VidAvDetect: A Deepfake-Inspired Vision Transformer Approach for Detecting Real Humans vs. AI-Avatars in Video Streams

Btissam Acim, Hamid Ouhnni, Nassim Kharmoum, Soumia Ziti
Computer Science Department-Intelligent Processing and Security of Systems (IPSS) Laboratory-Faculty of Sciences, Mohammed V University in Rabat, Morocco

***Abstract*—The pace of advancement in Generative AI has made it possible to realize highly realistic synthetic identities in the form of avatars for non-existent persons, thus paving the way for a paradigm beyond state-of-the-art deepfake attacks that aim to manipulate real identities in people. This rapidly emerging trend poses a challenge to digital media forensics in a most critical way, in terms of deciding whether a facial identity observed in a video clip represents a real human identity versus a fully synthetic identity created using advanced tools in the realm of Generative AI. To address this gap, we introduce VidAvDetect, a deepfake-inspired Vision Transformer approach specifically designed to discriminate real human faces from AI-generated avatars in video streams, addressing a novel identity-existence verification task. The proposed system integrates efficient frame sampling, robust facial preprocessing, patch-based embeddings, and global structural modeling through a transformer encoder, enabling the detection of subtle geometric and textural regularities characteristic of synthetic identities. Experimental results demonstrate strong performance, with training accuracy reaching 97–98%, video-level accuracy of 95.1%, a macro F1-score of 0.944, and a ROC-AUC of 0.991, confirming the model's robustness across heterogeneous real, manipulated, and fully synthetic datasets. By moving beyond manipulation detection to focus on identity-existence verification, VidAvDetect establishes a new methodological direction for transparency, regulation, and trust in modern digital media environments where AI-generated avatars increasingly resemble real humans.**

Keywords*—Vision transformer; deepfake; Artificial Intelligence (AI); Generative AI; AI Avatar; video streams

I. Introduction

The rapid rise of generative technologies has profoundly transformed the production and consumption of digital content. Beyond traditional deepfakes which rely on manipulating images of real individuals an entirely new category of artificial identities has emerged [1]: AI-generated avatars of non-existent people. These synthetic characters, produced by models such as GANs, diffusion architectures, and advanced video-avatar engines, replicate with remarkable realism human appearance, facial expressions, movements, and even emotions, despite having no authentic existence. Their presence is expanding across numerous domains [2], including advertising, marketing campaigns, promotional videos, digital media, and social platforms. They are now used to portray fictitious customers, fabricated witnesses, artificial presenters, or so-called experts, often without users being aware of their synthetic nature.

This evolution raises a major concern: as an increasing amount of video content features artificial identities presented as real individuals, the boundary between genuine humans and AI-generated avatars [3] becomes increasingly difficult for the public to discern. The growing use of these non-existent personas provides significant advantages to content producers, such as reduced costs [4], full control over messaging, instant multilingual production, and unlimited availability in interactive formats. However, for internet users, this opacity creates a problematic zone of ambiguity, undermining the reliability of information, weakening trust in media content, and opening the door to new forms of visual manipulation.

In this context, a central question emerges: how can an internet user determine whether the person they see in a video is a real human being or a non-existent AI-generated avatar? Although closely related to the evolution of deepfakes [5], this issue goes beyond the usual challenges of identity falsification. It concerns characters that are entirely generated, with no correspondence to any real person, and whose detection requires approaches fundamentally different from those used for traditional deepfake analysis. To date, no major study has directly addressed this question in the scientific literature [6], even though it is becoming essential for ensuring digital transparency and protecting users against the proliferation of undisclosed synthetic video content.

The fundamental innovation of VidAvDetect lies in its ability to tackle a new classification task focused not on detecting manipulated identities, but on identifying faces that belong to individuals who do not exist at all. This conceptual shift significantly expands the scope of synthetic-content detection and addresses the growing need for transparency in advertising and media environments. By relying on a transformer-based model with strong representational capabilities, our approach captures subtle yet recurrent divergences between real human faces and their artificial counterparts [7], even when the latter are produced with high visual fidelity.

Therefore, VidAvDetect introduces a new methodological direction capable of confronting the rise of artificial avatars [8] whose use is rapidly spreading across digital ecosystems. By emphasizing the detection of non-existent identities, this approach provides a concrete response to a phenomenon that remains understudied but is likely to become a major issue in the

verification, regulation, and interpretation of modern video content.

