these methods, derived from image classification models, emphasize category-level differences rather than the fundamental distinctions between real and fake images. To address these issues, a dual-granularity contrastive learning framework is introduced to manage intra-class variance and preserve transferability.

2.3 *High-Frequency Enhanced Forgery Detection Model*

The forgery detection model aims to enhance generalization through three core modules. In the Multi-scale High-frequency Feature Extraction module, high-pass filters are applied to low-level feature maps to enrich high-frequency features. This involves converting RGB input images into high-frequency residual images using SRM filters, resulting in multi-scale high-frequency feature maps. The attention guides subsequent operations on raw feature maps. The Dual Cross-modality Attention module captures interactions between low-frequency textures and high-frequency noises. By measuring correlations and generating attention maps, it re-weights features in different modalities at multiple scales. [1]

The overall model architecture includes entry, middle, and exit flows. The entry flow extracts features, the middle flow incorporates dual cross-modality attention, and the exit flow fuses features for final classification. High-level feature fusion employs channel-wise attention, and training utilizes the AM-Softmax Loss.

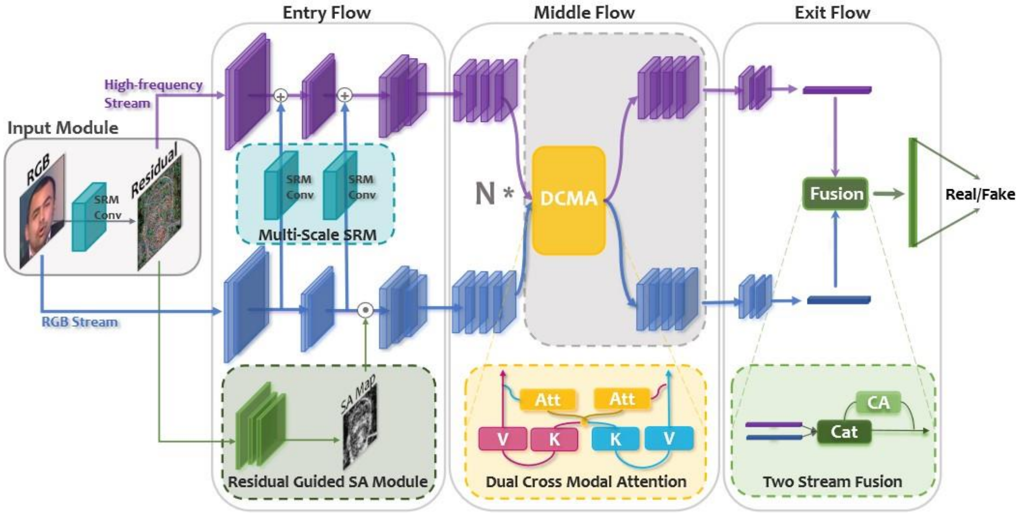


Figure 1: The suggested model follows a two-stream architecture to handle both RGB images and high-frequency noises. In the initial phase, the model captures multi-scale high-frequency features and residual guided spatial features. Moving to the intermediate stage, the model incorporates DCMA modules to facilitate interaction between feature maps from the RGB and noise modalities. Ultimately, an attention-based fusion mechanism is employed to combine features from both streams for the final classification.[1]

3. Proposed Method

In this segment, we present two novel methods for detecting forgery clues based on frequency analysis. These methods include frequency-aware decomposition (discussed in Section 3.1) and local frequency statistics (explained in Section 3.2). Following that, we introduce the cross-attention two-stream collaborative learning framework proposed in this study.

3.1 Frequency-Aware Decomposition (FAD)

FAD method for image forgery detection. Traditionally, studies used hand-crafted filter banks in the spatial domain, but these methods were limited in covering the entire frequency domain and adapting to forgery patterns. FAD is proposed as a novel approach to adaptively partition the input image in the frequency domain using learnable filters. These decomposed frequency components are then transformed back to the spatial domain and inputted into a convolutional neural network for comprehensive forgery pattern detection.

