On the Use of Implicit Representations for Deepfake Detection

Miguel Leão(✉) and Nuno Gonçalves

Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
miguel.leao@isr.uc.pt

Abstract. The developments in home computers, united with the thousands upon thousands of images/videos of individuals present on the Internet, allowed for the proliferation of deepfaked media affecting the lives of private individuals and the dangerous spread of misinformation. Current state-of-the art detection methods show impressive results. However the development of improved generation methods overcomes them, as there are generalization difficulties. This paper explores the viability of use of implicit representations of facial videos on deepfake detection. Implicit representations offer computer vision tasks a new paradigm of research, possibly offering alternatives to the current methods based on the color space or frequency domain. This work investigates the use of Sinusoidal Representation Networks (SIRENs) to show a significant difference between Fréchet Video Distance (FVD) scores obtained from bonafide videos and their SIREN reconstruction and deepfake videos and their SIREN reconstruction. This result leads to the conclusion that the SIREN representation of a video can be used as input for a deepfake detection method, opening a new avenue of research.

Keywords: Deepfakes · implicit representation · significance test · Fréchet Video Distance

1 Introduction

Deepfakes are introduced in 2017, showing image manipulations resulting from deep learning algorithms. While rudimentary at first, deepfakes are evolving ever closer to near undetectable manipulations in both images and videos.

Although they may be used for entertainment or other nonharmful purposes, the concern is with the material made targeting private individuals or public figures. The advances of video deepfakes in conjunction with audio deepfakes may tarnish the good reputation of someone or be used to spread misinformation. As such, it is vital to keep up with the development of deepfake generations, innovating in the detection front as much as possible.

Current deepfake analysis is confronted with challenges related to the variety of not only the different manipulations that are considered as deepfakes but also the models used to create said manipulations. These manipulations are identity swaps where a source identity receives the target's facial information, expression swap where the target's expression is manipulated according to the sources

N. Gonçalves et al. (Eds.): IbPRIA 2025, LNCS 15937, pp. 227–238, 2026.
https://doi.org/10.1007/978-3-031-99565-1_18

information, attribute manipulation where the target's attributes are modified and full face synthesis where a new identity is generated. These manipulations give different cues to be explored for deepfake detection and the different methods used for their creation give another set of cues to be explored, resulting in detection methods having difficulty with generalization.

Early detection approaches relied on the flaws of the generative methods. These could be based on the search for visual-cues where poor quality deepfakes had visible errors or temporal inconsistencies between frames of a video allowed for easier detection. With the improvement of the generative methods, these visual cues become subtler or disappear altogether. The forensic approach continues on the frequency domain searching for artifacts left by the model in the frames, but these approaches have difficulties as soon as the video is compressed, losing much of the information that allowed for accurate detection. In anticipation of a future where these 'clues' are no longer present, the search must be carried out elsewhere.

Following the logic of forensic approaches in both the color and frequency space, this work investigates the use of the implicit space in the problem of deepfake detection. Implicit representations have recently offered new research avenues for image analysis, translating a scene usually in a coordinates-based representation, that allows for detailed reconstructions of the original. Using Sinusoidal Representation Networks (SIRENs) [27] the video frames of the Deepfake Detection Challenge Dataset (DFDC) [3] were translated to the implicit space and analyzed.

This work uses Fréchet Video Distance (FVD) [29] between the original DFDC videos and their respective SIREN reconstruction, to show a significant difference in the average FVDs of the bonafide and deepfake pairs. We expect that this work might open new avenues of research for the deepfake detection problem.

2 Literature Review

2.1 Deepfakes

Deepfake generation begins with an auto-encoder approach, where images from persons A and B are used for the training of their respective auto-encoders, sharing the training weights between encoders while maintaining the decoders completely separate, as shown in [1]. When optimization is finished, the images of person A can be encoded with the shared encoder but then decoded with the person B decoder to produce the final image.

Since this initial approach, deepfake generation methods have evolved to use Generative Adversarial Networks (GANs) [7], adopting different approaches depending on the desired manipulation. These manipulations are identity swapping, expression swapping, attribute manipulation, and full identity synthesis.

Recently, innovations in identity swapping have tried to solve problems with undesired attribute swapping between identities [26] or non-identity attributes being removed in the final result [24].