From a research perspective, the emergence of fully AI-generated avatars introduces a critical and still insufficiently addressed scientific challenge. Although current methods of detecting deepfakes are mainly concentrated on detecting visual manipulation techniques used for real human images, their underlying assumption, based on which the image of a human being has to exist in reality, does not continue to hold true in the face of non-existent images of synthetic identities. To address this gap, this work introduces VidAvDetect, which shifts the detection objective from manipulation analysis to identity-existence verification in video streams. In addition to this, by making use of the Vision Transformer architecture and video-level aggregation method, it becomes possible for the proposed solution to offer an informed and specialized method for making distinctions between human identities and purely artificial avatars in digital media.

The remainder of this paper is structured as follows. Section II reviews the related literature; Section III describes the proposed VidAvDetect model; Section IV reports the experimental results; Section V presents the discussion, including ethical considerations; and Section VI provides the conclusion along with future research directions.

II. Related Work

The detection of synthetic media has been extensively studied in the context of deepfakes [9], where the identity of a real individual is modified using generative networks, encoder–decoder architectures, or facial transfer techniques. Existing models primarily rely on convolutional networks, frequency-domain analysis, or attention mechanisms to identify artifacts introduced during face swapping, reenactment, or warping operations. These methods consistently assume the presence of an authentic human identity that has been altered. However, such approaches become inadequate in the face of a new generation of synthetic identities: AI-generated avatars of non-existent person. Unlike traditional deepfakes [10], these faces are created entirely by diffusion models, StyleGAN architectures, or animated avatar engines, without any correspondence to a real person. They do not exhibit the typical irregularities associated with manipulation but instead display coherent and visually consistent features, making them significantly more difficult to detect. Recent studies on AI-generated faces focus mainly on static images and do not address the video dimension nor the verification of whether the displayed identity corresponds to an actual human.

Another line of research examines synthetic profiles on social networks by analyzing behavioral patterns, metadata, or interaction traces. However, these works do not involve visual analysis and do not aim to distinguish real human faces from AI-generated avatars.

Vision Transformer–based models have demonstrated strong performance in deepfake detection, due to their ability to capture global relationships and complex facial structures. Nevertheless, these models are trained exclusively on datasets containing manipulated versions of real identities [11], not fully synthetic faces. None of them is designed to discriminate between a real human and a non-existent AI-generated avatar.

To the best of our knowledge, no prior work has proposed a method capable of determining, within a video, whether the observed face belongs to a real person or to an identity entirely generated by AI [12]. This gap constitutes a notable missing link in the scientific literature and highlights the need for a dedicated approach such as VidAvDetect to address this emerging problem.

III. Proposed Model

The VidAvDetect model aims to determine whether a face extracted from a video sequence belongs to a real individual or to an AI-generated avatar representing a non-existent person. The approach relies on a Vision Transformer, selected for its ability to capture global facial structures and long-range dependencies that are essential for distinguishing authentic human faces from synthetic identities generated [13] by diffusion models, GANs, or avatar-animation engines. As illustrated in Fig. 1, the pipeline begins with the preprocessing and extraction of faces from video frames, after which each face is divided into patches and analyzed by the transformer to capture the textural and structural coherence characteristic of fully generated faces. VidAvDetect then encodes each face and classifies it as real or synthetic [14], while a temporal aggregation step integrates frame-level predictions to deliver a final video-level decision. This process makes VidAvDetect a robust and efficient model for detecting AI-generated non-existent humans in video streams an ability not offered by conventional deepfake detection methods.

