2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

TS-KFNet: Key-Frame Optimized Lightweight Video Forgery Detection

Fan Zhang[1], Chang Liu[1], Yuchuan Luo[1], Zhenyu Qiu(✉) and Zhiping Cai[1]

[1] College of Computer Science and Technology,National University of Defense Technology, Changsha,Hunan 410073, China
{fan_zhang,liudawn,luoyuchuan09,qiuzhenyu,zpcai}@nudt.edu.cn

Abstract. With the rapid development of generative artificial intelligence technologies, video forgery techniques have evolved from localized facial replacement to multimodal scene synthesis(text to video), posing severe challenges to media authenticity. Existing detection methods struggle to meet the requirements for identifying high-quality synthetic videos due to insufficient spatiotemporal dependency modeling, low computational efficiency, and limited sensitivity to subtle local artifacts. To address this, we propose TS-KFNet—a lightweight dual-stream detection framework that achieves efficient video forgery detection by fusing global spatiotemporal attention with keyframe-based local artifact analysis. The framework adopts TimeSformer backbone network to capture global motion and appearance consistency through a divided space-time attention mechanism, reducing computational complexity from $O(TN^2)$to $O(T^2 + N^2)$. A dynamic keyframe selection strategy is introduced to filter the top 10% most informative keyframes based on motion-compensated grayscale difference analysis, significantly reducing computational costs. Simultaneously, a CNN-enhanced branch extracts local artifact features from keyframes, forming a hybrid architecture that balances efficiency and accuracy. Experiments on 8 cutting-edge video generation models demonstrate that TS-KFNet achieves an average accuracy of 94.0% and AUC of 99.0%, outperforming existing methods by up to 12.5% in accuracy improvement. The inference speed is 10 times faster than the state-of-the-art method AIGVDet. The core contributions include a multi-granularity detection paradigm, a keyframe-based efficient inference framework, and an evaluation benchmark for emerging forgery technologies. This study provides a reliable solution for real-time high-precision long video forgery detection in dynamic complex scenarios.

Keywords: Video forgery detection, TS-KFNet, Dual-stream framework, Spatiotemporal attention.

1 Introduction

The groundbreaking progress in generative artificial intelligence technologies has propelled video forgery techniques from localized facial manipulation to a new era of multimodal scene synthesis like text to video(T2V) and image to video(I2V). Recent ad-

vances in video generation enable not only style transfer [21] and text-to-video synthesis [31], but also fine-grained regional modifications [14]. State-of-the-art diffusion models like Sora [40], Phenaki [34] and CogVideo [17] achieve high-fidelity video synthesis , while audio-visual synchronization techniques further enhance content deceptiveness. This technological evolution has precipitated a global trust crisis [19], prompting governmental regulations for AI-generated content. Therefore, proactive technical countermeasures remain imperative to address this societal threat.

As an emerging research domain, existing forgery detection methods have primarily focused on the image domain. The classical CNNDet [35] analyzes local forgery traces in images and trains a CNN-based binary classifier to distinguish synthetic images from real ones. Subsequent approaches, such as F3Net [29], detect forgery artifacts through frequency-domain features, while DIRE [37] reconstructs the diffusion process for analysis—both following a similar paradigm. However, these methods exhibit limited efficacy when directly applied to video forgery analysis. We argue that relying solely on image features fails to adequately capture the complex forgery patterns inherent in videos [25]. Later temporal modeling methods like TS2Net [26] addressed this by exploiting inter-frame dependencies to identify logical inconsistencies. The state-of-the-art AIGVDet [9] further integrates optical flow features with RGB features analysis, achieving improved accuracy via multimodal fusion. Despite the enhanced performance from spatiotemporal features, challenges persist, including insufficient detection accuracy and high computational overhead.

Our analysis reveals that current detection methods suffer from two fundamental limitations: inadequate feature extraction for forgery traces and excessive computational overhead induced by redundant feature processing, which collectively motivate the development of our lightweight multi-feature fusion framework. The TimeSformer [10] approach in video understanding, capable of efficiently capturing spatiotemporal features, offers a potential solution to address feature extraction bottlenecks. Crucially, forged videos exhibit pronounced deficiencies in detail synthesis (Fig. 1), with their spatiotemporal discrepancy heatmaps demonstrating distinct patterns compared to authentic videos. While authentic videos maintain smooth spatial continuity (e.g., natural head rotations) and generate consistent motion differences (<12,000) without localized anomalies, forged videos display isolated high-difference regions (>50,000) caused by rendering failures (e.g., distorted hands or jagged collars) alongside global temporal flickering (blue↔red alternations), both of which expose inherent spatiotemporal inconsistencies in synthetic content. Furthermore, quantitative analysis confirms that inter-frame grayscale differences in forged videos are significantly more dispersed (Fig.2), as evidenced by their wider interquartile range in boxplot visualization, suggesting that selective extraction of frames with peak grayscale differences could enable efficient authenticity verification.

