Interpretable DeepFake Detection Based on Frequency Spatial Transformer

Tao Luan [1], Guoqing Liang [2] and Pengfei Pei [3]
[1]The Institute of Software, Chinese Academy of Sciences
[2]Taiyuan Coal Gasification (Group) Co., Ltd. No. 29 Heping South Road, Wanbailin District, Taiyuan City, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100085, China
peipengfei@iie.ac.cn

***Abstract*—In recent years, the rapid development of DeepFake has garnered significant attention. Traditional DeepFake detection methods have achieved 100% accuracy on certain corresponding datasets, however, these methods lack interpretability. Existing methods for learning forgery traces often rely on pre-annotated data based on supervised learning, which limits their abilities in non-corresponding detection scenarios. To address this issue, we propose an interpretable DeepFake detection approach based on unsupervised learning called Find-X. The Find-X network consists of two components: forgery trace generation network (FTG) and forgery trace discrimination network (FTD). FTG is used to extract more general inconsistent forgery traces from frequency and spatial domains. Then input the extracted forgery traces into FTD to classify real/fake. By obtaining feedback from FTD, FTG can generate more effective forgery traces. As inconsistent features are prevalent in DeepFake videos, our detection approach improves the generalization of detecting unknown forgeries. Extensive experiments show that our method outperforms state-of-the-art methods on popular benchmarks, and the visual forgery traces provide meaningful explanations for DeepFake detection.**

***Index Terms*—Interpretable DeepFake Detection, Unsupervised Learning, Forgery Traces, Frequency-spatial Traces**

I. Introduction

The malicious use of DeepFake technology can inflict harm on personal reputation and property [1], [2]. Existing detection methods have made significant strides in accuracy. However, traditional DeepFake detection methods only provide probability values, lacking interpretability. Furthermore, these methods perform poorly on independently tested unrelated datasets [3], [4]. Some approaches attempt to yield visual results and interpretability for detection. They rely on pre-annotated forgery regions, visualizing DeepFake's forgery traces through supervised learning [2], [5], [6], [7]. Nevertheless, depending on pre-annotated forgery traces limits detection performance for unknown forgery methods. Hence, there is a need to explore unsupervised and interpretable DeepFake detection methods.

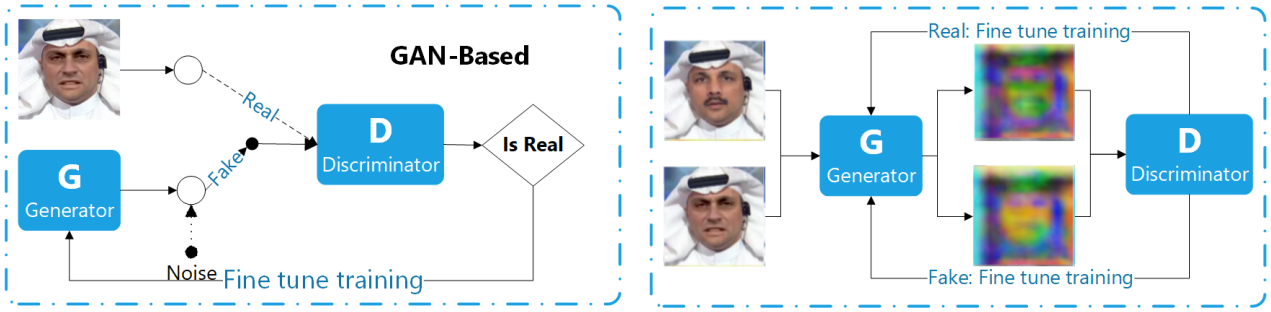
Learning to visualize inconsistency traces from datasets lacking annotated forgery traces presents a challenge, particularly for unknown facial inconsistency forgery traces. We observe that DeepFake-generated facial regions inevitably exhibit boundary artifacts, and the pixel distribution in forged regions differs from the original sources, leading to distinct statistical feature differences. These differences between the forged and original regions are widely present in various DeepFake videos. Based on these observations, we detect forged edges in the spatial domain and analyze pixel statistical distributions. Additionally, we capture frequency domain region correlations to provide interpretable visualizations of forgery traces.

Our method's design philosophy is illustrated in Figure 1. In Figure 1a, we employ a GAN-based approach, generating realistic images through the adversarial interplay between the generator network G and the discriminator network D. In Figure 1b, we adopt a cooperative approach, generating unsupervised visual forgery traces through non-adversarial collaboration between the generator network G and the discriminator network D. We address this challenge through a two-stage learning network.

