Spatiotemporal Consistency-Based Deep Forgery Detection

Peng Wen[1*]

International School, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. With the rapid development of generative artificial intelligence (AIGC) and deep learning technology, deep forgery technology based on generated adversarial network (GAN) and diffusion model can generate extremely realistic facial videos and images, posing a serious threat to the social trust system and information security. Existing research shows that fake videos often show subtle inconsistencies in space-time dimensions, such as light changes between frames, texture jitter and abnormal motion blur. These "space-time inconsistency" have become an important basis for identifying deep forgery. This article summarizes the deep forgery detection method based on the characteristics of space-time consistency by systematically sorting out existing research. First, explain the role of space-time consistency characteristics in video forgery detection; secondly, review the typical detection framework from the three dimensions of time domain, airspace and depth, including abnormal detection based on time-domain light flow, feature fusion methods based on spatial attention, and anti-deception models combining gradient and depth estimation; then analyze commonly used Data sets and evaluation indicators; finally summarize the shortcomings of existing research and look forward to the future direction, such as multimodal fusion and cross-domain generalization. This article aims to provide systematic research reference and development ideas for deep forgery video detection.

1 Introduction

Deep forgery technology is a generative method that uses deep neural networks to synthesize, tamper with or replace facial images or videos. In recent years, a major breakthrough has been made in the generative method based on the generation of confrontation network (GAN) and diffusion model. The generated video is almost difficult to distinguish from the real video in terms of facial expressions, lighting conditions and action coherence [1]. Although such technologies have a positive impact on areas such as film production and virtual character modeling, their abuse has led to the increasing proliferation of fake political videos, pornographic content and false information, posing a serious threat to social security [2].

Deep forgery technology is mainly divided into two categories: facial replacement and facial replay [3]. Facial replacement technology replaces the facial area of the source image with the facial area of the target image, but only rendering the facial area often leads to fusion

* Corresponding author's email: wenpeng@bupt.edu.cn

artifacts at the boundary, which is inconsistent in color and texture between the face and the background [3]. The facial replay technology aligns the expression and movements of the source face to align it with the target face. Although identity information is retained, this method can easily lead to a lack of time coherence in facial expression changes [3]. There is a space-time inconsistency in both technologies, which provides a key breakthrough for detection and research.

At present, deep forgery detection technology faces two major challenges: first, the coverage of the diversified forgery types generated by various forgery technologies is insufficient; second, the robustness of the detection method is not strong enough when it comes to degraded scenarios such as image blur, noise and compression [4]. The detection method based on space-time consistency responds to these challenges by identifying the physical inconsistency of forged content in spatial textures and time motion patterns. This method has become an important research direction in the field of in-depth forgery detection.

2 The role of spatio-temporal consistency in deep forgery detection

Spatio-temporal consistency in the video refers to the continuity of objects in terms of spatial texture and temporal movement, covering facial dynamics, expression transition and light consistency [1]. In real videos, these features follow the principles of natural physics - such as the coordination of muscle movement when facial expressions change, the stability of light intensity and direction between frames, and the geometric consistency of three-dimensional facial structures when moving [5]. In contrast, synthetic videos often show space-time inconsistency due to problems such as frame-by-frame generation, mismatch of synthetic parameters or insufficient modeling of physical laws [6].

In the spatial domain, fake videos often have problems such as local texture anomalies and geometric inconsistencies. For example, facial forged images generated by the confrontation network may have blurred texture or unnatural pixel distribution in detail areas such as the eyes and mouth [7], while facial replacement technology often leads to differences in color saturation and noise patterns between the forgery area and the background [8]. The model proposed by Chen and others accurately locates such spatial forgery areas through the spatial attention module [2], thus verifying the validity of spatial consistency characteristics.

Time inconsistency is particularly prominent in the characteristics of deep falsified videos. For example, facial expression mutations, intermittent head movements, sudden changes in lighting conditions and other artifacts often appear in forged content [6]. The time-domain surface frame (t-SF) method proposed by Ciamarra et al. captures the physical discontinuity generated by frame-by-frame synthesis by analyzing the chronological evolution of surface norms, tangents and light changes [4]. In addition, facial micro-actions such as blinking and fine muscle contraction in real videos show stable time laws, and these actions are often difficult to reproduce accurately in synthetic videos, resulting in dynamic distortion [9].

In summary, the spatial and temporal consistency characteristics comprehensively characterize the intrinsic differences between real video and fake video from the three dimensions of spatial texture, time motion and three-dimensional depth, thus providing a key discrimination basis for deep forge detection [1].