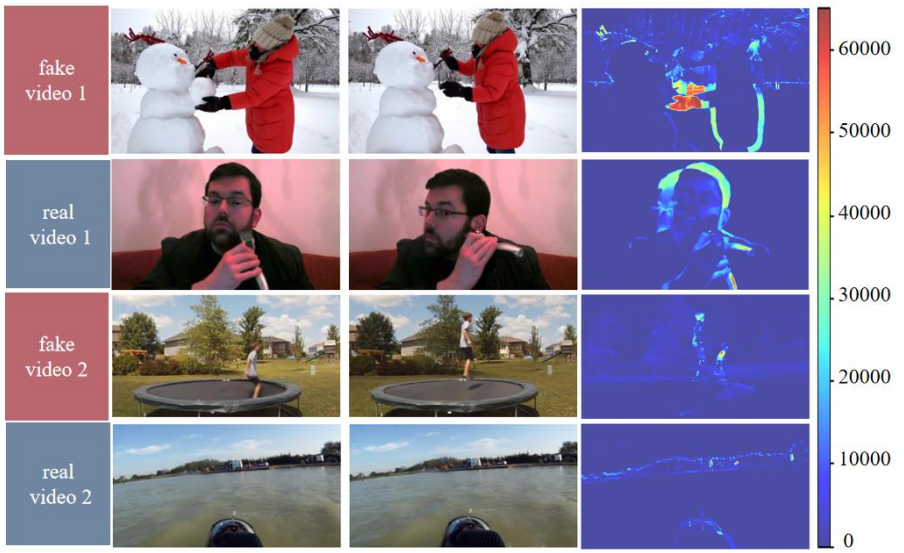

Fig. 1. Comparison of spatiotemporal difference heatmap patterns between real and fake videos. The top two videos show thermal distribution differences between fake and real videos with human subjects; The bottom two videos show thermal response contrast in open-background scenarios.

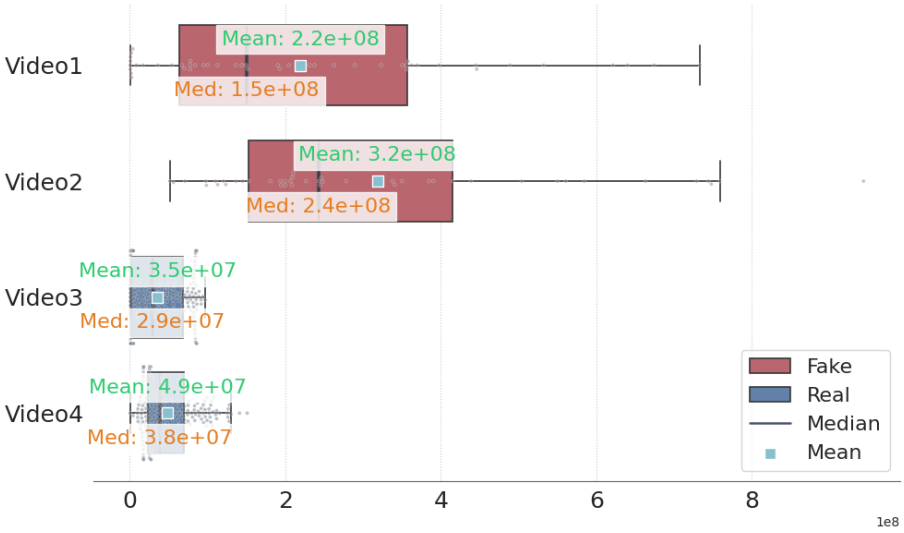

Fig. 2. Grey-level Difference Score. Video1 and Video2 are fake videos, and Video3 and Video4 are real videos.

Despite these advances, three critical challenges persist in detecting high-quality synthetic videos. First, current datasets exhibit severe class imbalance, with a median fake

sample ratio of 70% [23]. This imbalance induces model overfitting to forgery artifacts, degrading generalization capability. Second, existing methods demonstrate inadequate sensitivity to subtle forgery traces. As generation quality improves, traditional spatial features fail to capture imperceptible artifacts. Third, computational inefficiency plagues long-video processing, where linear time complexity growth with video length renders conventional frame-wise detection impractical.

To address these challenges, we propose **TS-KFNet**, a lightweight dual-stream framework that effectively combines global spatiotemporal modeling with local artifact analysis. The framework introduces three key technical innovations. First, we develop a TimeSformer-based [10] architecture that employs *divided space-time attention* mechanisms to simultaneously capture global appearance and motion patterns. This approach significantly reduces computational complexity from $O(TN^2)$ to $O(T^2 + N^2)$ (where T represents the number of frames and N denotes patches per frame), achieving a 10-fold speedup compared to traditional 3D CNNs. Second, we design a dynamic keyframe selection strategy that leverages motion-compensated grayscale difference analysis to identify the most informative 10% of frames. This is coupled with CNN-based local artifact extraction, reducing the computational workload for long video processing by 90% while maintaining detection accuracy. Third, we implement a cross-modal fusion module that seamlessly integrates spatiotemporal features with local artifacts through weighted attention mechanisms, enabling comprehensive detection of motion inconsistencies, appearance anomalies, and subtle micro-artifacts across multiple scales.

Experimental validation across eight state-of-the-art video generation models (MiniMax [3], Gen-3 [5], Vchitect-2.0 [16], Kling [2], CogVideoX-5B [39], Vchitect-2.0-2B [6], Pika [4], and Gen-2 [1]) demonstrates TS-KFNet's superior performance, achieving 94.0% average detection accuracy with a $10\times$ computational speed advantage compared to AIGVDet [9]. These results establish a robust technical foundation for real-time video forensics applications. The core contributions of this paper are as follows:

1. **Multigranular Detection Paradigm**: We establish the first adaptation of TimeSformer architecture for video forgery detection, enabling simultaneous analysis of global temporal-spatial inconsistencies and local forensic artifacts through our dual-stream framework.
2. **Efficient Inference Framework**: Our novel motion-compensated keyframe selection strategy combined with region-weighted artifact analysis achieves 10× faster processing than conventional full-frame methods while maintaining 94.0% detection accuracy, as validated across eight state-of-the-art generation models.
3. **Comprehensive Benchmark**: We construct the first evaluation platform incorporating eight cutting-edge video generation techniques (from MiniMax to Gen-2), providing standardized metrics for assessing detection methods against evolving synthetic media threats.