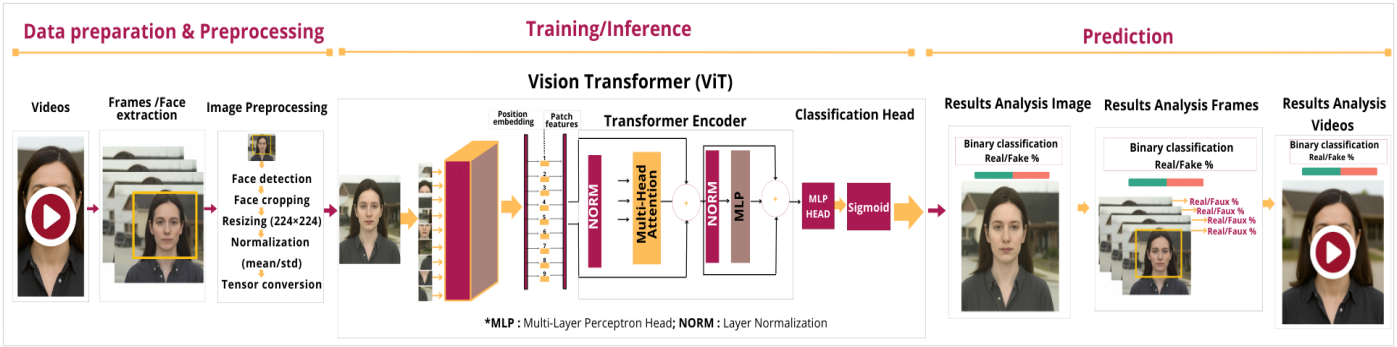

Fig. 1. Architecture of the VidAvDetect approach for real-human vs. AI-avatar detection.

A. Overview of the VidAvDetect Approach

The proposed VidAvDetect approach aims to address a growing challenge in digital content authentication [15]: distinguishing real human identities from AI-generated avatars representing non-existent individuals. With the increasing use of such synthetic characters in advertising, promotional videos, and online communication, the system seeks to provide a reliable mechanism for verifying whether a face appearing in a video truly corresponds to an authentic person. VidAvDetect operates by extracting facial regions from video frames, normalizing them through a series of preprocessing steps, and encoding them using a Vision Transformer capable of capturing the global structures and facial regularities that differentiate real humans from fully synthetic identities. This methodological combination gives VidAvDetect a level of precision and robustness particularly suited to modern contexts where humans and AI-generated avatars have become visually difficult to distinguish.

B. Motivation for Choosing Vision Transformer

The choice of a ViT as the backbone of the VidAvDetect approach is motivated by the need to capture subtle global patterns that distinguish real human faces from AI-generated avatars. Unlike CNN-based architectures that operate primarily through local receptive fields, ViT processes images as sequences of patches, enabling it to model long-range spatial dependencies and globally coherent structures across the entire face. This capability is particularly relevant for detecting synthetic identities [16], which often exhibit consistent textures, uniform facial alignment, and regular geometric patterns produced by diffusion models, GAN architectures, or avatar-generation engines.

Moreover, ViT has demonstrated strong generalization abilities on heterogeneous datasets and shows robustness when dealing with high-fidelity synthetic content, where local manipulation artifacts are minimal or absent. Its self-attention mechanism allows the model to compare facial regions holistically, capturing identity-level irregularities that traditional deepfake detectors may overlook.

These strengths make Vision Transformers especially suitable for distinguishing synthetic faces that do not originate from real individuals a requirement central to the VidAvDetect task. In addition, ViT [17] integrates seamlessly into a modular pipeline, offering scalability for video-level prediction through the aggregation of frame-based embeddings. This combination of global feature modeling, robustness to varied generative styles, and compatibility with video inference provides a solid foundation for reliable discrimination between real human identities and AI-generated avatars.

C. Data Preparation and Face Preprocessing

The data preparation stage relies on a diverse collection of video sources combining real human recordings, deepfake benchmarks, and fully synthetic AI-generated avatars. This heterogeneous composition ensures that the model is exposed to various real-world conditions, manipulation techniques, and generative styles. Table I summarizes the datasets used in our experiments, including the number of videos, total duration, and distribution between real and synthetic samples.

TABLE I. OVERVIEW OF THE VIDAVDETECT VIDEO DATASET

Source	Number of videos	Total duration (minutes)	Fake	Real
FaceForensics++	100	60	50	50
DFDC	150	90	75	75
DeepfakeTIMIT	40	25	20	20
Celeb-DF	30	20	15	15
Synthesized (AI)	20	12	20	0
RealVids2020	30	18	0	30
UCF101	40	25	0	40

Each video is sampled into frames at a fixed extraction rate to ensure a balanced temporal representation across all sources. For every extracted frame, a face detection module is applied to localize the facial region of interest. The detected faces are then cropped and resized to a standardized spatial resolution of 224 × 224 pixels, ensuring consistency with the Vision Transformer input format.

Preprocessing also includes pixel-value normalization (mean/std) and tensor conversion, producing uniform model-ready inputs regardless of the variability present in the original datasets. This step is essential for removing irrelevant background information and focusing the model exclusively on facial structures. It also harmonizes samples originating from different datasets, improving the stability and generalization capacity of the VidAvDetect model during training and inference.