3 Classification of detection methods based on spatio-temporal consistency

At present, detection methods based on space-time consistency can be divided into three categories: temporal consistency analysis, spatial consistency analysis and joint space-time modeling, as shown in Figure 1.

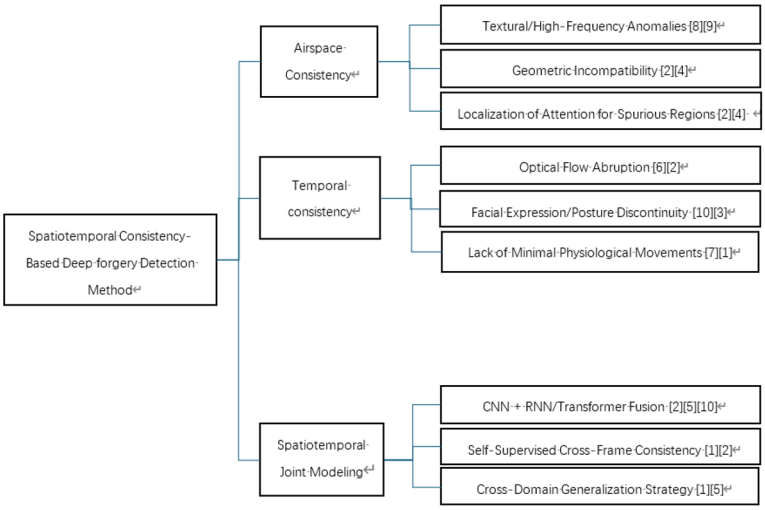

Fig. 1. Categorization of Spatio-Temporal Consistency Detection Methods (Picture credit: Original).

3.1 Temporal consistency-based detection methodology

The time consistency detection method mainly analyzes the time evolution pattern in the video frame sequence to identify the inconsistency between frames caused by content tampering. Such methods usually use circular neural networks, light flow analysis or time feature modeling techniques to reveal abnormal changes in interframe features (such as motion, light and facial expressions). Early studies found that there was a significant deviation in the time distribution of physiological behaviors such as blinking in the deeply forged video, thus establishing a basic physiological a priori for time consistency analysis [10].

The timing surface frame (t-SF) method proposed by Ciamarra et al. captures forgery anomalies by analyzing the timing evolution of surface frames in video sequences. This method uses information such as surface vectors, tangents and light changes to build local and global surface characterization, and uses bidirectional LSTM to learn the time series to detect the physical discontinuity caused by frame-by-frame generation. Experiments on the FaceForensics++ (FF++) data set show that the average detection accuracy of the method combined with local surface frames (t-SF8) reaches 89.7%, which is significantly better than the non-timed method [4].

Zheng et al. proposed a full-time sequence convolution network (FTCN), which forces the model to focus on learning time sequence characteristics by setting the size of the spatial convolution kernel to 1, so as to reduce excessive dependence on spatial artifacts. This method introduces timing Transformer to capture long-term dependencies. When conducting the cross-validation experiment on the FF++ data set, the average AUC reaches 99.7%, and shows excellent generalization ability for types that have never been forged [6].

Although the existing method can capture time discontinuity through surface geometry modeling and complete time feature learning, its essence stems from the dynamic inconsistency problem in video sequences. Gu et al. systematically characterized such problems by systematically modeling local time dynamic anomalies [11].

Guera and Delp proposed a deep forgery detection method based on circular neural networks (RNNs), which uses convolutional neural networks (CNNs) to extract frame-level features, and then uses RNNs to model the time dependency between frames, so as to capture the inconsistency between frames in deep forgery videos. Yeh et al. further combined ResNeXt with RNNs to achieve a detection accuracy of 85.13% and an F1 score of 0.9 on data sets such as DFDC and FF++, verifying the effectiveness of time modeling in deep forgery detection [9].

In addition, light flow analysis is a key technology for detecting time motion abnormalities. A light flow detection method based on convolutional neural networks (CNN) was proposed to identify fake videos by analyzing the abnormal patterns of the light flow field between frames[12]. Chen et al. used light flow to characterize time characteristics and combined it with the spatial attention module to realize the integration of space-time characteristics. They achieved an AUC value of 0.981 on the FF++ (LQ) data set, which is better than the traditional method [2].