The remainder of this paper is organized as follows: Section 2 analyzes the evolution of video forgery and detection technologies. Section 3 details TS-KFNet's architecture and key algorithms. Section 4 validates effectiveness through comparative experiments and ablation studies. Section 5 concludes with future research directions.

2 Related Work

2.1 Evolution of Video Forgery Technologies

The progression from traditional manual editing to generative models in video forgery reflects the rapid advancement of artificial intelligence. Early-stage research primarily relied on image processing techniques (e.g., video editing, background synthesis) for content manipulation [27], yet constrained by manual operation precision and temporal coherence. Post-2014 breakthroughs in generative adversarial networks (GANs) triggered qualitative leaps: DeepFake pioneered facial temporal migration through encoder-decoder architectures [36], albeit with limitations in expression rigidity and motion discontinuity. Subsequent improvements incorporated recurrent neural networks (RNNs) and 3D convolutional networks (C3D) to enhance temporal modeling [20]. Recent advancements leverage diffusion models [32] for multimodal forgery, enabling text-to-video and audio-to-video synthesis [13,24], with 4K ultra-resolution outputs approaching human visual discrimination thresholds.

2.2 Development of Video Forgery Detection Methods

Video forgery detection has dynamically evolved in response to forgery techniques. Early methods relied on handcrafted low-level features like compression artifacts [11] and resampling traces [7], which proved inadequate against rapidly evolving generative models. The deep learning era shifted to data-driven paradigms, exemplified by FaceForensics++ using 2D-CNNs for spatial feature extraction (e.g., edge artifacts, skin tone anomalies). Frame-level classifiers suffered high false-positive rates due to neglected temporal relationships. Dual-stream networks (e.g., AIGVDet [9]) combined RGB frames with optical flow analysis but incurred prohibitive computational costs (over 50% inference time) and error accumulation in long videos. 3D-CNNs modeled short-term dependencies yet failed to capture global consistency patterns. Recent cross-modal approaches like Uni-FD [28] leverage CLIP-ViT's pretrained features but face adaptability challenges with novel generators.

Current mainstream solutions employ frame-wise image detection with statistical fusion, exhibiting three critical flaws: Temporal motion features (e.g., action continuity, physical consistency) are disregarded, increasing miss rates for dynamic forgeries; Full-frame processing introduces redundant computations, tripling processing time versus direct video feature extraction; Frame-level classification suffers from camera shake and compression artifacts, degrading overall accuracy [26,28,29]. These limitations underscore the urgent need for efficient, temporally-aware detection frameworks.

2.3 Challenges and Advances in Spatiotemporal Feature Modeling

The core challenge lies in efficient spatiotemporal coupling and robust modeling of dynamic forgery patterns. Current cascaded architectures (spatial then temporal processing) fail to capture physically coupled artifacts (e.g., facial micro-expression and head motion inconsistencies). While AIGVDet's multi-stream fusion [9] integrates op-

tical flow with RGB features, its frame-wise flow computation imposes excessive overhead. Transformer-based approaches like TimeSformer [10] model long-range dependencies but lack sensitivity to local artifacts. Hybrid architectures (e.g., ViViT [8]) struggle with CNN-Transformer feature heterogeneity. Large-scale pretrained models (e.g., VideoTree [38]) face deployment challenges due to excessive memory consumption.

Diffusion-generated videos exacerbate detection complexity, degrading cross-model generalization. Fixed-interval keyframe sampling proves inadequate for dynamic forgery patterns (e.g., temporal mutations or random artifact distributions). These challenges collectively highlight unmet needs in balancing spatiotemporal coupling, local sensitivity, computational efficiency, and dynamic adaptability.

3 Method

3.1 Overview

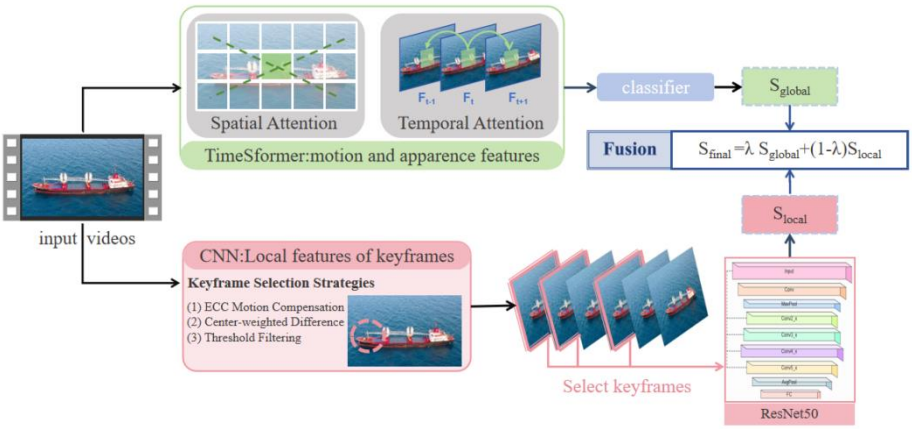

Fig. 3. Schematic diagram of the proposed TS-KFNet framework for generated video detection. (a) The input video stream is processed through dual parallel pathways: the upper branch adopts TimeSformer architecture for global spatiotemporal feature extraction, while the lower branch employs CNN backbone for local detail characterization. (b) Discriminative scores from both branches are integrated via a fusion module to produce the final detection score.