Specifically, we propose an interpretable DeepFake detection method named Find-X, consisting of two networks: the Forgery Traces Generation network (FTG) and the Trace Judgment Classification network (FTD). Initially, FTG generates multi-view visualized forgery traces of edges, pixels, and regions, denoted as g. Subsequently, the visualized forgery traces g generated by FTG are input to FTD to classify them as either real or fake, producing classification results d. These supervised binary classification results of d are then fed back to FTG to generate improved visualized forgery traces g. Since the generation of g relies solely on the binary classification results d, without the need for pre-annotated forgery traces, our method is applicable to a wide range of DeepFake detection tasks.

Extensive experiments conducted on multiple datasets demonstrate that our method effectively visualizes various forgery traces in different types of DeepFake videos and outperforms existing methods in terms of detection. In summary, the main contributions of this paper are as follows:

- We propose an unsupervised DeepFake detection method, Find-X, which incorporates interpretable visualizations of forgery traces by leveraging multi-view learning of forgery features.
- We achieve interpretable results using unsupervised learning methods and distinguish forgery traces produced by different techniques. Compared to existing interpretable DeepFake detection methods, our approach does not rely on pre-annotated forgery traces, making it suitable for a



(a) Adversarial GAN-Based Generation (b) Cooperative Visual Trace Generation

Fig. 1: (a) GAN-based approach to generate unsupervised images through the adversarial interplay between the generation network G and the discriminator network D. (b) We adopt a cooperative approach to facilitate the generation of superior forged trace images through the collaboration of the generation network G and the discriminator network D.

wide range of DeepFake visualization and interpretation tasks.

- We conduct experiments on multiple datasets and compare our method with state-of-the-art approaches. The experimental results demonstrate that Find-X outperforms existing methods in terms of detection. Additionally, Find-X effectively visualizes forgery traces of different types of DeepFake videos and provides corresponding interpretable visualizations.

II. Related Works

A. Traditional DeepFake Forgery Detection

Traditional approaches mainly aim to improve binary classification accuracy [8], [1], [9], [10]. LRNet [9] detects DeepFakes by analyzing facial movements and subtle unnatural expressions in temporal dimension. MA [1] utilizes multiple attention mechanisms to capture features from various facial regions, enhancing detection accuracy. DropoutViT [11] introduces a spatio-temporal Dropout Transformer for data augmentation, improving model robustness. Inconsistency-based methods exploit physiological characteristics to detect forged videos. Early approaches focused on visual cues from low-quality DeepFake videos [8], [12], such as blink frequency and facial symmetry. However, with improved video quality, some obvious forgery traces have been addressed. Therefore, recent methods incorporate visual, audio, and motion features for DeepFake detection[13], [14], [15].

B. Interpretable DeepFake Forgery Detection

Research on the interpretability of DeepFake detection has received significant attention [16], [2], [7], [6], [5]. Face X-Ray [16] detects forged image blending boundaries by decomposing input images into a mixture of two sources. RFM-Net [2] improves detection performance by masking sensitive facial regions and focusing on informative areas. MaskRelation [7] captures relational information from masked facial regions, reducing redundancy. FakeLocator [6] exploits flaws in GAN-generated faces to detect full-resolution face forgery videos. ISTVT [5] incorporates a spatio-temporal video transformer for robust DeepFake detection, capturing spatial artifacts and temporal inconsistencies. However, these methods either lack generality or visual interpretability, limiting their practical use.

III. Method

Find-X is a frame-level forgery detection method. As depicted in Fig. 2, Find-X comprises two networks: FTG and FTD. FTG enhances multi-view features through frequency-spatial aware branches and utilizes PoolFormer to extract multi-scale forgery traces. FTD judges the results generated by FTG and provides feedback on the supervised learning outcomes to the FTG generation module, enabling the learning of forgery trace features without relying on pre-annotated data. Since FTD's discrimination between real and fake depends on the visualized forgery traces output by FTG, the binary classification results of FTD contribute to generating more reliable forgery traces by FTG.