3.2 Detection methodology based on spatial-domain features and attention mechanism

The spatial consistency detection method mainly identifies spatial characteristic abnormalities in single-frame images or local areas. By using techniques such as attention mechanism, texture analysis or frequency domain feature extraction, these methods can accurately locate the spatial anomalies of the forged area, such as texture blur, color distortion and geometric distortion. The study shows that the generation model still has obvious shortcomings in fully reproducing the statistical distribution characteristics and high-frequency characteristics of real images. This limitation provides key theoretical support for the forgery detection method based on frequency domain analysis and noise level analysis [13].

The space-time consistency and attention mechanism fusion model proposed by Chen and others uses the ResNet trunk network to extract spatial features, locates the forged area through the spatial attention module, and captures interframe changes in combination with the time attention module. By integrating shallow texture enhancement features and deep semantic characterization, the model can not only effectively identify forged traces, but also maintain strong generalization ability. The experimental results on the FF++ and DFDC data sets show that the model not only achieves excellent detection performance, but also shows significant advantages in memory efficiency and computing cost [2].

Jia et al. proposed an inconsistent perception wavelet double-branch network, which enhances the forgery characteristics through static wavelet decomposition (SWD), and uses a double-branch architecture to predict image-level and pixel-level forgery labels respectively. Cross-branch features are integrated through bilinear pooling. When evaluated on the FF++ (c40 compression) data set, the method achieved an accuracy rate of 88.96% and a true positive rate (TPR) of 79.03% when FPR = 0.1, which is significantly better than baseline models such as Xception [8].

Luo et al. proposed a general detection model based on high-frequency characteristics, which uses SRM filters to extract multi-scale high-frequency noise characteristics. By designing a residual-guided spatial attention module, the RGB feature extractor is guided to focus on forged traces; at the same time, the cross-modal attention module is used to integrate high-frequency features and RGB features. The model showed excellent performance in the FF++ cross-method experiment, verifying the role of high-frequency features in alleviating texture overfitting and improving generalization ability [7]. A multi-attention depth forgery detection method, which combines local regional texture characteristics with high-level semantic characteristics. By using the attention mechanism to enhance the feature description

of the forged area, this method has achieved remarkable results in single-frame detection. However, this method ignores the time information between frames and needs to be combined with time modeling technology to improve the detection performance at the video level [5].

3.3 Space-time joint modeling approach

The one-dimensional detection method has inherent limitations. The spatial domain method is susceptible to the interference of dynamic forgery, while the time domain method is susceptible to the interference of compression artifacts [1]. Space-time joint modeling technology can achieve more comprehensive consistency verification by combining spatial texture characteristics with temporal motion characteristics, and achieve the best balance between performance and robustness [1].

The point 4D transformer (P4Transformer) network proposed by Fan and others integrates space-time local structures through 4D convolutional embedding, and uses Transformer architecture to capture global appearance and motion patterns, thus avoiding the complexity involved in point tracking. The test on the MSR-Action3D data set shows that the action recognition accuracy of this method has reached 90.94%, which proves the effectiveness of joint space-time modeling in dynamic sequence analysis [12].

The detection model proposed by Amada et al. for the electronic "know your customers" (eKYC) system integrates the identity consistency characteristics between registered images and videos through auxiliary difference comparison (ADC) and time difference comparison (TDC), and uses circular neural network (RNN) to model it. Time relationship. The AUC of the model on the DFDCp data set reached 80.24% and demonstrated the strong robustness of image degradation [3].

SSTNet detects synthetic facial forgery by integrating spatial, implicit analysis and temporal characteristics. The framework uses LSTM to model time dynamics and uses Xception for convolutional feature extraction, showing powerful performance on the ICASSP 2020 data set [2].

3.4 Comprehensive comparative analysis of methods.

This chapter systematically sorts out the depth forgery detection methods based on space-time consistency, and deeply analyzes the principles, feature descriptions, advantages and limitations and application scenarios of various methods from the three dimensions of spatial domain, temporal domain and joint space-time modeling. The comparative analysis in Table 1 shows that the single-modal method has obvious shortcomings in detection accuracy, interpretability and robustness, while combined space-time modeling shows a stronger ability to adapt to complex scenarios in reality by integrating multi-dimensional features, which indicates the future development direction for in-depth forgery detection research.

Deep forgery detection usually uses two classification evaluation indicators, including accuracy rate, area under the subject's working characteristic curve (AUC) and F1 score, to evaluate detection accuracy and generalization ability.