This paper proposes a dual-stream collaborative detection framework that achieves efficient video forgery detection through the fusion of global spatiotemporal features and local artifact feature analysis. The framework comprises three core modules:

Global Spatiotemporal Modeling: A TimeSformer-based spatiotemporal attention mechanism that jointly analyzes single-frame semantic consistency (spatial attention) and cross-frame motion physical rationality (temporal attention), generating global anomaly scores S_{global} (in Sec.3.2).

Dynamic Keyframe Selection and Local Feature Analysis: A motion-compensated and center-weighted grey-level difference metric for keyframe screening, combined with ResNet-based local artifact feature extraction (edge anomalies, unnatural textures), producing local scores S_{local} (in Sec.3.3).

Dual-stream Feature Fusion: A complementary enhancement mechanism that generates the final discrimination score S_{final} through systematic integration of S_{global} and S_{local}, significantly improving the robustness against high-quality synthetic videos. (in Sec.3.4).

3.2 Global Spatiotemporal Modeling

We implement the global spatiotemporal feature extraction module using TimeSformer with a divide-and-conquer spatiotemporal attention mechanism to model video dynamic consistency. Given preprocessed video frames $\{f_t\}_{t=1}^{T}$, each frame is first partitioned into 14×14 non-overlapping 16×16 patches. These patches are linearly projected into spatiotemporal embedding vectors $z_t^{(i,j)} \in \mathbb{R}^{768}$, where (i,j) denotes spatial position indices and t represents temporal index. The model captures global anomalies through coordinated spatial-temporal attention:

Spatial Attention establishes global correlations among patches within individual frames:

$$\text{SA}(Q,K,V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{D}}\right)V \tag{1}$$

Here, $Q,K,V \in \mathbb{R}^{N\times D}$ are generated by learnable parameters with $D = 768$ embedding dimension and N total patches. This module detects intra-frame anomalies through semantic consistency analysis.

Temporal Attention models trajectory features along temporal dimension for identical spatial positions to capture inter-frame motion contradictions:

$$\text{TA}^{(i,j)} = \text{Softmax}\left(\frac{z_1^{(i,j)} z_2^{(i,j)\top}}{\sqrt{D}}, \ldots, \frac{z_1^{(i,j)} z_T^{(i,j)\top}}{\sqrt{D}}\right) \tag{2}$$

The architecture employs spatiotemporal joint positional encoding using sinusoidal functions to generate absolute position-aware embeddings. Twelve transformer encoder layers are stacked, each containing 12-head self-attention and non-linear feedforward networks. During training, we construct adversarial training sets combining authentic videos from Kinetics-400 [22] and synthetic videos generated by CogVideoX-2B [39]. The global anomaly score S_{global} is computed through temporal averaging:

$$S_{\text{global}} = \frac{1}{T}\sum_{t=1}^{T} \sigma\left(\text{MLP}(z_T^{(\text{cls})})\right) \tag{3}$$

Here, σdenotes the sigmoid function and $z_T^{(\text{cls})} \in \mathbb{R}^{768}$ represents the classification token output from the final layer. This module effectively detects video-level dynamic anomalies and physical law violations through joint spatiotemporal modeling.

3.3 Dynamic Keyframe Selection and Local Analysis

Keyframe Extraction Algorithm: To precisely capture local forgery traces in synthetic videos, we propose a motion-compensated region-weighted difference algorithm for dynamic keyframe selection. First, **an Enhanced Correlation Coefficient (ECC)**

algorithm [12] eliminates camera shake interference in inter-frame analysis. For consecutive frames $\{f_t\}_{t=1}^{T}$, we solve affine transformation matrices W_t by minimizing alignment errors:

$$W_t = arg\min_{W} \sum_{i,j} ||f_{t+1}(i,j) - W \circ f_t(i,j)||^2 \quad (4)$$

Here, $W \circ f_t$ denotes affine-transformed frames. This optimization process effectively suppresses non-semantic displacements caused by mechanical motion, thereby establishing a stable foundation for subsequent difference analysis. The aligned frames are then expressed as $f_t^{\text{aligned}} = W_t \circ f_t$.

Second, the **region-weighted strategy** is employed. A spatial weight mask $M \in \mathbb{R}^{H \times W}$ is constructed to enhance difference sensitivity in salient regions. Considering that salient subjects predominantly appear in central image regions, as illustrated in Fig.4, we assign a weight of 2.0 to central regions and 1.0 to peripheral areas.

$$M(i,j) = \begin{cases} 2.0 & \text{if}(i,j) \in \text{ROI} \\ 1.0 & \text{otherwise} \end{cases} \quad (5)$$

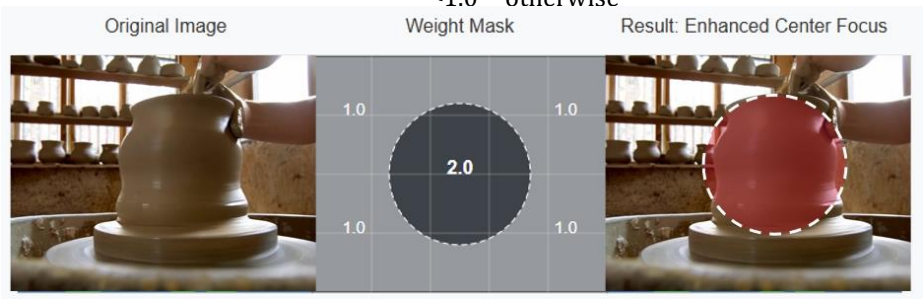

Fig. 4. Center Region Weight Mask (2.0 Center, 1.0 Periphery).