A. Forgery Traces Generation

1) Face Preprocessing: Face preprocessing is a commonly used technique in DeepFake detection to enhance detection accuracy. In order to ensure fairness and reproducibility, we utilize the open-source face recognition tool MTCNN [17] and integrate it into our publicly available implementation of Find-X. Additionally, we adopt a criterion of selecting faces with larger pixel width and higher quality from the input videos to further improve the preprocessing stage.

2) Spatial-Aware Branch: We extract forgery boundary traces and pixel statistical features from the spatial-aware branch. To capture forgery boundary traces, we utilize edge detection techniques such as the Sobel operator [18] and the Laplacian operator [19]. We specifically choose the Laplacian operator [19] for its rotational invariance property. For pixel statistical feature extraction, we employ the SRM operator [20] to enhance the detection of abnormal pixel distributions within manipulated videos, as it is sensitive to the continuous statistical attributes of pixels. The spatial-aware module is implemented by incorporating these spatial operators into the CNN filter kernels.

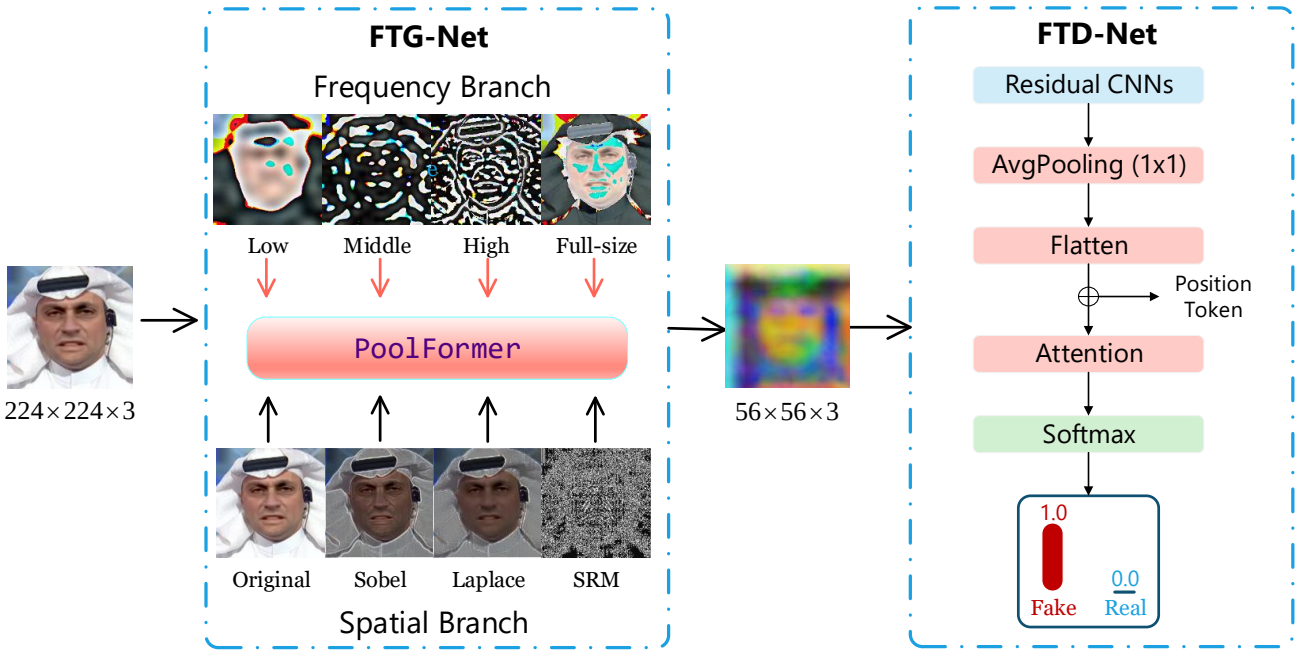


Fig. 2: Overview of the proposed framework architecture diagram for IFTV-Net, which includes two network FTG and FTD. FTG generates forged traces, and then inputs them into FTD to output the probability of authenticity.