The synthesis of new identities is a topic of interest not only in relation to deepfakes, but to facial biometrics as a whole. The generation of completely synthetic datasets makes it possible to solve the challenges of collecting data, the privacy of the participants, or the lack of balance between the demographics present [16]. Topics such as Presentation Attack Detection (PAD), have an interest in the generation of deepfakes since these may be use to spoof a biometrics system [20].

With much of facial synthesis being based on StyleGANs [12] and it's predecessors, recent improvements have been made by exploiting the latent space from StyleGANs2 to compensate for a lack of robustness to the source frame's facial expressions and head pose [19] or by moving away from GANs to the more recent diffusion models [37].

With such developments on the generation of deepfakes, there is an equivalent effort in deepfake detection research. Older methods based on spatio-temporal approaches are improved upon by searching for more minute inconsistencies such as the disturbance created by face movements [34]. Other approaches look for inconsistencies in biological signals with examples being inconsistencies with the gaze [21] or using remote photoplethysmography (rPPG) [33] as an alternative to the visual signals.

The use of information found in the frequency domain, allows researchers to search for non-visual cues, usually artifacts left from the generative method to then base the detection method on this information alone [28] or in a multi-modal approach, fusing frequency level information with some other [30].

However, the fear of unseen deepfakes leads research into looking for greater generalization capabilities, or focusing on the explainability of models, to better understand how and why decisions are made [13]. Mining for specific clues related to the artifacts left by the manipulation approach has given good results for deepfake detection but tends to be model specific. By giving different backbones to the model [15] or by adapting the texture and spectrum analysis to the input image [14], the models generalize better. Other researchers claim that a hurdle for generalization is implicit identity leakage [4] and aim to reduce the impact of the identity factor on deepfake detection, while others [10] go the opposite direction, by comparing the explicit and implicit identities of face images.

Another approach to deepfake detection is a proactive one. Assuming that a perfect detection model exists that would be able to detect any deepfake created by any method, the damage that said deepfake could do before being labeled as such would not be immediately countered. As such, researchers propose methods to prevent the creation of deepfakes through adversarial attacks. The standard approach is to insert noise into the image making it difficult for manipulation to occur, but this also makes any analysis of the "protected" image difficult as well. To combat this [38] uses an information-containing adversarial perturbation, that associates the image to a database and encodes into the distorted image, a message that links back to the unadulterated image present in the database.

The concern is with images uploaded to the public through social media for example. These social media platforms usually compress media that is uploaded,

causing the adversarial information to lose effect. While there are efforts to reduce the effect this compression might have [23,32] uses a more direct approach in modeling the adversarial perturbations according the compression found on social media platforms, as to reduce the negative impact the compression might have on the adversarial attack.

2.2 Implicit Representations

Implicit Neural Representations have recently offered a new approach to study visual data by parameterizing a segment of media, such as an image or video segment through a neural network. Neural Radiance Fields (NeRFs) [17] which outputs a scene through a 5D function of spatial coordinates and viewing directions, was improved upon to create dynamic NeRFs [5] for the creation of facial avatars.

Concurrent with NeRFs, SIRENs [27] are introduced as a continuous implicit representation, that not only boasts better representation, but is able to apply to its derivatives i.e. the derivative of a SIREN is a SIREN itself, which allows for further applications. The authors use SIRENs to solve a number of problems such as the Poisson Equation or the Helmholtz and Wave Equation, but most importantly for this paper, they point out the use of SIRENs for image fitting, achieving good results in reconstructing not only the neural image but also the first and second derivatives, and extend the applications to both video and audio.

While there are a number of works related to improving talking head models through NeRFs with techniques such as audio guided synthesis [8] or motion-assisted synthesis [36], as far as this review was able to find, there are no works that leverage implicit representations for deepfake detection.

3 Method

3.1 Implicit Representation

An image is represented as a function $I : \Omega \subset R^2 \rightarrow C$, where Ω is the image's domain and C is the color space. The image is then parameterized with a coordinate-based neural network $I_\theta : R^2 \rightarrow C$ with parameters θ. To train the neural image I_θ so that it approximates I, the model optimizes the following objective:

$$\int_\Omega (I - I_\theta)^2 \, dx.$$

The coordinate-based network is a sinusoidal multilayer perceptron (MLP) $f_\theta(p) : R^n \rightarrow R^m$, defined as a composition of d sinusoidal layers:

$$f_\theta(x) = W_d \circ f_{d-1} \circ \cdots \circ f_0(x) + b_d,$$

where each layer $f_i(x_i) = \sin(W_i x_i + b_i) = x_{i+1}$, with $W_i \in R^{n_{i+1} \times n_i}$ being the weight matrices, and $b_i \in R^{n_{i+1}}$ being the biases. The collection of these

parameters defines θ. The integer d denotes the depth of the network, and n_i refers to the width of the layers.