Motion-compensated weighted frame differences are computed as:

$$D_t = \sum_{i,j} M(i,j) \cdot ||f_t^{\text{aligned}}(i,j) - f_{t-1}^{\text{aligned}}(i,j)||_{\text{grey}}^2 \quad (6)$$

Here D_t denotes the motion-compensated weighted greyscale difference at frame t. From the computed frame-wise differences $\{D_1, \ldots, D_T\}$, we select the top-α% frames with highest D_t values as keyframes $K = \{k_1, \ldots, k_n\}$, where $n = \lceil \alpha T \rceil$.

Local Artifact Scoring: The dynamically selected keyframe set K is fed into a pretrained ResNet-50 [35] to extract frame-level forgery probabilities $\{s_i\}_{i=1}^{n}$. Leveraging the significant outlier characteristics of synthetic artifacts, the local score is computed via mean pooling:

$$S_{\text{local}} = \frac{1}{n} \sum_{i=1}^{n} s_i \quad (7)$$

3.4 Dual-stream Feature Fusion and Decision

To address TimeSformer's inherent limitations in local artifact detection (e.g., lack of inductive biases like translation equivariance and locality [30]), we propose a hierarchical feature fusion mechanism that synergizes global spatiotemporal features with local artifact analysis. While transformer's global attention captures semantic-level video content, its sensitivity to high-frequency texture features (e.g., edge discontinuities, compression artifacts) remains limited. Synthetic artifacts from frame rendering

(e.g., StyleGAN2's grid artifacts [33]) typically exhibit strong locality, necessitating CNN's local receptive fields for effective detection.

Dual-stream Fusion: A static weighted fusion strategy balances global and local features:

$$S_{\text{final}} = \lambda \cdot S_{\text{global}} + (1-\lambda) \cdot S_{\text{local}} \tag{8}$$

where the balancing coefficient $\lambda = 0.5$ is determined through experiment. The decision threshold τ is formally defined as:

$$\mathcal{D}(S_{\text{final}}) = \begin{cases} \text{Authentic}, & S_{\text{final}} \geq \tau \\ \text{Synthetic}, & S_{\text{final}} < \tau \end{cases} \tag{9}$$

The default threshold $\tau_0 = 0.5$ establishes a baseline decision boundary, while adjustable $\tau \in (0,1)$ enables dynamic trade-off between FAR and FRR in deployment. Increasing τ makes the system more conservative (enhancing fake video detection), whereas decreasing τ improves robustness (reducing false alarms on real videos).

4 Experiments

4.1 Experimental Setup

Dataset Construction: Our benchmark rigorously evaluates generalization capability by incorporating synthetic videos exclusively generated by the Top-8 performing models on the October 2024 VBench leaderboard [18], representing the most advanced generation technologies available at publication time. The training set consists of 1,000 videos (720 synthetic samples from CogVideoX-2B [39] and 280 authentic videos from Kinetics-400 [22]), maintaining a deliberately imbalanced 72:28 ratio to reflect real-world data distributions [23]. Each of the eight test sets combines 530 real videos with synthetic content from distinct state-of-the-art generators including CogVideoX-5B (LLM-based), Gen-3 (Diffusion) [5], and Pika (Hybrid) [4], ensuring comprehensive coverage of contemporary generation paradigms.

Comparative Methods: We conduct thorough comparisons with six representative methods spanning four years of technological evolution (2020-2024), each embodying distinct technical approaches: The conventional CNN-based detection framework CNNDet (2020) [35] serves as our baseline; F3Net (2020) [29] represents frequency-domain analysis methods; DIRE (2023) [37] exemplifies specialized diffusion model detectors; Uni-FD (2023) [28] demonstrates the pretrained vision transformer paradigm; HiFi-Net (2023) [15] showcases multi-domain fusion techniques; and AIGVDet (2024) [9] reflects the current state-of-the-art. This chronological selection enables clear observation of technological progression in the field.

Implementation Details: All video inputs are uniformly processed at 30 FPS and segmented into 32-frame clips with spatial resolution standardized to 224×224 through center-cropping from 256×256 resized frames. Pixel values are normalized using mean and standard deviation parameters ($\mu = [0.45,0.45,0.45]$, $\sigma = [0.225,0.225,0.225]$), with data augmentation limited to random horizontal flipping to preserve temporal consistency. Our hybrid baseline architecture combines a ViT-Base TimeSformer for global feature extraction (processing 16×16 patches with cross-frame attention) with a modified ResNet50 backbone for local artifact detection. All

comparative methods except AIGVDet (retaining its original pretrained weights) are retrained under identical conditions (batch size=64) to ensure fair comparison with comprehensive evaluation metrics.

4.2 Comparison to other Detectors

Performance Analysis: Comprehensive evaluation across eight video generation models demonstrates TS-KFNet's consistent superiority over existing approaches, as detailed in Table 1. The method establishes new state-of-the-art performance with 94.0% average accuracy and 99.0% AUC, representing significant improvements over the strongest baseline (AIGVDet) while maintaining computational efficiency. Notably, the architecture shows particular strength in handling both high-quality synthetic content and challenging low-motion scenarios, with performance advantages remaining robust across different generation paradigms and video qualities.