3) Frequency-Aware Branch: The frequency-aware branch is dedicated to capturing forgery traces that are not easily detectable in the spatial domain. The Discrete Cosine Transform (DCT) is used to divide an image into small blocks consisting of different frequencies. During quantization, high-frequency components that are less perceptible are discarded, while the low-frequency components are retained for image reconstruction. The DCT frequency domain information varies across videos from different sources. Leveraging this observation, we convert the video frame information into DCT frequency domain information using varying block sizes to identify inconsistencies between genuine and manipulated regions. PyTorch was employed to implement CNN filters based on DCT, enabling the transformation of spatial signals into the frequency domain and extraction of frequency features from videos. Different block sizes of DCT are beneficial for capturing various types of information. Smaller block sizes are suitable for capturing details and motion information, while larger block sizes are more appropriate for capturing broader spatial information. To capture distinct DCT characteristics of the video, we incorporated four different block sizes of DCT filters: full-size, large-size, medium-size, and small-size. The strong correlation features exhibited by DCT facilitate the detection of forged traces that are imperceptible in the spatial domain.

4) Multi-Scale Feature Learning: We utilize PoolFormer to gradually extract multi-view features of ”edges,” ”pixels,” and ”regions” from coarse to fine. PoolFormer[21] is a multi-scale image feature extraction network that combines the advantages of CNN’s local feature friendliness and ViT’s sequence feature friendliness. PoolFormer effectively leverages various meta-information to enhance fine-grained recognition performance.

B. Forgery Traces Discrimination

We design a compact classification model, FTD, consisting of a Transformer module and several simple residual CNN modules. This is because FTD receives forged traces generated by FTG, which are of size 56x56, smaller than the original video features. Since FTG replaces the feature extraction stage of traditional DeepFake detection methods, there is no need for a larger classification network. Additionally, employing a smaller classification network facilitates easier feedback of results to the FTG network.

IV. Experiments

A. Experimental Setup

1) Dataset: We select several popular datasets to evaluate our method. FaceForensics++ (FF++)[22] offers a variety of different forgery methods, making it particularly suitable for assessing the visualized forgery results of Find-X with various types of forgery traces. FF++ dataset includes 1000 real videos and an equal number of forged videos generated by state-of-the-art DeepFake methods, namely Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and FaceShifter (FSh). The dataset provides a balanced distribution of original and forged videos. The DeepFakeDetection dataset (DFD) [23] contains videos from 28 actors, consisting of 363 original videos and 3068 fake videos generated by basic

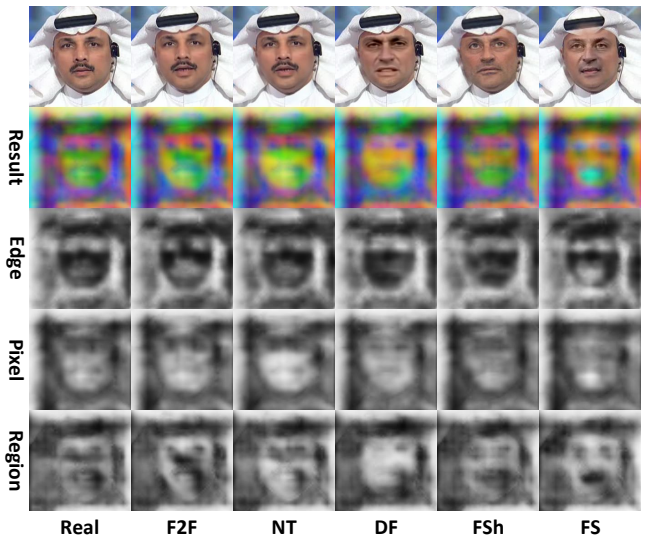


Fig. 3: Visualization results of five different forgery methods and real videos in the FF++ dataset from multiple perspectives, including 'edges', 'pixels', and 'regions'.

DeepFake methods. The dataset offers a varied set of actors for evaluation purposes. The Celeb-DF dataset [24] is part of the DeepFake Detection Challenge and consists of 590 real videos and 5639 fake videos with high visual quality. The dataset is used for developing and evaluating DeepFake detection algorithms, providing a large number of synthetic clips for analysis.

2) Baseline Methods: To evaluate the robustness of our method in binary classification detection across videos of various qualities, we compare it with six existing methods on all sub-datasets of the FF++ dataset, including Face X-ray[16], Xception[25], F3-Net[26], EfficientNet-B4[27], MA(Xception)[1], and MA(Efficient-B4) [1]. Additionally, we assess the performance of our method compared to the latest methods on different datasets, including Celeb-DF, DFD, and FF++ (DF) datasets. The state-of-the-art methods for comparison are Xception[22], DILNet[28], Grad-CAM[29], DIANet[30], STIL[31], FInter[32], and ViTHash [33].