With the neural image defined by θ, the RGB values for any pixel of a reconstructed image are given by the value of f_θ at x coordinates.

Through the method used in [25], the neural images of each frame of the subject's face is obtained. The individual frames are then joined into a reconstructed video.

3.2 Distance Between Original and Reconstructed Videos

This article proposes to show that there is a difference between how reliable the neural reconstruction of a video is for bonafide and deepfake video cases, so that it can be used to detect the latter. This is measured through Fréchet Video Distance (FVD).

FVD is proposed as an improvement on common video analysis approaches such as Peak Signal-to-Noise-Ratio (PSNR) or Structural Similarity (SSIM) [31] claiming that these lack for the temporal coherence of the video, aside from the video quality itself. It is based on the principal of Fréchet Inception Distance (FID) [9], commonly used for image analysis, where the distance between the real world data distribution P_R and the distribution defined by the generative model P_G is defined by:

$$d(P_R, P_G) = min_{X,Y} E|X - Y|^2$$

where the minimization is over all random variables X and Y with distributions P_R and P_G respectively. With the data distribution being represented as a multivariate Gaussian using a suitable feature space, the previous equation can be solved as:

$$d(P_R, P_G) = |\mu_R - \mu_G|^2 + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R \Sigma_G)^{\frac{1}{2}})$$

where μ_R and μ_G are the means and Σ_R and Σ_G are the co-variance matrices of P_R and P_G. This representation is obtained from an Inflated 3D ConvNet (I3D) [2], and the distance between videos is obtained. In our work, we obtained the FVD through the implementation used in [6].

3.3 Computational Effort

While implicit representations offer a lightweight alternative to vision problems, obtaining the representation itself comes with a computational cost. This paper was conducted using two NVIDIA GeForce RTX 2080 Ti, obtaining the SIREN representation for one 256×256 facial image using one of these GPUs takes approximately 10 s.

With the current setup, a single ten second video, at 30 FPS, takes roughly 50 min to process. At the scale of over 120,000 videos, with the 2 GPUs used, it would take over 5 years to just obtain the data to train a potential model.

While reducing the number of frames that are represented is a possible solution, it carries the risk of information loss. Section 5 addresses this topic in order to mitigate the heavy computational power needed for the training.

4 Experiments and Results

4.1 Dataset

The Deepfake Detection Challenge (DFDC) [3] is a self-designated third generation dataset featuring 23,654 videos from 960 actors hired for this purpose, from which 104,500 fake videos are created using various deepfake creation methods.

These include a Deepfake Auto Encoder (DFAE) model with a shared encoder but two isolated decoders, one for each identity, and a Neural Talking Heads (NTH) [35] model comprised of a metalearning stage and a fine-tuning stage.

It also includes deepfakes generated from FSGAN [18] which applies an adversarial loss to generators for reenactment and inpainting, and trains additional generators for face segmentation and Poisson blending and StyleGAN [11] which is modified to produce a face swap between a given fixed identity descriptor onto a video by projecting this descriptor on the latent face space. Finally, certain videos from the previous categories are processed with a sharpening filter to improve the quality of the final video and certain videos receive vocal deepfakes with the method presented in [22].

As previously mentioned, processing the whole dataset would require a large computational effort. As such the results were obtained from a total of 2048 videos, randomly selected from the dataset, split evenly between bonafides and deepfakes with a 256×256 window over the facial region. While ideally, the selection of deepfake videos would be made as to get an even distribution of videos generated by the methods mentioned previously, this is currently not possible as the dataset does not provide that information. The random selection expects a somewhat even distribution but cannot guarantee it (Fig. 1).