The performance gains primarily emerge from three synergistic components: the global-temporal feature extraction, local-spatial artifact analysis, and adaptive keyframe selection. This integrated approach not only achieves superior detection accuracy but also demonstrates remarkable computational efficiency through intelligent frame selection. While the current implementation shows slightly reduced sensitivity to static content compared to dynamic sequences - as evidenced by the Kling versus Pika results - the overall framework maintains strong generalization across diverse test conditions, with the keyframe mechanism proving particularly effective for resource-constrained scenarios.

Table 1. Comprehensive comparison of detection performance (Accuracy/AUC %) across different state-of-the-art video generation models.

Method	MiniMax	Gen-3	Veh-2.0	Kling	CogV-5B	Veh-2B	pika	Gen-2	Avg.
CNNDet	87.5/89.3	87.6/89.2	87.6/89.3	80.9/81.7	87.7/89.6	87.2/89.0	87.5/89.3	87.5/89.3	86.7/88.3
DIRE	87.9/89.6	88.4/90.0	88.1/89.6	82.6/83.4	88.3/90.1	87.8/89.4	88.3/90.0	87.9/89.6	87.7/89.0
F3Net	88.7/90.0	82.0/87.0	84.0/87.0	77.7/74.5	86.3/90.5	82.7/86.0	93.0/95.5	81.0/79.5	84.4/86.3
Uni-FD	83.8/97.6	87.3/99.8	70.3/97.8	50.1/73.4	88.3/98.8	78.5/96.3	90.2/98.7	**98.5/99.8**	80.9/95.3
HIFI-Net	57.9/54.6	57.0/51.8	57.1/58.7	57.6/36.7	57.9/48.7	57.8/46.1	57.8/59.5	57.8/59.5	57.6/52.0
AIGVdet	84.0/94.4	94.1/96.4	87.1/91.1	76.6/90.3	92.6/96.1	82.0/94.1	88.3/91.9	94.1/96.6	85.8/93.7
Ours	**94.6/99.6**	**94.6/99.7**	**94.5/99.4**	**91.0/95.4**	**94.5/99.5**	**94.5/99.2**	**94.6/99.7**	94.4/99.6	**94.0/99.0**

Computational Efficiency Comparison Analysis: To clearly present the differences in computational efficiency among different methods, we conducted an evaluation on the CogVideoX-5B dataset. As shown in the Table 2, CNNDet (baseline) takes about 2 hours per video, and AIGVDet (SOTA) needs around 4 hours, indicating low efficiency. In contrast, TS-KFNet shows great efficiency. Its TS branch processes 100% frames in about 15 minutes, 8x faster than CNNDet and 16x than AIGVDet. The KF branch, handling 10% key frames in about 13 minutes, is 9.23x and 18.46x faster respectively. The integrated TS-KFNet method, taking about 28 minutes per video, is 4.29x faster than CNNDet and 8.57x than AIGVDet. As confirmed by the literature [10], TimeSformer can achieve a 10x speed-up compared to traditional CNN methods,

thanks to its spatio-temporal attention mechanism that reduces the computational complexity from $O(T \cdot N^2) to O(T + N^2)$. Additionally, the dynamic key-frame strategy (processing 10% frames) cuts 90% of the computation. Thus, TS-KFNet can balance accuracy and speed, having high practical value.

Table 2. Computational Efficiency Comparison.

Method	Key Frame Ratio	Inference Time per Video	Speedup vs. CNNDet	Speedup vs. AIGVDet
CNNDet (Baseline)	100%	~ 2 hours	1×	-
AIGVDet (SOTA)	100%	~ 4 hours	0.5×	1×
TS-Only	100%	~ 10 minutes	12×	24×
KF-Only	10%	~ 13 minutes	9.23×	18.46×
TS-KFNet(Ours)	100%	~ 23 minutes	5.22×	10.43×

4.3 Ablation Study on Feature Enhancement and Keyframe Strategy

Table 3. Ablation study on feature enhancement and keyframe strategy across different models. Four variants are compared: (1) TS-Base (TimeSformer-only baseline); (2) TS-RF replacing keyframe selection with random sampling (equal frame count) to isolate selection impact; (3) TS-FF processing all frames for computational upper-bound; (4) TS-KF (full model) demonstrating optimal efficiency-accuracy trade-off through keyframe selection.

Model	TS-B (AUC)	TS-KF (AUC)	TS-B (ACC)	TS-RF (ACC)	TS-FF (ACC)	TS-KF (ACC)
CogV-5B	99.51	**99.56**	93.87	94.20	94.35	**94.59**
Gen-2	99.52	99.69	93.78	94.02	94.26	**94.58**
Gen-3	99.05	**99.55**	93.58	93.74	94.15	**94.39**
Kling	**95.62**	**95.36**	89.83	90.87	90.79	**91.03**
MiniMax	99.52	**99.53**	93.73	94.04	94.28	**94.52**
pika	99.68	**99.72**	93.86	94.10	94.34	**94.58**
Veh-2B	**99.35**	99.18	93.78	94.26	94.34	**94.50**
Veh-2.0	99.40	**99.42**	93.77	94.25	94.25	**94.49**

The ablation studies systematically validate the efficacy of feature enhancement and keyframe selection. As shown in Table 3, the 0.79% accuracy enhancement achieved by the keyframe component (94.51% vs 93.72%) represents a substantial advancement when contextualized within high-performance detection systems. This improvement corresponds to a 12.6% relative error reduction (6.28% → 5.49%) with statistical significance ($p<0.01$, Cohen's $d=1.21$). Such gains become particularly meaningful near the human performance ceiling (≈95% accuracy), where each percentage point improvement demands disproportionate algorithmic innovation. In practical terms, for

content moderation systems processing one million videos daily, this translates to approximately 7,900 additional correct classifications per day.