3) Implementation Details: Our model is implemented by PyTorch, and the code has been released on GitHub. We use *ffmpeg* to extract frames from videos and train the model on a single NVIDIA RTX 3090 24GB GPU card. Each model is trained for 2-5 epochs depending on the scale of the dataset. We apply the Adaptive Moment Estimation (ADAM) optimizer with a learning rate of $1e-5$, which is computationally efficient, requires less memory, and performs well on large-scale datasets.

B. Visualization Results Analysis

1) Visualization Results of Different Forgery Methods: Figure 3 presents the evaluation results on the FF++ dataset, showcasing the visual interpretability provided by Find-X for videos generated using five different forgery methods (DF, F2F, FS, NT, and FSh) based on a common real video. The colored 'Result' represents the overall forgery traces on the manipulated face. 'Edge' indicates forgery traces on the face's edges, 'Pixel' depicts forgery traces based on pixel distribution, and 'Region' showcases localized results after frequency-domain feature enhancement of the entire manipulated face region. The visual analysis reveals distinct differences between the real video and the forged videos, with the latter exhibiting abnormal facial distribution and deficiencies in facial features. Find-X effectively discerns inconsistencies across different forged videos in an unsupervised manner. For Deepfakes, rectangular forged regions stand out, as indicated by the 'Region' part of the DF results. F2F exhibits inconsistencies in facial expressions, particularly evident in the 'Region' part. FS demonstrates significant distortions in the overall region of the manipulated face, with notable distortions in the eyes and mouth. NT shows pronounced alterations near the mouth region, leading to significant changes in the mouth features. FSh results indicate significant changes in facial structure, with minimal resemblance to a face in the 'Edge', 'Pixel', and 'Region' parts. The evaluation on the FF++ dataset demonstrates Find-X's ability to provide detailed visual explanations for different forgery methods, enhancing our understanding of the forged traces and their impact on facial features.

2) Impact of Branching on Results: Figure 4 presents the results of the ablation experiment conducted on the FF++ (DF) dataset to visualize the forgery traces. The visualization results without feature enhancement (None) exhibit limited interpretability and fail to capture facial features. The spatial branch shows sensitivity to structural changes in facial features, which is helpful for visualizing forgery traces. In contrast, the frequency branch effectively detects forgery in the mouth region but lacks detailed facial features. On the other hand, the twin branch (Twin) combines the advantages of both branches and can effectively visualize forgery traces. The experimental results indicate that the spatial and frequency branches enhance forged traces from multiple views, resulting in improved visualization of forged traces.

C. Comparison Experiments

To validate the accuracy of Find-X in binary classification, we compare its performance with state-of-the-art methods. We first compare the accuracy (ACC) of our method with the state-of-the-art methods on the FF++ dataset at different video qualities. As shown in Table I, our method outperforms the related methods, particularly in the evaluation of low-quality (LQ) videos. This is mainly attributed to the degradation of forgery traces caused by high compression, our proposed Find-X effectively mitigates the impact of video compression by enhancing the features of forgery traces.

Furthermore, we conduct comparative experiments with recent works on widely used datasets including Celeb-DF, DFD, and FF++ (F2F). For DFD, we train on the c23 compressed dataset and test on the c40 compressed dataset. For FF++, we train on the raw dataset and test on the c23 dataset. As shown in Table II, we achieve the best performance on

