Unmasking Sophisticated Video Deepfakes: A Spatio-Temporal Approach for Multi-Identity and Size-Varying Scenarios

Kongara Harsha Deep[1], Mogili Karunakar[2] , Mullapudi Bhargava Satya Narendra[3], Dr. K. Siva Kumar[4]

[1,2,3,4] Department of Computer Science and Engineering, R.V.R & J.C College of Engineering, Guntur, India

Abstract:

Deepfake video generation techniques are evolving rapidly, creating highly realistic manipulated content that poses significant societal risks. Detecting these fakes, especially in complex, real-world videos, remains a major challenge. Current detection methods often struggle when videos feature multiple individuals or when faces appear at widely varying sizes. Many approaches rely on simple frame averaging or focus only on the most prominent face, potentially missing subtle or localized manipulations and ignoring crucial temporal inconsistencies. To overcome these limitations, we present a novel video deepfake detection framework designed for robustness in challenging scenarios. Our approach uniquely integrates a Convolutional Neural Network (CNN) backbone, capturing fine-grained spatial details, with a Spatio-Temporal Transformer architecture adept at modeling temporal dynamics. Critically, we introduce an Identity-Aware Attention mechanism. This allows the model to process face sequences corresponding to different individuals independently within the Transformer, enabling effective analysis of multi-person videos without resorting to naive post-hoc aggregation. Furthermore, we incorporate two specialized embedding strategies: Temporal Coherence Embeddings that preserve the correct temporal ordering and relationships of faces, even across different identities appearing concurrently, and Relative Size Embeddings that explicitly encode the scale of each detected face relative to the video frame. Our experiments, particularly on the diverse ForgeryNet dataset, demonstrate state-of-the-art performance, showing a marked improvement (up to 14% AUC) in videos containing multiple people compared to existing methods. The framework also shows strong generalization capabilities across different forgery types and datasets, highlighting its potential for practical deployment. [Optional: We plan to release our implementation to facilitate further research.

Introduction:

(I.) The Evolving Deepfake Threat

The past few years have witnessed an explosion in the quality and accessibility of deepfake technology. Fueled by advances in generative models like GANs, NeRFs, and Diffusion Models [CITE 1, 2, 3, 4], creating hyper-realistic manipulated videos of individuals is becoming increasingly feasible. While these tools have creative applications, their malicious use—for disinformation campaigns, personal harassment, fraud, and undermining democratic processes [CITE 5]—presents a serious and growing threat. Distinguishing authentic footage from sophisticated fakes is now a critical task for maintaining digital trust.

(II.) Limitations of Current Video Deepfake Detection

In response, numerous deepfake detection methods have emerged. However, many existing approaches exhibit significant limitations when confronted with the complexities of real-world videos:

1. **Over-reliance on Spatial Artifacts:** Many detectors analyze videos on a frame-by-frame basis, focusing primarily on spatial inconsistencies introduced by the generation process [CITE 9, 10, 11, 12]. While useful, this often neglects

the rich temporal information inherent in videos. Detecting subtle flickering, unnatural motion, or inconsistent dynamics across frames is crucial but often overlooked [CITE 6, 7, 8].

2. **The Aggregation Problem:** Frame-based methods typically require an aggregation step (e.g., averaging or taking the maximum score) to arrive at a single video-level prediction. The choice of aggregation strategy can heavily influence the final outcome, making these methods brittle and potentially unreliable [CITE 8, 13]. An ideal system should perform *internal* aggregation, learning to weigh evidence across the entire video context.

3. **The Multi-Identity Challenge:** Real-world videos frequently contain multiple people. An attacker might only manipulate one individual, hoping the presence of authentic faces will mask the forgery [CITE 14]. Methods analyzing the video *en bloc* or averaging scores across all faces risk being deceived. Processing each identity track separately can be computationally expensive and still leads back to the aggregation problem. There's a clear need for methods that can inherently handle and reason about multiple identities simultaneously within a single analysis pass.

4. **Ignoring Face Scale Information:** Detection pipelines typically detect faces and resize them to a standard input size for a classifier. This normalization discards potentially valuable information about the face's original size relative to the frame. Scale variations might correlate with manipulation quality or provide context clues that could improve robustness.

5. **Poor Generalization:** Many detectors learn artifacts specific to the forgery methods present in their training data. Consequently, they often fail to generalize to unseen manipulation techniques or different datasets, limiting their practical utility [CITE 15-22].