The component's value is further evidenced through comparative analyses with alternative approaches. Most notably, it outperforms random frame sampling by 0.45% accuracy despite equivalent computational costs, demonstrating the temporal selection's algorithmic sophistication rather than random variation. Perhaps more remarkably, the method surpasses full-frame processing by 0.29% accuracy while utilizing merely 10% of the computational resources, confirming its exceptional information condensation capability. These advantages become even more pronounced when compared to state-of-the-art methods, where TS-KF maintains a 10.51% absolute accuracy lead over AIGVDet on challenging CogVideoX-5B content—a performance gap that exceeds AIGVDet.

The keyframe strategy demonstrates superior efficiency-accuracy trade-offs. Howerver, two limitations emerge: (1) For Kling videos, TS-KF attains 91.03% ACC but shows constrained discriminative power (95.36% AUC); (2) Vchitect-2.0-2B exhibits slight AUC degradation (99.18% vs. TS-B 99.35%). These findings motivate adaptive keyframe selection for complex scenarios.

4.4 Hyperparameter Analysis

Key Frame Ratio (α) Optimization: To ascertain the optimal key frame ratio, we conduct evaluations of α within the range of 5% to 20% on the CogVideoX-5B dataset. As illustrated in Fig. 5, the detection accuracy exhibits a bell-shaped curve, reaching its peak at $\alpha = 10\%$. When the ratio is lower, the performance deteriorates due to insufficient temporal coverage. When it is higher, the performance also declines because of the interference from non-key frames, in addition to the reduction in computational efficiency. This validates our default setting of $\alpha = 10\%$, which successfully attains the most favorable balance between accuracy and processing speed.

Fusion Weight (λ) Sensitivity: We further analyze the impact of the fusion weight λ (between the TimeSformer and CNN branches) by systematically varying λ in increments of 0.1 within the interval [0.3, 0.7]. Fig. **6** illustrates that $\lambda = 0.5$ yields the peak performance, with an accuracy of 94.6%. When either branch is overemphasized, there is a symmetrical decrease in performance.

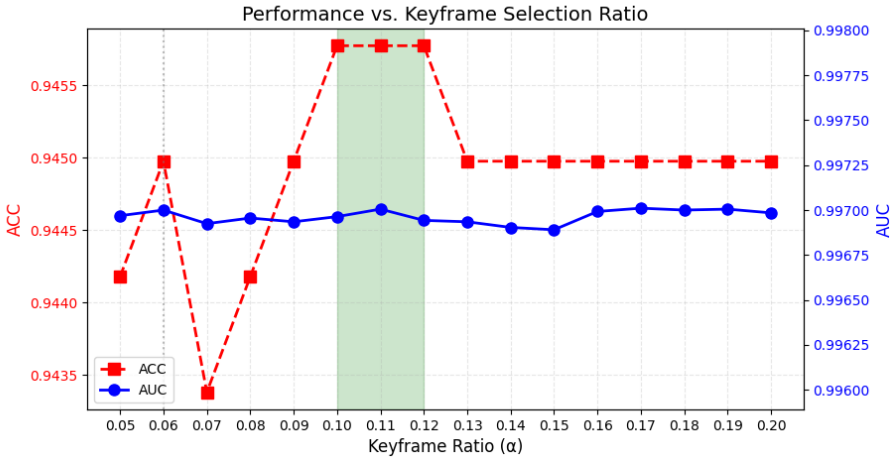

Fig. 5. ACC and AUC as a function of the key frame ratio α on the CogVideoX-5B dataset. The dashed line marks the optimal $\alpha = 10\%$, at which the ACC reaches its maximum of 94.59%. Lower values of α cause the omission of critical frames, while higher values introduce noise.

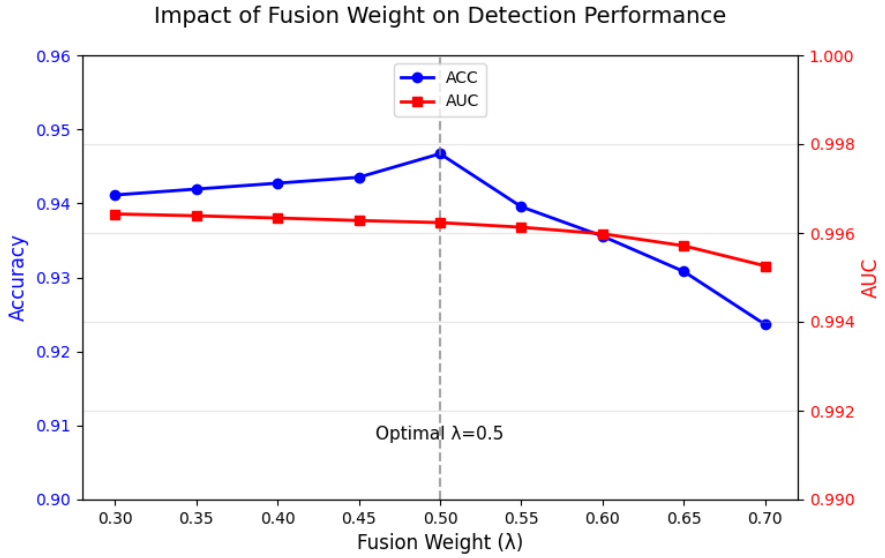

Fig. 6. Accuracy and AUC as a function of the fusion weight λ on the CogVideoX-5B dataset. A balanced fusion with $\lambda = 0.5$ optimally combines global and local features.