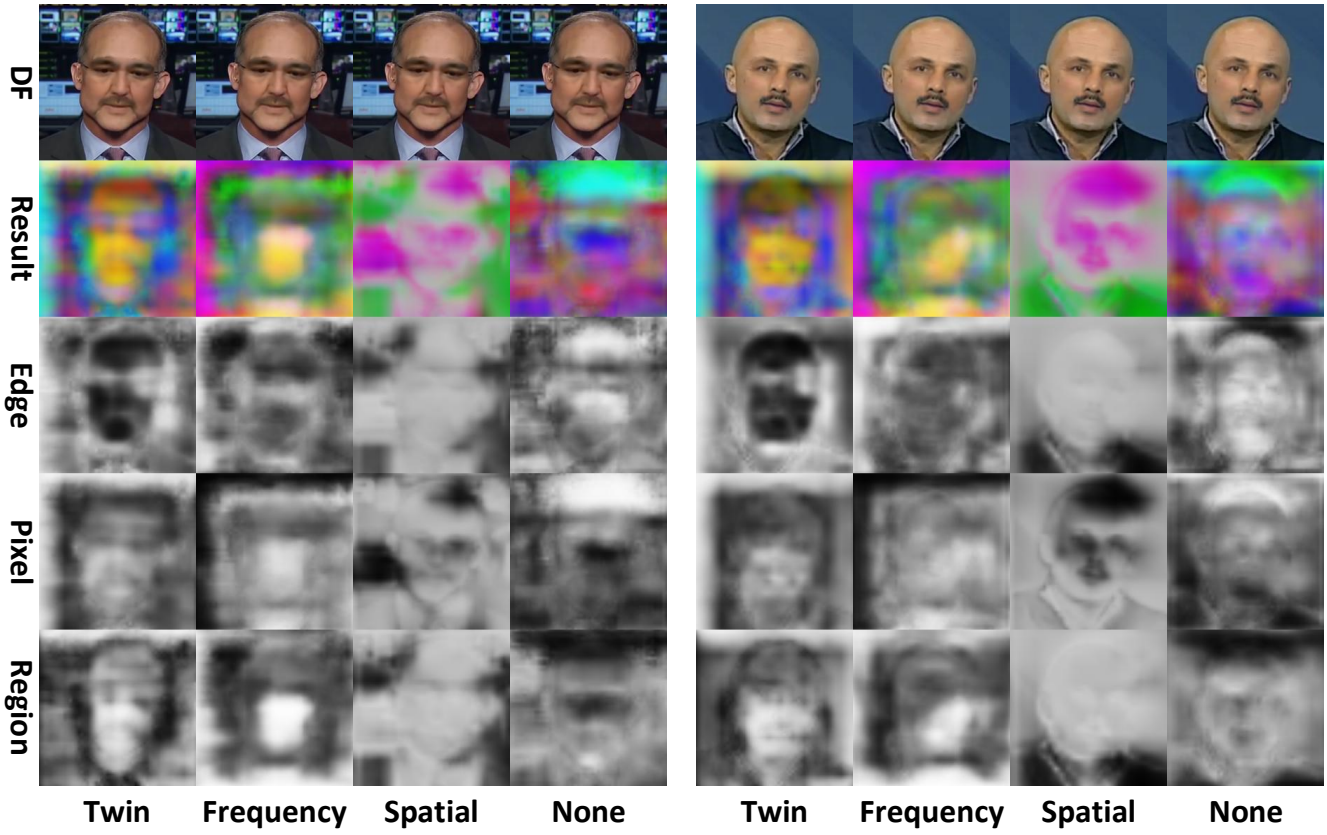


Fig. 4: Evaluation of the visual explanations of the spatial-frequency twin-branch through ablation experiments on the FF++ (DF) dataset.

TABLE I: Comparison experiment of fine-grained accuracy (ACC) with recent works on FaceForensics++ High Quality (HQ) and Low Quality (LQ) datasets.

Method	FF++ (HQ)				FF++ (LQ)				Celeb-DF
	DF	F2F	FS	NT	DF	F2F	FS	NT	
Xception	98.9	98.9	99.6	95.0	96.8	91.1	94.6	87.1	99.4
I3D	92.9	92.9	96.4	90.4	91.1	86.4	91.4	78.6	99.2
LSTM	**99.6**	99.3	98.2	93.9	96.4	88.2	94.3	88.2	95.7
TEI	97.9	97.1	97.5	94.3	95.0	91.1	94.6	90.4	99.1
ADDNet-3d	92.1	83.9	92.5	78.2	90.4	78.2	80.0	69.3	95.2
S-MIL	98.6	99.3	99.3	95.7	96.8	91.4	94.6	88.6	99.2
S-MIL-T	**99.6**	99.6	**100.0**	94.3	97.1	91.1	96.1	86.8	98.8
STIL	**99.6**	99.3	**100.0**	95.4	98.2	92.1	97.1	91.8	99.8
VTN	**99.6**	99.3	99.6	95.4	97.9	92.1	95.7	90.4	99.3
ISTVT	**99.6**	99.6	**100.0**	96.8	98.9	96.1	97.5	92.1	99.8
Ours	99.4	**100.0**	99.8	**99.4**	**99.4**	**99.6**	**99.5**	**98.5**	**99.9**

TABLE II: Comparison with state-of-the-art methods on three public datasets: Celeb-DF, DFD and F2F of FF++.