D. Detailed Description of Model Blocks

The VidAvDetect architecture is composed of several sequential modules, each responsible for transforming the raw video input into a robust video-level authenticity prediction. This subsection describes the function of each block, from frame extraction to transformer-based encoding and classification [18].

- Frame Extraction Module: Each video is uniformly sampled to extract a limited number of frames, reducing computational cost and response time while maintaining sufficient temporal coverage. The number of extracted frames is computed as:

$$Frames_{extracted} = \frac{Video\_Duration \times FPS}{Sampling\_Rate}$$

- Video_Duration corresponds to the video length in seconds, FPS indicates the frame rate, and Sampling_Rate defines how frequently frames are selected; for example, selecting one frame every ten in a 60-second video recorded at 30 FPS reduces computation and improves response time while still preserving sufficient temporal coverage.
- Face Detection and Preprocessing Block: Sampled frames undergo face detection, and the extracted facial regions are cropped, resized to 224×224, normalized, and converted into tensors to produce consistent model-ready inputs.

- Patch Embedding and Tokenization: Each preprocessed face is divided into fixed-size patches, which are flattened and projected into embedding vectors enriched with positional information to form the token sequence for the transform.
- Vision Transformer Encoder: The transformer layers process the patch tokens through self-attention and feed-forward operations, enabling the model to capture global structural and textural cues distinguishing real human faces from synthetic ones.
- Classification Head: The final representation is passed through an MLP and a sigmoid function to generate a probability indicating whether the detected face is real or AI-generated.
- Video-Level Decision: Frame-level predictions are aggregated using temporal fusion to produce a stable and reliable video-level authenticity decision.

E. Comparative Analysis of Methods

To better situate the contribution of VidAvDetect within the landscape of facial authenticity analysis [19], we provide a concise comparison with existing deepfake detection approaches. While traditional models are primarily designed to identify manipulations applied to real faces [20], VidAvDetect focuses on determining whether the identity itself is real or AI-generated. Table II summarizes the key conceptual differences between existing detectors and the proposed VidAvDetect approach.

The comparison in Table II highlights the conceptual shift introduced by VidAvDetect. While traditional detectors focus on detecting modifications applied to real faces, VidAvDetect assesses whether the identity itself is real or artificially generated. This orientation aligns with the growing use of synthetic AI-generated avatars across digital media.

TABLE II. COMPARISON BETWEEN DEEPFAKE DETECTORS AND VIDAVDETECT

Aspect	Existing Deepfake Detectors	VIDAVDetect
Goal	Detect manipulated faces	Detect real vs AI-generated identities
Assumption	The person exists	The person may not exist
Feature Focus	Local artifacts	Global structural & synthetic cues
Target Content	Altered real faces	Fully synthetic AI avatars
Strength	Good for manipulation detection	Strong for identity-existence verification
Limitation	Fail on fully generated faces	Requires diverse synthetic data

IV. EXPERIMENTAL RESULTS

To evaluate the relevance of the proposed VidAvDetect approach, we designed a use case representative of current challenges in video-based identity verification. With the increasing presence of AI-generated avatars across advertising [21], social platforms, and promotional media, determining whether the person appearing in a video stream is real or synthetically generated has become essential for transparency and content reliability. The use case presented in Fig. 2 illustrates this setting by showing how VidAvDetect processes video inputs through frame sampling, facial extraction, and transformer-based classification to produce an identity authenticity assessment indicating whether the depicted individual corresponds to a real human or an AI-generated avatar.

Fig. 2. Experimental use case of the VidAvDetect approach.

Fig. 2 provides an overview of the complete VidAvDetect workflow, from video-stream ingestion to final authenticity prediction. The illustration shows how an advertisement video is first decomposed into uniformly sampled frames, then processed through face detection, cropping, and normalization before being analyzed by the transformer-based classifier.

The final stage aggregates frame-level probabilities to produce a video-level decision, here identifying the individual as an AI-generated avatar with a confidence score of 95%. This end-to-end flow highlights the practical value of VidAvDetect for real-world video streams, where users' viewers, regulators, or platforms need instant and reliable verification of whether a person shown on screen is real or artificially generated.