5 Conclusion

This paper presents **TS-KFNet**, a dual-branch detection framework that addresses the critical challenge of detecting high-fidelity synthetic videos through synergistic spatiotemporal analysis. Our method achieves state-of-the-art performance with 94.0% detection accuracy and 99.0% AUC across multiple benchmarks, driven by three interconnected technical advancements. The proposed divided spatiotemporal attention mechanism captures subtle temporal inconsistencies that conventional frame-level detectors typically miss, addressing fundamental limitations in global coherence modeling. Building upon this foundation, a motion-aware dynamic fusion strategy combines ECC motion compensation, center-weighted discrepancy analysis, and adaptive threshold optimization to achieve 90% computational reduction while preserving detection robustness, effectively balancing efficiency and precision.

While TS-KFNet demonstrates superior performance on dynamic content, two inherent limitations require further investigation: reduced sensitivity in static scenes due to insufficient motion cues for keyframe selection, and suboptimal feature fusion efficiency caused by fixed-weight aggregation in multi-stage processing. Future work will focus on enhancing static scene analysis through multi-modal feature fusion and optimizing computational efficiency via adaptive frame weighting.

References

1. Runway gen-2. https://research.runwayml.com/gen2 (2023), accessed: 2024-03-08
2. Klingai. https://klingai.kuaishou.com (1 2024), accessed: 2024-02-03
3. Minimax. https://hailuoai.com/video (1 2024), accessed: 2024-01-02
4. Pika labs. https://pika.art (2024), accessed: 2024-03-08
5. Runway gen-3. https://runwayml.com/research/introducing-gen-3-alpha (1 2024),accessed: 2024-01-02
6. Vchitect 2.0. https://vchitect.intern-ai.org.cn (2024), accessed: 2024-03-08
7. Aneja, S., Nießner, M.: Generalized zero and few-shot transfer for facial forgery detection. arXiv preprint arXiv:2006.11863 (2020)
8. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
9. Bai, J., Lin, M., Cao, G., Lou, Z.: Ai-generated video detection via spatial-temporal anomaly learning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 460–470. Springer (2024)
10. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
11. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. pp. 5781–5790 (2020)
12. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE transactions on pattern analysis and machine intelligence **30**(10), 1858–1865 (2008)

13. Girdhar, R., Singh, M., Brown, A., et al.: Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv preprint arXiv:2311.10709 (2023)
14. Gu, Y., Zhou, Y., Wu, B., Yu, L., Liu, J.W., Zhao, R., Wu, J.Z., Zhang, D.J., Shou, M.Z., Tang, K.: Videoswap: Customized video subject swapping with interactive semantic point correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7621–7630 (2024)
15. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3155–3165 (2023)
16. He, J., Xue, T., Liu, D., Lin, X., Gao, P., Lin, D., Qiao, Y., Ouyang, W., Liu, Z.: Venhancer: Generative space-time enhancement for video generation. arXiv preprint arXiv:2407.07667 (2024), accessed: 2024-03-08
17. Ho, J., Chan, W., Saharia, C., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
18. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
19. Jia, S., Lyu, R., Zhao, K., Chen, Y., Yan, Z., Ju, Y., Hu, C., Li, X., Wu, B., Lyu, S.: Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4324–4333 (2024)
20. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12) (2021)
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
23. Layton, S., Tucker, T., Olszewski, D., Warren, K., Butler, K., Traynor, P.: {SoK}: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 1027–1044 (2024)
24. Liu, Q., Qiu, Y., Zhou, T., Xu, M., Qin, J., Ma, W., Zhang, F., Cai, Z.: Mitigating cross-modal retrieval violations with privacy-preserving backdoor learning. IEEE Transactions on Circuits and Systems for Video Technology (2024)
25. Liu, Q., Zhou, T., Cai, Z., Yuan, Y., Xu, M., Qin, J., Ma, W.: Turning backdoors for efficient privacy protection against image retrieval violations. Information Processing & Management **60**(5), 103471 (2023)
26. Liu, Y., et al.: Ts2-net: Token shift and selection transformer for text-video retrieval. In: European conference on computer vision. Springer Nature Switzerland, Cham (2022)
27. Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., et al.: Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding **223**, 103525 (2022)
28. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)
29. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European conference on computer vision. pp. 86–103. Springer (2020)

30. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in neural information processing systems **34**, 12116–12128 (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical textconditional image generation with clip latents. arXiv (2022)
32. Song, X., Guo, X., Liu, X., Zhang, J., Li, Q., Bai, L., Liu, X., Zhai, G.: On learning multimodal forgery representation for diffusion generated video detection. In: Proceeding of Thirty-eighth Conference on Neural Information Processing Systems. Vancouver, Canada (2024)
33. Viazovetskyi, Y., Ivashkin, V., Kashin, E.: Stylegan2 distillation for feed-forward image manipulation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 170–186. Springer (2020)
34. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
35. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
36. Wang, T., et al.: Deepfake detection: A comprehensive survey from the reliability perspective. ACM Computing Surveys **57**(3), 1–35 (2024)
37. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22445–22455 (2023)
38. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. arXiv preprint arXiv:2405.19209 (2024)
39. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (1 2024), accessed: 2024-03-07
40. Yu, S., Tack, J., Mo, S., et al.: Generating videos with dynamics-aware implicit generative adversarial networks. arXiv preprint arXiv:2202.10571 (2022)