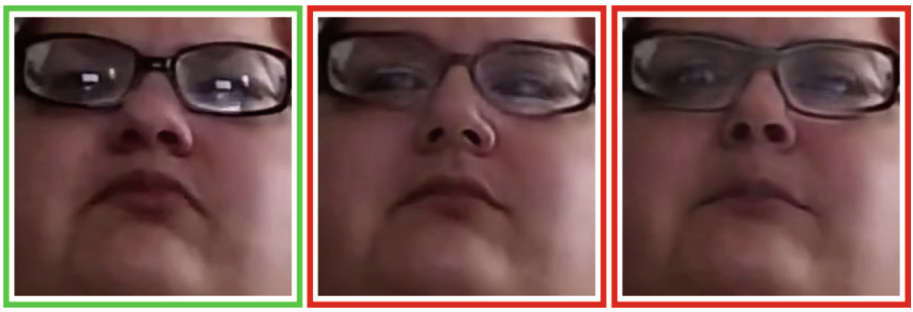

Fig. 1. Example of the 256×256 windows over a bonafide video frame (in green) and two deepfakes generated from it (in red) from the DFDC dataset. (Color figure online)

4.2 SIREN Reconstructions

The SIREN models were trained for 1000 epochs for each frame, resulting in a reconstruction that shows no differences to the naked eye, for both deepfake and bonafide videos, even for the ones scoring the highest FVD scores, as shown in Figs. 2.

Fig. 2. Comparison between an original frame (left) from a video, it's SIREN reconstruction (center) and their difference (right), for bonafide cases in green and deepfake cases in red. (Color figure online)

Although the reconstructions do not show visible differences when analyzed, it is possible to find the areas in the image where the reconstructions is not perfect. Analyzing these areas together with additional information from the scene can give insights into the problem.

This would greatly benefit from a labeling effort on the dataset to properly analyze if and how different conditions affect the SIREN representation for bonafide and deepfake cases.

4.3 Testing the Hypothesis

The average FVD scores obtained show that SIREN reconstructions for the bonafide videos have lower fidelity to their original video than in the deepfake videos, achieving higher FVD scores, as shown in Fig. 3. Higher FVD scores mean higher distance between video pairs i.e. worse reconstructions.

To confirm that the data shows that SIREN reconstructions could be used for deepfake detection, specifically that the neural reconstructions of bonafide material have lower fidelity than deepfake reconstructions, a one tail significance test is conducted. First the null hypothesis is defined as there is no difference

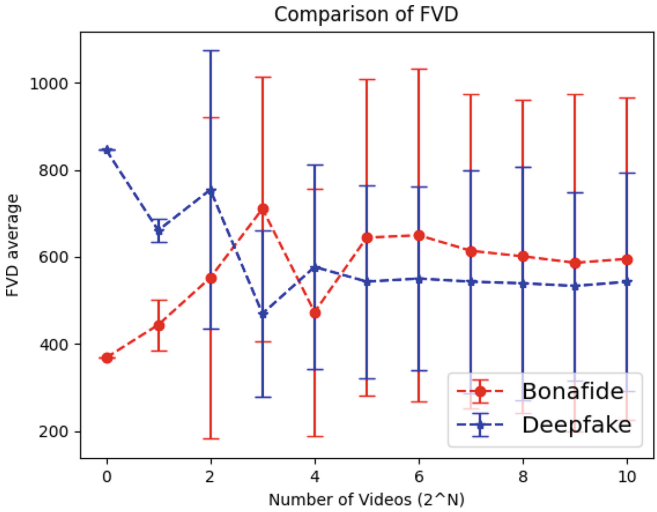

Fig. 3. Average FVD scores between bonafide videos and their SIREN reconstruction and deepfake videos and their reconstructions. The averages are given for 2^N video pairs of each category with the final values being $\mu = 596.95$ and $\sigma = 372.23$ for bonafide pairs and $\mu = 538.09$ and $\sigma = 245.29$ for deepfake pairs.

between the distribution of FVD scores for the original and SIREN reconstruction of bonafide videos and deepfake videos, or that the FVD scores for bonafide videos are lower then the deepfake scores, i.e. SIRENS achieve better reconstructions on bonafide videos than deepfake ones:

$$H_0 : \mu FVD_{bonafide} <= \mu FVD_{deepfake}$$

and present the alternative hypothesis that the bonafide video pairs score higher FVDs, therefore have worse fidelity, than the deepfake video pairs:

$$H_a : \mu FVD_{bonafide} > \mu FVD_{deepfake}$$

conducting the test with a significance level of $\alpha = 0.01$, the p-value result is equal to $1.145e-5$ giving $p < \alpha$, thus rejecting the null hypothesis.

4.4 Discussion

The data shows that SIREN reconstructions bonafide videos have lower fidelity than the reconstructions of deepfake videos. This could suggest that bonafide videos contain richer information, which is lost during manipulation.