A. Binary Detection: Real Human vs. AI Avatar (Fake)

Fig. 3 presents the joint evolution of the training loss and accuracy of the VidAvDetect model during the binary classification task Real Human vs AI-Generated Avatar. This curve provides insight into the model's convergence behavior, training stability, and its ability to effectively separate authentic human faces from fully synthetic AI-generated identities.

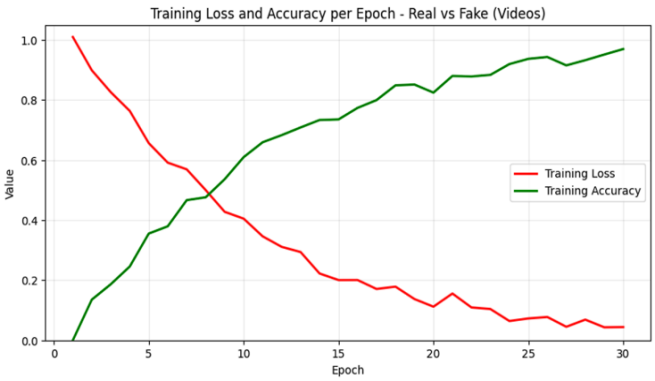

Fig. 3. Real vs. AI-avatar binary detection results.

As shown in Fig. 3, the training loss decreases steadily across epochs and eventually stabilizes at a very low value, while the accuracy consistently increases and reaches approximately 97–98% by the end of training. This trend demonstrates that VidAvDetect learns to discriminate real human faces from AI-generated avatars with high reliability, indicating strong class separability and robust generalization on the Real Human vs AI-Generated Avatar task.

B. Confusion Matrix and Metrics

To further evaluate the effectiveness of VidAvDetect in distinguishing real human identities from AI-generated avatars, we report the quantitative performance on the video-level binary classification task. This analysis includes detailed classification metrics precision, recall, F1-score, and overall accuracy as well as the confusion matrix shown in Fig. 4, allowing us to assess not only the global performance of the model but also its class-wise behavior. These results shown in Fig. 5 provide a clear view of how reliably VidAvDetect generalizes to unseen video streams.

As shown in Fig. 5, VidAvDetect achieves a strong and well-balanced performance across both classes. The model correctly classifies 44 out of 46 real identities and 34 out of 36 AI-generated avatars, resulting in an overall accuracy of 95.1 %, a macro-averaged F1-score of 0.944, and a ROC-AUC of 0.9505. These results confirm that the model consistently captures the discriminative patterns separating real human faces from artificially generated ones.

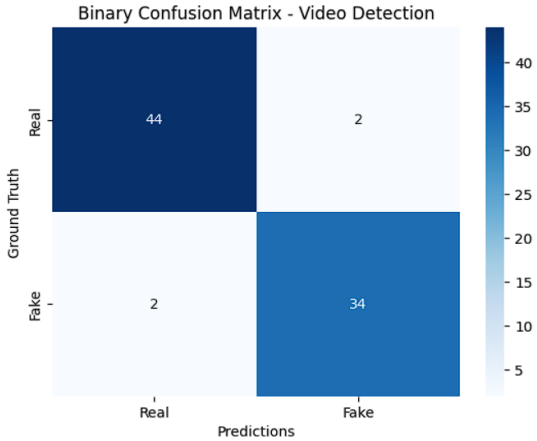

Fig. 4. Binary video detection confusion matrix.

	precision	recall	f1-score	support
Real	0.94	0.94	0.94	36
Fake	0.96	0.96	0.96	46
accuracy			0.95	82
macro avg	0.95	0.95	0.95	82
weighted avg	0.95	0.95	0.95	82

Accuracy: 0.9512
Precision: 0.9444
Recall: 0.9444
F1 Score: 0.9444
MCC: 0.9010
Cohen's Kappa: 0.9010
ROC AUC: 0.9505

Fig. 5. VidAvDetect classification performance results.

C. ROC–AUC Evaluation

To further assess the discriminative capacity of VidAvDetect on the binary classification task Real Human vs. AI-Generated Avatar, we computed the Receiver Operating Characteristic (ROC) curve, as shown in Fig. 6.

This metric provides a threshold-independent evaluation of the model's ability to separate the two classes by plotting the true positive rate against the false positive rate across all decision thresholds. A higher Area Under the Curve (AUC) value indicates stronger separability and more reliable decision-making.