Specifically, N binary base filters are manually designed to partition the frequency domain into low, middle, and high-frequency bands. Three additional learnable filters are added to adaptively enhance the base filters. The frequency filtering is achieved through a dot-product between the frequency response of the input image and the combined filters, where a sigmoid function is applied to squeeze the result within the range of -1 to +1. The decomposed image components are obtained through this process, leveraging Discrete Cosine Transform (DCT) for its frequency distribution characteristics. [2]

Empirically, the DCT-based FAD is considered compatible with compression artifacts commonly found in forgery patterns. The DCT power spectrum of natural images reveals a non-uniform distribution, concentrated mostly in the low-frequency area. The base filters are applied to divide the spectrum into N bands, with added learnable filters providing adaptability. In practice, N is set to 3, representing low, middle, and high-frequency bands. The proposed approach aims to enhance forgery detection by effectively capturing frequency-related features and patterns.

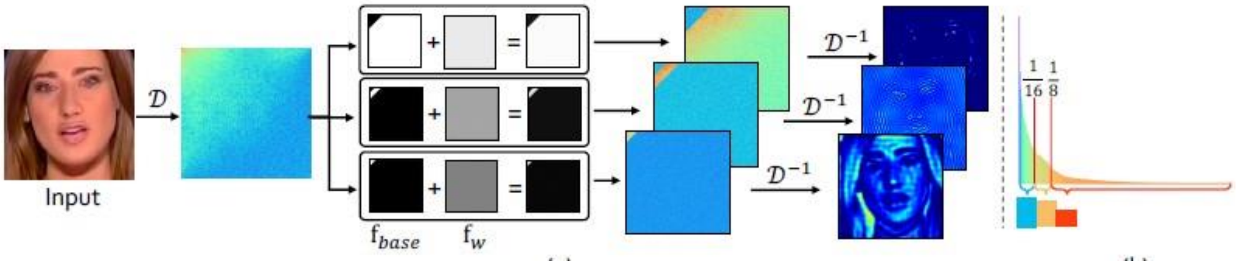


Figure 2: FAD utilizes Discrete Cosine Transform to identify significant frequency components and extracts comprehensive information by concatenating multiple frequency band components, providing a detailed analysis of the 1D power spectrum distribution.[2]

3.2 ***Local Frequency Statistics (LFS)***

While the previous Frequency-aware Decomposition (FAD) method offered a frequency-aware representation compatible with CNNs, it faced challenges in directly utilizing frequency information in the spatial domain. Acknowledging the difficulty of extracting forgery artifacts directly from spectral representations using CNN features, a new approach, local frequency statistics (LFS), is proposed. LFS employs Sliding Window Discrete Cosine Transform (SWDCT) on RGB images to extract localized frequency responses, followed by mean frequency response calculation at learnable frequency bands. The resulting statistics are then transformed into a spatial map for input into a convolutional neural network, such as Xception, facilitating the discovery of high-level forgery patterns. The LFS method aims to provide a localized aperture for detecting detailed abnormal frequency distributions, achieving a reduced statistical representation with a smoother distribution.