Table 1. A Comparative Analysis of Feature-Based Approaches for Deep forgery Detection

Method Type	Advantages	Limitations	Applications	Leading-edge methodologies
Spatial Domain Features	Accurate localization of local forgeries, lightweight model	Difficult to handle dynamic forgeries,	Forensic detection of image forgeries	Jia [8], Luo [7]

		sensitive to compression		
Temporal Domain Features	Can capture motion anomalies and semantic inconsistencies	Sensitive to noise, weak in regional localization	Forensic detection of video forgeries	Zheng [6]
Spatio-Temporal Fusion	Strong generalization, high robustness, adaptable to complex scenarios	High model complexity, high training cost	Comprehensive detection, practical applications	Chen [2], Amada [3]

4 Commonly used datasets

The deep forgery detection data set is an important basis for method verification. The existing data set has a significant diversity in terms of forgery type, resolution and compression level to meet the evaluation needs of different application scenarios [1]. Among them, FaceForensics++ (FF++) is a publicly available video forgery data set, covering a variety of facial processing technologies and different compression levels. This data set is widely used in the comparative evaluation of deep forgery detection methods. It is the most widely used benchmark data set. Its diversified forgery techniques and compression quality characteristics can effectively evaluate the robustness of the model to forgery types and compression artifacts [4, 8, 14]. The Deep Forgery Detection Challenge (DFDC) data set is known for its large-scale and wide resolution range, which is suitable for performance verification in complex reality scenarios [2, 3]. Celeb-DF v2 is famous for its high-quality synthetic content and is often used in cross-data set generalization evaluation [3, 7]. Korea's deep forgery detection data set (KoDF) covers a wide range of identity information and video samples, mainly for practical application scenarios such as electronic customer authentication (eKYC) [3]. In addition, the data set including OULU-NPU, SiW and Replay-Attack is mainly designed for facial anti-deception research, and provides verification support for printing, playback and 3D mask attacks for detection methods [5]. It is worth noting that Deep Forgery Detection Challenge (DFDC) officially released by Kaggle is a large-scale benchmark test covering a variety of real recording conditions and forged technologies, and has become a key criterion for evaluating the generalization ability of models [15]. Table 2 shows the comparative analysis of commonly used datasets and characteristics for deep forgery detection.

Table 2. Comparative Analysis of Commonly Used Datasets and Characteristics for Deep Forgery Detection.

Data set	Unique characteristics	Reference
FaceForensics++ (FF++)	This benchmark data set covers common forgery techniques, such as deep forgery and face-to-face forgery, supports multiple compression levels, and represents one of the most widely used reference standards in the field.	[2]
Celeb-DF	The video has high visual fidelity and is accompanied by subtle defects, which is highly similar to the real Internet scene.	[9]
DFDC	This data set covers a variety of scenarios and racial characteristics, which is very suitable for evaluating the generalization ability.	[3]
WildDeep forgery	It comes from real tampered videos on the Internet, containing a lot of noise and uncontrolled variability.	[1]
FakeAVCeleb	Contains cross-modal forgery samples, integrating audio, video and port synchronization components.	[6]

LipForensics	Specially used to verify the consistency of mouth type and voice synchronization.	[6]

5 Current challenges and future research directions

5.1 Current challenges

With the breakthrough of cutting-edge technologies such as diffusion models and neural radiation fields (NeRF), the visual authenticity of deep forged videos has been significantly improved, while the traditional forged traces have gradually weakened, which has greatly increased the difficulty of detection [3]. At the same time, the existing detection models often show poor generalization ability in the face of unprecedented tampering technology and videos in real social media environments [9]. In addition, most methods still rely mainly on the visual characteristics of a single mode, and it is difficult to effectively identify multimodal forgery involving inconsistency between audio, lip action and semantic content [6]. Although the introduction of space-time modeling can improve the detection performance, it will greatly increase the complexity of the model, thus hindering its actual deployment in real-time monitoring and mobile device applications [2]. Although some studies try to achieve multimodal depth forgery detection through lip-shaped action and sound synchronization, these methods usually rely on specific forgery assumptions or are limited to controlled environments, which leads to their robustness and generalization capabilities in real social media scenarios and cross-domain conditions [16].

5.2 Future research directions

Future research directions can focus on integrating more explainable a priori knowledge - such as the physiological principles of controlling light transmission and facial muscle movement - into the detection model to enhance the ability to identify forgery mechanisms [4]. At the same time, the development of a multimodal depth forgery detection method that combines audio, text, semantics and visual information is expected to improve the identification performance of complex form forgery [6]. At the methodological level, combining self-supervised learning and comparative learning is a key way to enhance the cross-field generalization ability of the model and reduce the dependence on large-scale forgery samples [12]. From an application perspective, the implementation of lightweight model architecture through technologies such as model pruning and knowledge distillation is expected to improve deployment efficiency and practical application value [8]. The latest research shows that strategies such as potential space enhancement can significantly improve the generalization ability of the model to unknown counterfeit technology, providing new ideas for responding to rapidly evolving generative technologies [17].