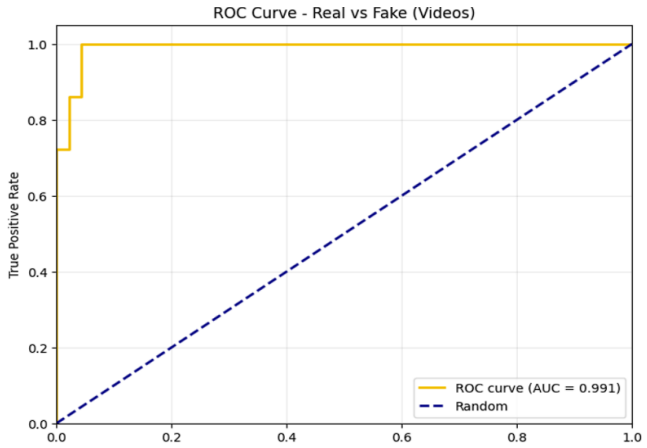

Fig. 6. ROC–AUC curve for binary detection.

VidAvDetect achieves an exceptionally high AUC of 0.991, indicating almost perfect separability between real human faces and AI-generated avatars. The ROC curve remains near the top-left corner, confirming that the model maintains a very high true positive rate even when the decision threshold is varied. These results reinforce the robustness of VidAvDetect and align consistently with the previously reported accuracy (~95%), precision, recall, and confusion matrix performance, demonstrating the model's reliability in distinguishing authentic identities from synthetic ones.

V. Discussion

The experimental findings demonstrate that VidAvDetect addresses a critical gap not covered by existing deepfake-detection research. Traditional deepfake detectors focus on

identifying manipulations applied to real human identities [22] such as face swapping, reenactment, or warping and therefore rely on the presence of local artifacts generated during the manipulation process.

These approaches are inherently limited when confronted with fully AI-generated identities, which are produced natively by GANs and diffusion models and therefore do not exhibit the typical distortion patterns exploited by classical algorithms.

The numerical results highlight the effectiveness of the proposed approach. The model achieves an almost perfect separability between real and synthetic identities, as reflected by the ROC–AUC of 0.991. This indicates that VidAvDetect detects global structural signatures geometric regularities, statistical homogeneity of textures, and the absence of natural micro-variations that persist across avatars generated by different architectures.

The balanced classification performance (44/46 real faces and 34/36 AI-generated avatars correctly identified) combined with a macro-F1 score of 0.944 shows that the system avoids both over-detection and under-detection, a limitation frequently observed in conventional deepfake detectors confronted with high-realism synthetic faces.

The video-level accuracy of 95.1% further demonstrates the advantage of incorporating temporal aggregation. Whereas many existing studies are restricted to static images, VidAvDetect maintains its discriminative ability across frames, enabling consistent identity-existence verification in real video streams. This is particularly relevant as modern AI-avatar engines produce smooth, artifact-free video sequences that would evade detectors based solely on frame-level texture inconsistencies.

The results collectively indicate that the Vision Transformer architecture plays a decisive role in the model's performance. By capturing long-range dependencies and holistic facial structures, the transformer overcomes the limitations of CNN-based approaches tied to local receptive fields. This capability becomes crucial [23] as generative models continue to evolve toward near-photorealism.

Nevertheless, the discussion must acknowledge that technological progress in diffusion-based avatar generation may gradually reduce the visual discrepancies currently detectable. Ensuring long-term robustness will require periodic dataset expansion, model retraining, and potentially the inclusion of multimodal cues such as audio-visual synchrony, micro-expression dynamics, or physiological consistency.

Overall, VidAvDetect introduces a methodological shift from manipulation detection to identity-existence verification, representing a significant advancement for authenticity assessment in digital media. As artificially generated personas become increasingly indistinguishable from real individuals, such models will be essential for maintaining transparency, trust, and integrity across online environments.

In addition to these technical considerations, the growing use of highly realistic AI-generated avatars also raises important ethical and legal issues. Such synthetic identities can compromise the credibility of digital content by enabling undisclosed impersonation, fabricated testimonies, or misleading representations that appear visually indistinguishable from real individuals.

Ensuring that viewers can verify whether a portrayed identity truly exists is therefore essential for preserving trust, preventing misuse, and supporting emerging regulatory requirements that mandate transparency in the dissemination of AI-generated media.