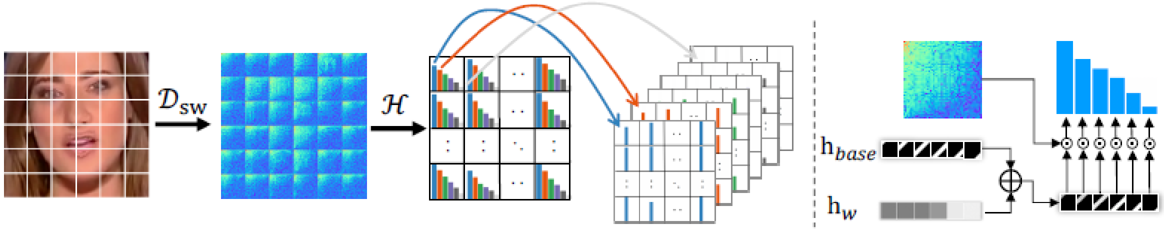


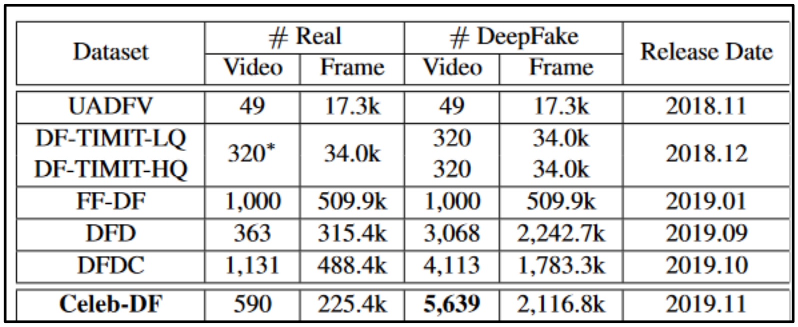
Figure 3: The introduced Local Frequency Statistics (LFS) method is designed for obtaining statistical information from the local frequency domain. SWDCT refers to the application of Sliding Window Discrete Cosine Transform, and H signifies the adaptive gathering of statistics on each grid. [2]

In Section 3.1 and Section 3.2, the FAD and LFS modules are introduced for extracting forgery clues based on frequency awareness from distinct yet inherently linked perspectives. We posit that these two types of clues exhibit differences but are mutually supportive. Consequently, we advocate for a collaborative learning framework, enhanced by cross-attention modules, to systematically integrate features from the two streams of FAD and LFS.

4. Experimental Work

4.1 *Dataset*

The Celeb-DF dataset comprises 590 authentic videos and 5,639 DeepFake videos, totaling over two million frames. The average duration of all videos is around 13 seconds, maintaining a standard frame rate of 30 frames per second. Authentic videos are sourced from publicly available YouTube interviews featuring 59 celebrities, representing diverse genders, ages, and ethnicities. Of the authentic videos, 56.8% feature male subjects and 43.2% female subjects. Age distribution includes 8.5% aged 60 and above, 30.5% between 50-60, 26.6% in their 40s, 28.0% in their 30s, and 6.4% younger than 30. Ethnicity distribution comprises 5.1% Asians, 6.8% African Americans, and 88.1% Caucasians. Authentic videos display variations in face sizes, orientations, lighting, and backgrounds. DeepFake videos are created by swapping faces among the 59 subjects, resulting in MPEG4.0 format videos. A comparison with other DeepFake datasets is presented in Table 1.



Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
UADFV	49	17.3k	49	17.3k	2018.11
DF-TIMIT-LQ DF-TIMIT-HQ	320*	34.0k	320 320	34.0k 34.0k	2018.12
FF-DF	1,000	509.9k	1,000	509.9k	2019.01
DFD	363	315.4k	3,068	2,242.7k	2019.09
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	**5,639**	2,116.8k	2019.11

Table 1: Basic Information about various DeepFake Datasets

Evaluation Metrics: Our evaluation framework employs the Accuracy score (Acc) and the Area Under the Receiver Operating Characteristic Curve (AUC) as primary metrics. Acc, inspired by celeb-DF, is a widely used metric in face forgery detection tasks. Specifically, for single-frame methods, we compute the average accuracy scores across each frame in a video. Additionally, we incorporate AUC scores, as seen in face X-ray, and similarly average them for single-frame methods.

Implementation Details: In our experiments, we utilize Xception pretrained on ImageNet as the backbone for the proposed F3-Net. The newly introduced layers and blocks are initialized randomly. The networks undergo optimization using SGD, with a base learning rate of 0.002, and employ the Cosine learning rate scheduler.

To assess the applicability of our methods in video-based scenarios, we integrate LFS and FAD into existing video-based models. The training was conducted for 20 epochs. An epoch typically refers to one complete pass through the entire training dataset. In this case, the training process involved 20 iterations through the entire dataset, with each iteration (epoch) consisting of multiple. The batch size each epoch consists of 32 batches.

Model	**Database**	**AUC**
F3-Net	FaceForensics++	98.10
F3-Net	Celeb-DF	96.24

Table 2: Cross-database evaluation from FF++ and Celeb-DF

4.2 *Compared with Previous Detection Methods:*

In our investigation of DeepFake detection methods, we considered nine distinct approaches, selected based on the availability of both code and DNN models that could be applied to the Celeb-DF dataset.

Two-stream: This method employs a two-stream CNN architecture with GoogLeNet InceptionV3 as its foundation. Trained on the SwapMe dataset, it serves as a benchmark for general-purpose image forgery detection.