(III.) Our Proposed Approach: Context-Aware Temporal Video Forensics

Motivated by these shortcomings, we propose a new deepfake detection framework specifically designed to address the challenges of multi-identity, size-varying, and temporally complex videos. Our goal is to move beyond simple frame analysis towards a holistic spatio-temporal understanding. Our core contributions are:

- **Integrated Spatio-Temporal Analysis:** We combine a CNN feature extractor with a Spatio-Temporal Transformer (inspired by architectures like TimeSformer [CITE 48], adapted for this task). This allows capturing both detailed spatial forgery traces and their evolution over time.

- **Identity-Aware Attention:** We introduce a novel attention mechanism within the Transformer that processes sequences of face tokens *conditioned on identity*. It effectively allows the model to track and analyze inconsistencies for each person separately while still integrating information for a unified video-level prediction.

- **Relative Size Embedding:** We propose a new embedding technique that explicitly informs the model about the relative size of each detected face within its original frame, allowing it to potentially learn size-dependent forgery patterns or improve robustness to scale changes.

- **Temporal Coherence Embedding:** We employ a positional embedding scheme carefully designed to encode the temporal sequence of faces, correctly handling multiple faces appearing in the same frame and maintaining consistent temporal relationships across different identities.

- **End-to-End Video-Level Prediction:** Thanks to the identity-aware attention and internal aggregation within the Transformer (using mechanisms like a CLS token), our model directly outputs a single, robust prediction for the entire video in one forward pass, eliminating the need for unreliable post-hoc aggregation schemes.

We rigorously evaluated our approach on the challenging ForgeryNet benchmark [CITE 49] and conducted extensive cross-forgery and cross-dataset experiments. The results show significant performance gains over existing methods, particularly in multi-identity scenarios, and demonstrate superior generalization capabilities. This suggests our framework offers a promising direction for building more reliable and practical video deepfake detection systems.

(IV.) Experimental Validation and Key Results

We conducted extensive evaluations to validate the effectiveness of our proposed framework, primarily using the diverse ForgeryNet dataset [CITE 49], known for its variety in forgery types, multi-identity scenarios, and face-frame area ratios. We compared our approach (instantiated with both EfficientNet-B0 and XceptionNet backbones, termed MINTIME-EF and MINTIME-XC respectively) against relevant state-of-the-art methods, including strong spatio-temporal models like SlowFast [CITE 60] and frame-based hybrid approaches like Cross Convolutional ViT [CITE 12].

Our key findings are summarized as follows:

1. **State-of-the-Art Performance:** On the ForgeryNet validation set, our MINTIME-XC model achieved state-of-the-art performance, outperforming existing methods in AUC and achieving accuracy comparable to the best spatio-temporal approaches (which typically only handle single identities) [Refer to MINTIME Table II]. MINTIME-EF also showed strong competitive results.

2. **Superior Multi-Identity Handling:** A crucial finding was the framework's exceptional performance on videos containing multiple identities. When evaluated exclusively on these challenging subsets, MINTIME-XC demonstrated a significant improvement, achieving up to 14% higher AUC compared to methods not explicitly designed for this scenario [Refer to MINTIME Table III]. This underscores the effectiveness of the Identity-Aware Attention mechanism. Frame-based methods like Cross Convolutional ViT struggled significantly on these multi-identity cases.

3. **Robustness Across Forgery Types:** Our method showed consistently strong detection rates (True Positive Rate) across the various forgery techniques present in ForgeryNet, while maintaining a high True Negative Rate on pristine videos. Unlike some methods that excel on certain artifact types but falter on others, MINTIME-XC displayed lower variance in performance across manipulations, indicating better robustness [Refer to MINTIME Table IV].

4. **Generalization Capabilities:**

○ **Cross-Forgery:** We tested generalization to unseen manipulation types by training on one category of forgeries (e.g., ID-Replaced) and testing on another (e.g., ID-Remained), and vice-versa. MINTIME-XC consistently outperformed prior work in these challenging settings, demonstrating a better ability to learn generalizable forgery indicators rather than overfitting to specific artifacts [Refer to MINTIME Table VI]. Training on diverse forgeries proved beneficial.

○ **Cross-Dataset:** When trained on ForgeryNet and evaluated on the DFDC Preview test set [CITE 70], MINTIME-XC achieved highly competitive generalization performance compared to methods trained on datasets more similar to DFDC (like FaceForensics++), showcasing its ability to adapt across different data distributions [Refer to MINTIME Table VII].