VI. Conclusion

VidAvDetect establishes a new direction in video-based identity authentication by addressing a problem that remains absent from existing deepfake-detection literature: determining whether the individual appearing in a video is an actual human or an AI-generated avatar with no real-world existence. Unlike conventional detectors that rely on identifying manipulation artifacts applied to genuine identities, the proposed model focuses on discriminating fully synthetic faces produced natively by diffusion models, GAN architectures, and advanced avatar-generation engines.

The experimental evaluation highlights the effectiveness of this approach. VidAvDetect achieves an overall video-level accuracy of 95.1%, with balanced classification performance 44/46 real faces and 34/36 AI-generated avatars correctly identified. Complementary metrics, including a macro F1-score of 0.944, precision and recall values of 0.944, and a ROC–AUC of 0.9505, confirm the model's ability to reliably separate authentic human faces from their synthetic counterparts. These results demonstrate that the Vision Transformer architecture effectively captures global structural cues and synthetic regularities that persist across diverse generative styles, while the temporal aggregation mechanism ensures stable predictions across continuous video sequences.

The model therefore provides a solid foundation for identity-existence verification in real video streams an increasingly crucial capability as AI-generated personas continue to proliferate in advertising, marketing, entertainment, social platforms, and promotional media. VidAvDetect offers users, platforms, and regulators a practical mechanism for assessing whether a video identity corresponds to a real individual, thereby contributing to digital transparency and strengthening the reliability of modern multimedia content.

Despite these promising results, this study presents certain limitations that should be acknowledged. First, the proposed approach relies exclusively on visual facial information and does not currently integrate multimodal cues such as audio signals, speech consistency, or physiological indicators, which may provide complementary evidence in complex real-world scenarios. Second, although the experimental evaluation covers heterogeneous datasets including real, manipulated, and fully synthetic videos, the diversity of emerging avatar-generation techniques continues to evolve rapidly, which may introduce new generative patterns not fully represented in the current data. Finally, the model focuses on frontal or near-frontal facial content, and its performance may be affected in cases involving extreme poses, occlusions, or low-quality video streams. Addressing these limitations constitutes an important direction for future research.

Future developments may incorporate multimodal features such as audio–visual synchrony, micro-expression dynamics, or temporal physiological cues to further enhance detection robustness as generative technologies evolve. In parallel, the growing presence of realistic AI-generated avatars raises significant ethical and legal considerations. Synthetic identities can be used to fabricate testimonies, impersonate experts, or disseminate misleading narratives that appear indistinguishable from authentic human communication. Ensuring that viewers can verify whether an identity genuinely exists is therefore essential to protect public trust, uphold transparency, and comply with emerging regulatory frameworks governing the disclosure of AI-generated media.

REFERENCES

[1] B. Acim, N. Kharmoum, S. N. Lagmiri, and S. Ziti, "The role of generative AI in deepfake detection: A systematic literature review," in Proc. Int. Conf. Smart Business and Technologies (ICSBT 2024), vol. 1330, Lecture Notes in Networks and Systems, S. N. Lagmiri, M. Lazaar, and F. M. Amine, Eds. Cham, Switzerland: Springer, 2025, pp. 349–357. doi: 10.1007/978-3-031-86698-2_31.

[2] B. Liu, H. Jiang, Y. Wang, and X. Xu, "A review of deepfake and its detection: From generative adversarial networks to diffusion models," Int. J. Intell. Syst., vol. 40, no. 5, pp. 12567–12592, 2025. doi: 10.1155/int/9987535.

[3] Y. Gong, J. Zhang, Y. Wang, and Z. Liu, "A contemporary survey on deepfake detection: Datasets, algorithms, and challenges," Electronics, vol. 13, no. 3, p. 585, 2024. doi: 10.3390/electronics13030585.

[4] F. Lago, C. Pasquini, R. Böhme, H. Dumont, V. Goffaux, and G. Boato, "More real than real: A study on human visual perception of synthetic faces," IEEE Signal Processing Magazine, vol. 39, no. 4, pp. 109–116, July 2021. doi:10.1109/MSP.2021.3120982.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4401–4410. doi: 10.1109/CVPR.2019.00453.

[6] A. Alsaeedi, A. AlMansour, and A. Jamal, "Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 6, 2025. doi:10.14569/IJACSA.2025.0160622.