Xception: Xception is a DeepFake detection method based on the XceptionNet model, with three variants trained on the FaceForensics++ dataset. These variants differ in the compression levels of H.264 videos (Xception-raw, Xception-c23, and Xception-c40).

MesoNet: MesoNet is a convolutional neural network (CNN)-based technique designed to capture mesoscopic properties of images. Two variants were explored: Meso4, which employs conventional convolutional layers, and MesoInception4, which utilizes more advanced Inception modules. The models were trained on unpublished DeepFake datasets curated by the authors.

HeadPose: The HeadPose method focuses on detecting DeepFake videos by scrutinizing inconsistencies in head poses. It utilizes a Support Vector Machine (SVM) model trained on the UADFV dataset.

FWA (Face Warping Artifacts): This method utilizes a ResNet-50 architecture to expose face warping artifacts induced by resizing and interpolation operations. The model is trained on self-collected face images. [4]

VA (Visual Artifacts): VA is a recent DeepFake detection method that targets visual artifacts in the eyes, teeth, and facial contours of synthesized faces. Two variants, VA-MLP and VA-LogReg, leverage different classifiers and are trained on an unpublished dataset, with real images from the CelebA dataset and DeepFake videos from YouTube.

Multi-task: This approach employs a CNN model to simultaneously detect manipulated images and segment manipulated areas as a multi-task learning problem. The model is trained on the FaceForensics dataset.

Capsule: Capsule utilizes capsule structures based on a VGG19 network for DeepFake classification. The model is trained on the FaceForensics++ dataset.[5]

DSP-FWA: An enhanced version of FWA, DSP-FWA incorporates a spatial pyramid pooling (SPP) module to better handle variations in the resolutions of the original target faces. This method is trained on self-collected face images.

In summary, these methods represent a diverse set of techniques, encompassing different CNN architectures, strategies for capturing various image properties, and approaches to identifying artifacts associated with DeepFake generation. The datasets used for training vary, including both publicly available datasets and unpublished datasets collected by the authors, contributing to a comprehensive exploration of DeepFake detection capabilities.

4.3 *Ablation studies*

To assess the effectiveness of the proposed methods (LFS, FAD), a quantitative evaluation was conducted on F3-Net and its variants.

The result, demonstrate consistent improvement in accuracy (Acc) and area under the curve (AUC) scores when incorporating FAD into the baseline (model 2). Further enhancement is observed when adding LFS (model 4) to the FAD-based model. The best performance is achieved by integrating MixBlock into the two-branch structure, yielding AUC scores of 96.24%. This is not progressive improvements affirm that FAD and LFS modules contribute significantly to forgery detection and are complementary. The introduction of MixBlock enhances collaboration between FAD and LFS, leading to additional gains.

F3-Net performs best at lower false positive rates (FPR), a challenging scenario for forgery detection systems. To gain insights into the effectiveness of the proposed methods, feature maps extracted by Xception and F3-Net. The discriminative capability of F3-Net is evident, with clear distinctions between real and forged faces in the feature distributions. In contrast, Xception's feature maps exhibit similarity and indistinguishability between real and forged faces.

5. Future Work

The presented work on F3-Net for face forgery detection is comprehensive and innovative. However, there are always possibilities for improvement and further development. Here are some suggestions:

- **Model Robustness through Data Augmentation:** Integrate advanced data augmentation techniques to simulate diverse real-world conditions, thereby improving the model's ability to generalize and detect face forgeries under various scenarios.
- **Optimized Training Dynamics:** Investigate adaptive learning rate schedules, architecture variations, and regularization techniques to fine-tune the training process, prevent overfitting, and potentially boost the overall performance of the F3-Net for face forgery detection.
- **Advanced Fusion Techniques:** Explore and implement sophisticated fusion methods within the cross-attention module, possibly incorporating attention mechanisms or additional contextual information to enhance collaboration between frequency-aware branches.