6 Conclusion

The rapid progress of generating adversarial networks and diffusion models has made a qualitative leap in image quality and time coherence of deep forgery video, which makes existing detection technology face greater challenges. Research shows that although synthetic video can present realistic visual effects at the macro level, it is often difficult to fully follow the physical laws and physiological characteristics of the real world in terms of spatial texture, dynamic changes and three-dimensional structure. This inherent defect leads to the inconsistency of time and space in the video, and this characteristic can be the key basis for identifying authenticity and fake content.

Based on the space-time consistency framework, this paper systematically sorts out the research progress of deep forgery detection technology. First, an in-depth analysis of the identification role of space-time characteristics in identifying fake videos. Then, from the three dimensions of time modeling, the combination of spatial characteristics and attention mechanism, and space-time joint modeling, the representative methods are classified and compared. At the same time, summarize the commonly used benchmark data sets and evaluation indicators to provide support for experimental verification. Finally, for the main challenges faced by the current method, including cross-domain generalization ability, multimodal modeling and computational complexity, this paper carries out in-depth discussions and looks forward to future research direction.

Comprehensive analysis shows that the single-dimensional detection method has obvious limitations in complex reality scenarios. In contrast, the joint modeling strategy of integrating space-time information shows stronger robustness and generalization capabilities. By integrating physical and physiological a priori knowledge, multimodal clues and efficient model architecture, the space-based consistency detection method has significant potential to improve the reliability and deployability of practical applications, thus providing more effective technical support for in-depth forgery content governance and video forensics.

References

1. W. Yao, P. Li, Y. Zhao, H. Wu, Review of research on face deepfake detection methods. J. Image Graph. 30, 2343–2363 (2025)
2. Y. Chen, N. Akhtar, N.A.H. Haldar, S. Mian, Deepfake detection with spatio-temporal consistency and attention, in Proceedings of the 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (IEEE, 2022), pp. 1–8
3. T. Amada, K. Kakizaki, T. Miyagawa, Y. Aoki, Robust deepfake detection for electronic know your customer systems using registered images, in Proceedings of the 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG) (IEEE, 2025), pp. 1–10
4. A. Ciamarra, R. Caldelli, A. Del Bimbo, Temporal surface frame anomalies for deepfake video detection, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), pp. 3837–3844
5. Z. Wang, Z. Yu, C. Zhao, Y. Qin, Deep spatial gradient and temporal depth learning for face anti-spoofing, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), pp. 5042–5051
6. Y. Zheng, J. Bao, D. Chen, F. Wen, Exploring temporal coherence for more general video face forgery detection, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021), pp. 15044–15054
7. Y. Luo, Y. Zhang, J. Yan, W. Liu, Generalizing face forgery detection with high-frequency features, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), pp. 16317–16326
8. G. Jia, M. Zheng, C. Hu, X. Sun, Inconsistency-aware wavelet dual-branch network for face forgery detection. IEEE Trans. Biom. Behav. Identity Sci. 3, 308–319 (2021)
9. J.F. Yeh, M.J. Shih, J.X. Yang, Y.H. Chang, Utilizing recurrent neural network for deepfake video detection, in Proceedings of the 2024 IEEE 6th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS) (IEEE, 2024), pp. 496–498

10. Y. Li, M.C. Chang, S. Lyu, In ictu oculi: Exposing AI-generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)
11. Z. Gu, Y. Chen, T. Yao, S. Ding, Delving into the local: Dynamic inconsistency learning for deepfake video detection, in Proceedings of the AAAI Conference on Artificial Intelligence 36, 744–752 (2022)
12. H. Fan, Y. Yang, M. Kankanhalli, Point 4D transformer networks for spatio-temporal modeling in point cloud videos, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), pp. 14204–14213
13. S.Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, CNN-generated images are surprisingly easy to spot… for now, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), pp. 8695–8704
14. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, G. Thies, M. Nießner, FaceForensics++: Learning to detect manipulated facial images, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), pp. 1–11
15. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The DeepFake Detection Challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397 (2020)
16. A. Haliassos, R. Mira, S. Petridis, M. Pantic, Leveraging real talking faces via self-supervision for robust forgery detection, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), pp. 14950–14962
17. Z. Yan, Y. Luo, S. Lyu, B. Li, Transcending forgery specificity with latent space augmentation for generalizable deepfake detection, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), pp. 8984–8994