[7] B. Acim, H. Ouhnni, N. Kharmoum, S. Ziti, and A. Ouajdouni, "The era of blockchains, deepfakes, and NFTs in therapeutic treatments: A comparative study," in Non-Fungible Tokens (NFTs) in Smart Cities: Advancements and Security Challenges, U. Khalil, M. Uddin, O. Malik, and A. Dandoush, Eds. Hershey, PA, 650 USA: IGI Global Scientific Publishing, 2026, pp. 119–152, doi: 10.4018/979-8-3693-8876-1.ch005.

[8] A. Hidouri, S. Jabbour, and B. Raddaoui, "On the enumeration of frequent high utility itemsets: A symbolic AI approach," in Proc. 28th Int. Conf. Principles and Practice of Constraint Programming (CP), LIPIcs, vol. 235, 2022, pp. 27:1–27:17. doi: 10.4230/LIPIcs.CP.2022.27.

[9] H. Ouhnni, B. Acim, M. Belhiah, K. El Bouchti, Y. Z. Seghroucheni, S. N. Lagmiri, R. Benachir, and S. Ziti, "The evolution of virtual identity: A systematic review of avatar customization technologies and their behavioral effects in VR environments," Front. Virtual Real., vol. 6, 2025. doi: 655 10.3389/frvir.2025.1496128.

[10] H. Wu, L. Leng, and P. Yu, "Learning Local Reconstruction Errors for Face Forgery Detection," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 11, 2024. doi:10.14569/IJACSA.2024.01511120.

[11] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, et G. Amato, "Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection," in Proc. 1st Int. Workshop on Multimedia AI against Disinformation (MAD 2022), Newark, NJ, USA, June 27–30 2022, ACM, New York, NY, USA, pp. 52– 58. doi:10.1145/3512732.3533582.

[12] Z. Xia, T. Qiao, M. Xu, X. Wu, L. Han, and Y. Chen, "Deepfake video detection based on MesoNet with preprocessing module," Symmetry, vol. 14, no. 5, Art. no. 939, 2022. doi:10.3390/sym14050939.

[13] B. Acim, N. Kharmoum, M. Ezziyyani, and S. Ziti, "Mental health therapy: A comparative study of generative AI and deepfake technology," in Proc. Int. Conf. Adv. Intell. Syst. Sustain. Dev. (AI2SD 2024), vol. 1403, Lecture Notes in Networks and Systems, M. Ezziyyani, J. Kacprzyk, and V. E. Balas, Eds. Cham, Switzerland: Springer, 2025, pp. 351–357. doi: 10.1007/978-3-031-91337-2_33.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[15] M. I. Ahmad, H. Alkhalifah, A. A. Khaleel, and A. A. Al-Khalifa, "Extracting Facial Features to Detect Deepfake Videos Using Machine Learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 4, 2025. doi:10.14569/IJACSA.2025.0160417.

[16] C. Ben Rabah, I. N. Petropoulos, R. A. Malik, and A. Serag, "Vision transformers for automated detection of diabetic peripheral neuropathy in corneal confocal microscopy images," Frontiers in Imaging, vol. 6, 2025. doi: 10.3389/fimag.2025.1542128.

[17] H. Mancy, M. Elpeltagy, K. Eldahshan, and A. Ismail, "Hybrid-Optimized Model for Deepfake Detection," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 16, no. 4, 2025. doi: 10.14569/IJACSA.2025.0160417..

[18] M. Elpeltagy, A. Ismail, M. S. Zaki, and K. Eldahshan, "A Novel Smart Deepfake Video Detection System," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 1, 2023. doi:10.14569/IJACSA.2023.0140144.

[19] H. A. Soudy, O. Sayed, H. Tag-Elser, R. Ragab, S. Mohsen, T. Mostafa, A. A. Abohany, and S. O. Slim, "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks," Neural Computing and Applications, vol. 36, no. 15, pp. 10245–10261, Aug. 2024. doi:10.1007/s00521-703 024-10181-7.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[21] Z. Wang, Z. Cheng, J. Xiong, X. Xu, T. Li, B. Veeravalli, and X. Yang, "A timely survey on Vision Transformer for deepfake detection," arXiv preprint, 2024. doi: 10.48550/arXiv.2405.08463.

[22] H. Wu, L. Leng, and P. Yu, "Learning local reconstruction errors for face forgery detection," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 11, 2024. doi:10.14569/IJACSA.2024.01511120.

[23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.