6. Conclusion

This paper introduced a novel and effective approach, F3-Net, for face forgery detection by leveraging the frequency-aware fusion of multi-scale features. The proposed method successfully addressed the challenges associated with diverse forgery manipulation techniques, providing a robust solution for the identification of manipulated facial images. Through the incorporation of a cross-attention mechanism and a carefully designed frequency decomposition strategy, F3-Net demonstrated state-of-the-art performance on benchmark datasets, showcasing its ability to generalize across various forgery scenarios. Moreover, the evaluation conducted on the FaceForensic++ dataset demonstrated that F3-Net achieved even more accurate results compared to the Celeb-DF dataset. This highlights the model's robustness and superior performance across a variety of datasets, further emphasizing its effectiveness in face forgery detection. The experimental results underscored the efficacy of F3-Net in comparison to existing methods, highlighting its superior accuracy and resilience to common face manipulation attacks. The comprehensive evaluation also revealed the model's proficiency in handling real-world variations and showcased its potential for deployment in practical applications where face forgery detection is crucial. The presented work contributes to the evolving field of digital forensics by providing a sophisticated solution to the escalating challenges posed by face manipulation techniques. F3-Net's combination of innovative design elements and superior performance positions it as a noteworthy advancement in the pursuit of accurate and reliable face forgery detection. As technology continues to advance, the findings and methodologies presented in this paper pave the way for future developments in the ongoing battle against digital image manipulation and fraudulent practices.

Acknowledgments

I extend my deepest appreciation to my advisor Dr. Jingtao Yao, whose invaluable insights, constructive feedback, and unwavering support have been instrumental in shaping this work.

References

[1] Luo, Y. (2021, March 23). *Generalizing Face Forgery Detection with High-frequency Features*. arXiv.org. https://arxiv.org/abs/2103.12376

[2]Qian, Y. (2020, July 18). *Thinking in frequency: Face forgery detection by mining Frequency-aware clues*. arXiv.org. https://arxiv.org/abs/2007.09355

[3] Frank, J., Eisenhofer, T., Schönherr, L., & Holz, T. (2020b). Leveraging frequency analysis for deep fake image recognition. *ResearchGate*. https://www.researchgate.net/publication/340049392_Leveraging_Frequency_Analysis_for_Deep_Fake_Image_Recognition

[4] Rössler, A. (2019, January 25). *FaceForensics++: Learning to Detect Manipulated Facial Images*. arXiv.org. https://arxiv.org/abs/1901.08971

[5] Taeb, M., & Chi, H. (2022). Comparison of Deepfake Detection Techniques through Deep Learning. *Journal of Cybersecurity and Privacy*, *2*(1), 89–106. https://doi.org/10.3390/jcp2010007

[6] Li, Y. (2020). *Celeb-DF: a Large-Scale challenging dataset for DeepFake forensics*. https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html

[7] Liu, H. (2021). *Spatial-Phase shallow Learning: Rethinking face forgery detection in frequency domain*. https://openaccess.thecvf.com/content/CVPR2021/html/Liu_Spatial-

Phase_Shallow_Learning_Rethinking_Face_Forgery_Detection_in_Frequency_Domain_CVPR_2021_paper.html

[8] Masi, I. (2020, August 8). *Two-branch recurrent network for isolating deepfakes in videos*. arXiv.org. https://arxiv.org/abs/2008.03412

[9] *Exploiting visual artifacts to expose deepfakes and face manipulations*. (2019, January 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8638330

[10] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2019). On the Detection of Digital Face Manipulation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.01717

[11] Theobalt, C. (2016, September 5). *US20180068178A1 - Real-time Expression Transfer for facial reenactment - Google Patents*. https://patents.google.com/patent/US20180068178A1/en

[12] Wang, X. (2020). *Face manipulation detection via auxiliary supervision*. https://www.semanticscholar.org/paper/Face-Manipulation-Detection-via-Auxiliary-Wang-Yao/9d86e6f5db94186e81fc7b1fc0d53ada8d03584a

[13] Holub, V., & Fridrich, J. (2013). Random Projections of Residuals for Digital Image steganalysis. *IEEE Transactions on Information Forensics and Security*, *8*(12), 1996–2006. https://doi.org/10.1109/tifs.2013.2286682

[14] *ImageNet: A large-scale hierarchical image database*. (2009, June 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/5206848

[15] Huang, H. (2017). *Wavelet-SRNET: a Wavelet-Based CNN for Multi-Scale Face Super Resolution*. https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Wavelet-SRNet_A_Wavelet-Based_ICCV_2017_paper.html