Method	Celeb-DF	DFD	FF++ (F2F)
Xception	99.4	-	95.5
DILNet	99.6	-	98.1
Grad-CAM	79.4	0.919	99.2
DIANet	-	-	90.4
STIL	99.6	-	99.6
FInter	90.5	-	95.7
ViTHash	99.4	0.963	99.9
Ours	**99.9**	**100**	**100**

several datasets compared to other methods, even reaching 100% accuracy.

D. Multi-class Evaluation Experiment

1) Multi-Class Results Evaluation: Distinguishing subtle differences between face-swapped videos of the same person generated by different forgery methods is challenging due to their imperceptible nature. We evaluate the interpretability of Find-X for multiple fake videos on five forgery methods (DF, F2F, FS, NT, and FSh) and real videos from the FF++ dataset. Table III shows that Find-X achieves high accuracy

TABLE III: Evaluation on FF++ with five different forgery methods by training on FF++ raw for multi-classification.

Compression	Training/Test Set (ACC)					
	Real	DF	F2F	FS	NT	FSh
Raw	99.4	99.3	99.0	99.4	99.1	99.3
C23	97.8	99.5	98.8	99.5	99.0	99.5
C40	31.8	99.8	96.7	99.1	99.0	99.8

TABLE IV: Ablation study with spatial and frequency branches on the FF++ dataset for multi-classification, we train on the raw data and test on the c23 data.

Branch	Training/Test Set (ACC)				
	DF	F2F	FS	NT	FSh
None	98.5	94.1	97.6	95.1	96.7
Frequency	98.0	94.8	98.3	86.1	93.7
Spatial	98.1	98.4	95.1	91.5	89.3
Twin	**99.5**	**98.7**	**99.5**	**98.9**	**99.5**

(average 95.3%, highest 99.8%) in differentiating between forgery types, demonstrating its effectiveness in discerning subtle differences.

2) Robustness of Compression: To evaluate the robustness of Find-X against video compression, we perform multi-class compression performance experiments on the FF++ dataset. As shown in Table III, the accuracy of detecting various types of forged videos slightly decreases as the compression ratio increases. Notably, the accuracy of c40 results for real videos significantly drops to 0.318 compared to forged videos. This discrepancy is likely due to the absence of potential forgery traces in real videos, which makes the enhancement of spatial and frequency features less effective. The experiment demonstrates the strong resilience of Find-X to video compression, with the effective enhancement of forged traces during the feature enhancement stage.

V. Ablation Study

Table IV illustrates the multi-classification results on the five subsets of the FF++ dataset, aiming to evaluate the necessity of the spatial and frequency branches. The models are trained on the raw data and tested on the c23 data. Using the classification results without feature enhancement (None) as the baseline, it is observed that the spatial branch (Spatial) has minimal impact on the results, while the frequency branch (Frequency) notably enhances the performance. The combined learning of the spatial and frequency branches (Twin) yields the best performance. The experiments demonstrate that joint learning of spatial and frequency branches for multi-view feature extraction can improve the detection accuracy.

VI. Conclusion

In this paper, we introduce Find-X, a novel framework consisting of two integrated networks, for visually explaining DeepFake detection results by highlighting forged traces. The approach goes beyond traditional methods that provide only probability values by offering intuitive visual explanations. By leveraging unsupervised forgery trace learning, our method provides visualizable and interpretable results, making it applicable to various types of DeepFake detection methods. By enhancing multi-view (edge, pixel, and region) features, Find-X improves detection accuracy and exhibits excellent visualization of forged traces. Extensive experiments demonstrate the effectiveness and robustness of Find-X compared to existing approaches. Our method provides valuable visual explanations for DeepFake detection.

References

[1] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *CVPR*. virtual: IEEE, 2021, pp. 2185–2194.

[2] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *CVPR*. virtual: IEEE, 2021, pp. 14 923–14 932.

[3] P. Pei, X. Zhao, Y. Cao, and C. Hu, "Visual explanations for exposing potential inconsistency of deepfakes," ser. Lecture Notes in Computer Science, X. Zhao, Z. Tang, P. C. Alfaro, and A. Piva, Eds., vol. 13825. Springer, 2022, pp. 68–82.

[4] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *ICCV*, Montreal, QC, Canada, 2021, pp. 15 088–15 097.

[5] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: interpretable spatial-temporal video transformer for deepfake detection," vol. 18, pp. 1335–1348, 2023.