The fact that the standard deviation for deepfake scores is lower, may also indicates a process of homogenization of the information. The DFDC is a large dataset, so as previously mentioned, its full translation into neural representations would, at this pace, take a not practical amount of time. However by

expanding this research over more videos from the dataset, it would give a clearer idea of how different attributes from these videos might affect the neural representation, or how certain deepfake generative methods impact the image.

This result could contribute to the development of a system for detecting deepfakes by learning how to distinguish between bonafide videos and deepfake videos by analyzing the original video and its neural reconstruction.

However, there is still work to be done in this area, particularly in understanding the influence of certain factors like resolution, the context of the video (e.g. how much of the frame does the face occupy), among other elements.

5 Conclusion and Future Work

5.1 Conclusion

This article presented the hypothesis of using implicit representations of facial videos to distinguish between bonafide and deepfake videos. Carrying out this first analysis with videos reconstructed from the SIREN representation, the FVD value between the original videos and their reconstructions was measured. These values were use to test the hypothesis that the bonafide reconstructions have lower fidelity to their original material when compared to the reconstruction of deepfake videos. Carrying out a significance test at a significance level of 99%, we were able to show that the null hypothesis was rejected. Although these are initial results, the hypothesis that we can use implicit representations to detect deepfakes seems promising.

5.2 Future Works

Having reached these conclusions, it is necessary to consider how to proceed. The end result of this research is expected to achieve state-of-the-art deepfake detection. There are still a number of obstacles to overcome, with problems such as data volume. It is still required to test if all frames from a video are required to achieve satisfactory results. This, among a battery of ablation tests, will be conducted as to conceive the "ideal" conditions to proceed with research.

While this paper revolves around videos reconstructed from their SIREN representations, it is to show a discernible distinction between deepfakes and bonafide material. Future work will be conducted as much as possible with the implicit representation itself.

Acknowledgments. This study has received funding from the Portuguese Fundação para a Ciência e Tecnologia, I.P. under the project BLOCKDFAKE - 2024.07681.IACDC. DOI:10.54499/2024.07681.IACDC.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7 (2018). https://doi.org/10.1109/WIFS.2018.8630761
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733 (2017). https://doi.org/10.1109/CVPR.2017.502
3. Dolhansky, B., et al.: The deepfake detection challenge dataset. arXiv abs/2006.07397 (2020)
4. Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., Ge, Z.: Implicit identity leakage: the stumbling block to improving deepfake detection generalization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3994–4004 (2023). https://doi.org/10.1109/CVPR52729.2023.00389
5. Gafni, G., Thies, J., Zollhofer, M., Niesner, M.: Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8645–8654. IEEE Computer Society, Los Alamitos, CA, USA (2021). https://doi.org/10.1109/CVPR46437.2021.00854
6. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in Fréchet video distance. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7277–7288. IEEE Computer Society, Los Alamitos, CA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.00695
7. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc. (2014)
8. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: audio driven neural radiance fields for talking head synthesis. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5764–5774 (2021). https://doi.org/10.1109/ICCV48922.2021.00573
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
10. Huang, B., et al.: Implicit identity driven deepfake face swapping detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4490–4499 (2023). https://doi.org/10.1109/CVPR52729.2023.00436
11. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4396–4405 (2019). https://doi.org/10.1109/CVPR.2019.00453
12. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN . In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8107–8116. IEEE Computer Society, Los Alamitos, CA, USA (2020). https://doi.org/10.1109/CVPR42600.2020.00813
13. Khalid, F., Javed, A., Malik, K.M., Irtaza, A.: Explanet: a descriptive framework for detecting deepfakes with interpretable prototypes. IEEE Trans. Biometrics Behav. Identity Sci. **6**(4), 486–497 (2024). https://doi.org/10.1109/TBIOM.2024.3407650