5. **Impact of Novel Components:** Ablation studies systematically validated the contributions of our novel components. Disabling the Size Embeddings led to a noticeable drop in performance, especially on videos with smaller face-frame ratios [Refer to MINTIME Tables VIII, IX]. Similarly, removing the Identity-Aware Attention and Temporal Coherence Embeddings negatively impacted performance on multi-identity videos [Refer to MINTIME Table X]. The identity sorting policy based on size during inference also proved beneficial compared to random or frequency-based sorting [Refer to MINTIME Table XI].

6. **Qualitative Insights:** Analysis of the attention maps within the Transformer revealed that the model learns to focus on the manipulated faces or identities within a video, providing a degree of interpretability and potentially allowing users to identify *which* parts of a video are deemed suspicious [Refer to MINTIME Figure 7].

(V.) Conclusion and Future Directions

In this work, we tackled the challenging problem of detecting sophisticated video deepfakes in complex, real-world scenarios often characterized by multiple individuals and varying face scales. Current methods frequently struggle with these complexities, relying on simplistic aggregation or ignoring vital temporal and scale information.

Our framework, integrating CNN features with a tailored Spatio-Temporal Transformer incorporating **Identity-Aware Attention**, **Relative Size Embeddings**, and **Temporal Coherence Embeddings**, directly addresses these limitations. It effectively processes spatio-temporal inconsistencies, distinctly handles multiple identities within a single forward pass, incorporates valuable face scale information, and eliminates the need for problematic post-hoc prediction aggregation.

The empirical results demonstrate state-of-the-art performance on the challenging ForgeryNet dataset, with particularly significant gains in multi-identity videos. Furthermore, our approach exhibits robust generalization across different forgery types and datasets, a crucial requirement for practical deployment. The attention mechanism also offers potential for explaining the model's predictions by highlighting suspicious faces or identities.

Looking ahead, several avenues warrant exploration. Integrating multi-modal information, particularly audio cues which often contain deepfake artifacts, could further enhance detection robustness. Adapting the framework to handle even larger numbers of identities efficiently and testing on newer, more challenging benchmark datasets will be important steps. Exploring the application of these principles to detect manipulations beyond face-swapping, such as full body synthesis or scene manipulation, represents another promising direction for future research.

References

[1] I. Goodfellow et al., "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst., vol. 27. Red Hook, NY, USA: Curran Associates, 2014, pp. 1–9.

[2] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the wild: Neural radiance fields for unconstrained photo collections," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 7210–7219.

[3] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron, "NeRFactor: Neural factorization of shape and reflectance under an unknown illumination," ACM Trans. Graph., vol. 40, no. 6, pp. 1–18, Dec. 2021.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10674–10685.

[5] M. Van Huijstee, P. Van Boheemen, and D. Das, Tackling Deepfakes in European Policy. Scientific Foresight Unit (STOA), European Parliamentary Research Service (EPRS), 2021.

[6] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15024–15034.

[7] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 14930–14942.

[8] S. Baxevanakis et al., "The mever deepfake detection service: Lessons learnt from developing and deploying in the wild," in Proc. 1st Int. Workshop Multimedia AI Against Disinformation, 2022, pp. 1–10.

[9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Dec. 2018, pp. 1–7.

[10] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in Proc.

4th ACM Workshop Inf. Hiding Multimedia Secur., Jun. 2016, pp. 5–10.
[11] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid transformer network for deepfake detection," in Proc. Int. Conf. Content-Based Multimedia Indexing, Sep. 2022, pp. 8–14.
[12] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in Image Analysis and Processing (ICIAP)—Part III. Lecce, Italy: Springer, 2022, pp. 219–229.
[13] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Investigating the impact of pre-processing and prediction aggregation on the deepfake detection task," in Proc. Truth Trust Online Conf. (TTO), Oct. 2020, pp. 1–11.
[14] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 10097–10107.
[15] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 5780–5789.
[16] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues," in Proc. Eur. Conf. Comput. Vis. (ECCV). Springer, 2020, pp. 86–103.
[17] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in Proc. ECCV. Springer, 2020, pp. 103–120.
[18] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don’t lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 5037–5047.
[19] L. Li et al., "Face X-ray for more general face forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 5000–5009.
[20] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-generated images are surprisingly easy to spot...for now," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8692–8701.
[21] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, and G. Amato, "Cross-forgery analysis of vision transformers and CNNs for deepfake image detection," in Proc. 1st Int. Workshop Multimedia AI Against Disinformation, Jun. 2022, pp. 52–58.
[22] D. A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, "On the generalization of deep learning models in video deepfake detection," J. Imag., vol. 9, no. 5, p. 89, Apr. 2023.
[23] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 2382–2390.
[24] M. Westerlund, "The emergence of deepfake technology: A review," Technol. Innov. Manag. Rev., vol. 9, no. 11, pp. 39–52, 2019.
[25] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial

images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1–11.
[26] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in Proc. Int. Symp. Comput., Consum. Control (IS3C), Dec. 2018, pp. 388–391.
[27] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, "MTD-net: Learning to detect deepfakes images by multi-scale texture difference," IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4234–4245, 2021.
[28] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14918–14927.
[29] Z. Hanqing, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 2185–2194.
[30] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F2Trans: High-frequency fine-grained transformer for face forgery detection," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 1039–1051, 2023.
[31] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114.
[32] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 347–356.
[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1800–1807.
[34] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16312–16321.
[35] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Dec. 2018, pp. 1–7.
[36] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1205–1207.
[37] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don’t lie: An audio-visual deepfake detection method using affective cues," in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 2823–2832.
[38] I. Amerini and R. Caldelli, "Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos," in Proc. ACM Workshop Inf. Hiding Multimedia Secur., New York, NY, USA, 2020, pp. 97–102.
[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[40] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, “ID-reveal: Identity-aware DeepFake video detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15088–15097.
[41] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deep-fake videos from appearance and behavior," in Proc. IEEE Int. Workshop Inf.

Forensics Secur. (WIFS), Dec. 2020, pp. 1–6.
[42] X. Dong et al., "Protecting celebrities from DeepFake with identity consistency transformer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 9458–9468.
[43] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, arXiv:2010.11929.
[44] A. Vaswani et al., “Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.
[45] N. Shazeer, "GLU variants improve transformer," 2020, arXiv:2002.05202.
[46] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv:1607.06450.
[47] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[48] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in Proc. ICML, vol. 2, no. 3, 2021, p. 4.
[49] Y. He et al., "ForgeryNet: A versatile benchmark for comprehensive forgery analysis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4358–4367.
[50] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," 2020, arXiv:1912.13457.
[51] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7183–7192.
[52] I. Perov et al., "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," 2021, arXiv:2005.05535.
[53] O. Fried et al., "Text-based editing of talking-head video," ACM Trans. Graph., vol. 38, no. 4, pp. 1–14, Aug. 2019.
[54] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7824–7833.
[55] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications," Data Min. Knowl. Discov., vol. 2, no. 2, pp. 169–194, 1998.
[56] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-V4, inception-resnet and the impact of residual connections on learning," in Proc. 31st AAAI Conf. Artif. Intell., 2017, pp. 4278–4284.
[57] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit., May 2018, pp. 67–74.
[58] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 815–823.
[59] B. Dolhansky et al., "The DeepFake detection challenge (DFDC) dataset," 2020, arXiv:2006.07397.
[60] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV),

Oct. 2019, pp. 6201–6210.
[61] W. Kay et al., "The kinetics human action video dataset," 2017, arXiv:1705.06950.
[62] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 200–210.
[63] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4724–4733.
[64] Z. Liu et al., “TEINet: Towards an efficient architecture for video recognition," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 11669–11676.
[65] X. Li et al., "Sharp multiple instance learning for deepfake video detection," in Proc. 28th ACM Int. Conf. Multimedia, Seattle, WA, USA, C. W. Chen, R. Cucchiara, Z. Zhang, and R. Zimmermann, Eds. ACM, 2020, pp. 1864–1872.
[66] Z. Gu et al., “Spatiotemporal inconsistency learning for deepfake video detection," in Proc. 29th ACM Int. Conf. Multimedia, Chengdu, China, 2021, pp. 3473–3481.
[67] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021, arXiv:2102.11126.
[68] S. Seferbekov. (2020). DFDC 1st Place Solution. [Online]. Available: https://github.com/selimsef/dfdc_deepfake_challenge
[69] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake detection scheme based on vision transformer and distillation," 2021, arXiv:2104.01353.
[70] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, arXiv:1910.08854.
[71] M. Caron et al., "Emerging properties in self-supervised vision transformers," in Proc. ICCV, Oct. 2021, pp. 9650–9660.
[72] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proc. CVPR Workshops, 2019, pp. 1–7.
[73] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS), Sep. 2019, pp. 1–8.
[74] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in Proc. CVPR Workshops, 2019, pp. 1–8.