[6] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu, "Fakelocator: Robust localization of gan-based face manipulations," vol. 17, pp. 2657–2672, 2022.

[7] Z. Yang, J. Liang, Y. Xu, X. Zhang, and R. He, "Masked relation learning for deepfake detection," vol. 18, pp. 1696–1708, 2023.

[8] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *CVPR*. virtual: IEEE, 2021, pp. 6458–6467.

[9] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *CVPR*. virtual: IEEE, 2021, pp. 3609–3618.

[10] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *ICCV*, Montreal, QC, Canada, 2021, pp. 15 003–15 013.

[11] D. Zhang, F. Lin, Y. Hua, P. Wang, D. Zeng, and S. Ge, "Deepfake video detection with spatiotemporal dropout transformer," in *ACM MM*, J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds., 2022, pp. 5833–5841.

[12] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, "Deepfacelab: A simple, flexible and extensible face swapping framework," *CoRR*, vol. abs/2005.05535, 2020.

[13] Y. Gu, X. Zhao, C. Gong, and X. Yi, "Deepfake video detection using audio-visual consistency," X. Zhao, Y. Shi, A. Piva, and H. J. Kim, Eds., vol. 12617. Melbourne, VIC, Australia: Springer, 2020, pp. 168–180.

[14] M. Li, Y. Ahmadiadli, and X. Zhang, "A comparative study on physical and perceptual features for deepfake audio detection," in *DDAM@MM 2022: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, 14 October 2022*, J. Tao, H. Li, H. Meng, D. Yu, M. Akagi, J. Yi, C. Fan, R. Fu, S. Lian, and P. Zhang, Eds. ACM, 2022, pp. 35–41.

[15] J. Xue, C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao, "Audio deepfake detection based on a combination of F0 information and real plus imaginary spectrogram features," in *DDAM@MM 2022: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, 14 October 2022*, J. Tao, H. Li, H. Meng, D. Yu, M. Akagi, J. Yi, C. Fan, R. Fu, S. Lian, and P. Zhang, Eds., 2022, pp. 19–26.

[16] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *CVPR*, Seattle, WA, USA, 2020, pp. 5000–5009.

[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[18] C. C. Lee, "Elimination of redundant operations for a fast sobel operator," vol. 13, no. 2, pp. 242–245, 1983.

[19] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H. Seidel, "Laplacian surface editing," in *Second Eurographics Symposium on Geometry Processing*, vol. 71, Nice, France, 2004, pp. 175–184.
[20] J. J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," vol. 7, no. 3, pp. 868–882, 2012.
[21] Q. Diao, Y. Jiang, B. Wen, J. Sun, and Z. Yuan, "Metaformer: A unified meta framework for fine-grained recognition," in *CVPR*. New Orleans, Louisiana, USA: IEEE, 2022.
[22] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, Seoul, Korea (South), 2019, pp. 1–11.
[23] N. Dufour and A. Gully, "Deepfakedetection dataset," https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, 2019.
[24] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *CVPR*. Seattle, WA, USA: IEEE, 2020, pp. 3204–3213.
[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*. Honolulu, HI, USA: IEEE, 2017, pp. 1800–1807.
[26] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *ECCV*, vol. 12357, Glasgow, UK, 2020, pp. 86–103.
[27] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," vol. 97, Long Beach, California, USA, 2019, pp. 6105–6114.
[28] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in *AAAI*. Virtual Event: AAAI Press, 2022, pp. 744–752.
[29] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *CVPR*. virtual: IEEE, 2021, pp. 16 317–16 326.
[30] Z. Hu, H. Xie, Y. Wang, J. Li, Z. Wang, and Y. Zhang, "Dynamic inconsistency-aware deepfake video detection," in *IJCAI*, Virtual Event / Montreal, Canada, 2021, pp. 736–742.
[31] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in *ACM MM*, H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, and B. Prabhakaran, Eds., Virtual Event, China, 2021, pp. 3473–3481.
[32] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, "Finfer: Frame inference-based deepfake detection for high-visual-quality videos," in *AAAI*, Virtual Event, 2022, pp. 951–959.
[33] P. Pei, X. Zhao, J. Li, Y. Cao, and X. Yi, "Vision transformer based video hashing retrieval for tracing the source of fake videos," *CoRR*, vol. abs/2112.08117, 2021.