14. Liu, J., Xie, J., Wang, Y., Zha, Z.J.: Adaptive texture and spectrum clue mining for generalizable face forgery detection. IEEE Trans. Inf. Forensics Secur. **19**, 1922–1934 (2024). https://doi.org/10.1109/TIFS.2023.3344293
15. Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., Kot, A.C.: Beyond the prior forgery knowledge: mining critical clues for general face forgery detection. IEEE Trans. Inf. Forensics Secur. **19**, 1168–1182 (2024). https://doi.org/10.1109/TIFS.2023.3332218
16. Melzi, P., et al.: Gandiffface: controllable generation of synthetic datasets for face recognition with realistic variations. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3078–3087 (2023). https://doi.org/10.1109/ICCVW60793.2023.00333
17. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021). https://doi.org/10.1145/3503250
18. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: subject agnostic face swapping and reenactment. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7183–7192 (2019). https://doi.org/10.1109/ICCV.2019.00728
19. Oorloff, T., Yacoob, Y.: Robust one-shot face video re-enactment using hybrid latent spaces of stylegan2. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 20890–20900 (2023). https://doi.org/10.1109/ICCV51070.2023.01915
20. Otroshi Shahreza, H., Marcel, S.: Template inversion attack using synthetic face images against real face recognition systems. IEEE Trans. Biometrics Behav. Identity Sci. **6**(3), 374–384 (2024). https://doi.org/10.1109/TBIOM.2024.3391759
21. Peng, C., Miao, Z., Liu, D., Wang, N., Hu, R., Gao, X.: Where deepfakes gaze at? Spatial-temporal gaze inconsistency analysis for video face forgery detection. IEEE Trans. Inf. Forensics Secur. **19**, 4507–4517 (2024). https://doi.org/10.1109/TIFS.2024.3381823
22. Polyak, A., Wolf, L., Taigman, Y.: TTS skins: speaker conversion via ASR. In: Interspeech (2019)
23. Qu, Z., Xi, Z., Lu, W., Luo, X., Wang, Q., Li, B.: DF-rap: a robust adversarial perturbation for defending against deepfakes in real-world social network scenarios. IEEE Trans. Inf. Forensics Secur. **19**, 3943–3957 (2024). https://doi.org/10.1109/TIFS.2024.3372803
24. Ren, X., Chen, X., Yao, P., Shum, H.Y., Wang, B.: Reinforced disentanglement for face swapping without skip connection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 20608–20618 (2023). https://doi.org/10.1109/ICCV51070.2023.01889
25. Schardong, G., et al.: Neural implicit morphing of face images. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7321–7330. IEEE Computer Society, Los Alamitos, CA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.00699
26. Shiohara, K., Yang, X., Taketomi, T.: Blendface: re-designing identity encoders for face-swapping. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7600–7610 (2023). https://doi.org/10.1109/ICCV51070.2023.00702
27. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 7462–7473. Curran Associates, Inc. (2020)

28. Tan, C., Zhao, Y., Wei, S., Gu, G., Wei, Y.: Learning on gradients: generalized artifacts representation for GAN-generated images detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12105–12114 (2023). https://doi.org/10.1109/CVPR52729.2023.01165
29. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: a new metric for video generation. In: DGS@ICLR (2019)
30. Wang, Y., Yu, K., Chen, C., Hu, X., Peng, S.: Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7278–7287 (2023). https://doi.org/10.1109/CVPR52729.2023.00703
31. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
32. Wu, H., Zhou, J., Tian, J., Liu, J., Qiao, Y.: Robust image forgery detection against transmission over online social networks. IEEE Trans. Inf. Forensics Secur. **17**, 443–456 (2022). https://doi.org/10.1109/TIFS.2022.3144878
33. Wu, J., Zhu, Y., Jiang, X., Liu, Y., Lin, J.: Local attention and long-distance interaction of RPPG for deepfake detection. Vis. Comput. **40**(2), 1083–1094 (2024). https://doi.org/10.1007/s00371-023-02833-x
34. Yin, Q., Lu, W., Li, B., Huang, J.: Dynamic difference learning with spatio-temporal correlation for deepfake video detection. IEEE Trans. Inf. Forensics Secur. **18**, 4046–4058 (2023). https://doi.org/10.1109/TIFS.2023.3290752
35. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9458–9467 (2019). https://doi.org/10.1109/ICCV.2019.00955
36. Zhang, W., Zhou, X., Cao, Y., Feng, W., Yuan, C.: Ma-nerf: motion-assisted neural radiance fields for face synthesis from sparse images. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 1757–1762 (2023)
37. Zhang, Y., et al.: Genface: a large-scale fine-grained face forgery benchmark and cross appearance-edge learning. IEEE Trans. Inf. Forensics Secur. **19**, 8559–8572 (2024). https://doi.org/10.1109/TIFS.2024.3461958
38. Zhu, Y., et al.: Information-containing adversarial perturbation for combating facial manipulation systems. IEEE Trans. Inf. Forensics Secur. **18**, 2046–2059 (2023). https://doi.org/10.1109/TIFS.2023.3262156