Transactions on Computational and Scientific Methods | Vo. 5, No. 4, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Transformer-Based Structural Anomaly Detection for Video File Integrity Assessment

Da Xu
Worcester Polytechnic Institute, Worcester, USA
dxu2@wpi.edu

Abstract: With the widespread use of video data in surveillance, security, and communication, the integrity of video files has become a key factor affecting data usability and trustworthiness. Compared to semantic analysis of video content, structural anomalies — such as frame loss, container errors, timestamp disorder, and bitstream corruption—have a more direct impact on video parsing and usage. However, existing methods often rely on rule-based checks or content-driven learning strategies, making them less robust for complex structural anomalies. To address this, we propose a Transformer-based method for video structural integrity anomaly detection. It employs structural embeddings and hierarchical attention to build multi-scale representations of video files, effectively identifying hidden anomalies in containers, frame sequences, and encoding parameters. The method takes structural metadata sequences as input and performs global modeling and local refinement using a multi-stage Swin Transformer architecture, producing corresponding integrity scores. Experiments on video datasets show that our method outperforms mainstream models in terms of detection accuracy, recall, and anomaly recognition stability, while maintaining strong generalization and execution efficiency across encoding formats and under low-resource conditions. This study not only introduces a novel modeling approach for video integrity analysis but also provides critical technical support for quality control and security protection in multimedia systems.

Keywords: Video file structure; anomaly detection; Swin Transformer; integrity analysis

1. Introduction

With the rapid development of multimedia technology, video data has become a vital carrier in fields such as information transmission, entertainment, surveillance, and security. Due to its high information density and intuitive expression, video is widely used in key scenarios like security monitoring, forensic analysis, intelligent transportation, online education, and medical diagnosis. In this context, the integrity of video data is crucial not only for its validity and usability but also for the reliability and safety of the system. However, during transmission, storage, or transcoding, the video file structure is prone to damage or tampering. This may lead to decoding failures, playback errors, or even content loss. Therefore, efficient and accurate detection of video file structure integrity is a fundamental requirement for ensuring multimedia data quality [1].

Current mainstream research focuses primarily on video content analysis, such as action recognition, object detection, and scene change detection. In contrast, studies on the structural integrity of video files remain limited [2]. In practical applications, structural issues often appear as metadata corruption, frame

sequence breaks, loss of encoded frames, or container format anomalies. These problems may not directly affect semantic understanding of the video content but can cause playback failures or parsing errors. In severe cases, the video information may become irrecoverable. Traditional integrity detection methods rely on parser errors, file format checks, or hard-coded rules. These methods struggle to handle complex and diverse structural anomalies, especially when facing unknown or non-standard video files. Their robustness and generalization capability are often insufficient .

Recently, Transformer models have shown remarkable performance in natural language processing, image recognition, and sequence modeling [3-5]. Unlike traditional convolutional networks, Transformers do not rely heavily on spatial locality. Instead, they use attention mechanisms to model global feature interactions flexibly. This makes them suitable for modeling high-dimensional sequential data like video file structures [6]. Logically, a video file can be viewed as a sequence of ordered data segments, including container structures, encoded frame information, timestamps, and bitstream blocks. These elements show strong sequential features and structural dependencies. Based on this, we propose a Transformer-based method for detecting structural integrity anomalies in video files. The goal is to discover deep patterns within the internal sequence structure and identify hidden anomalies or corruption.

This research holds both theoretical and practical significance. Theoretically, it extends the application of Transformers to non-traditional sequence tasks and explores their potential in modeling multimodal data structures. Practically, the method can be applied to tasks such as video data validation, anti-tampering detection, forensic evidence verification, and quality control in content delivery networks. It improves system response and handling of abnormal video files, enhancing overall data reliability and stability. With the Transformer's strong feature learning capabilities, the method may also achieve adaptive detection of abnormal patterns. This reduces reliance on manual rules and increases system intelligence.

In summary, this study on Transformer-based video structural integrity anomaly detection responds to the urgent need for trustworthy multimedia data. It also provides a new technical path for file-level quality control. By deeply mining the hidden patterns within video structures and building adaptive detection models, we aim to improve robustness and automation in video processing systems under complex conditions. This lays a solid foundation for secure and reliable multimedia information systems.

2. Related work

Research on video integrity detection mainly focuses on two directions: content consistency analysis and structural validity verification. The former relies on image processing or deep learning methods. It uses visual features and temporal consistency between frames to detect content-level tampering, such as frame insertion, replacement, or duplication. However, such methods often ignore the integrity of the underlying video file structure. They fail to handle issues like metadata corruption, abnormal container formatting, or missing bitstream frames. The latter focuses on verifying the video container format and structural compliance based on rules or protocol standards. Tools like FFmpeg are used for format parsing [7]. Yet, these methods show clear limitations when dealing with unknown or non-standard encoded videos.

With the widespread adoption of deep learning in multimodal data analysis, some studies have started applying neural networks to video structure anomaly detection. For example, some convert video metadata into sequential features. Models like recurrent neural networks (LSTM) [8] or graph neural networks (GNN) [9] are used to model structural dependencies and detect anomalies. These methods outperform traditional rule-based approaches in modeling capacity. However, their ability to capture long-range dependencies remains limited. When faced with high-dimensional and multi-level structural relationships, performance bottlenecks persist. Moreover, most existing methods rely heavily on

handcrafted features or specific encoding formats. This limits their generalization and adaptability in real-world scenarios.

In anomaly detection tasks, the Transformer architecture has gained attention for its strong global modeling ability and structure-independence [10]. It has shown remarkable success in non-visual sequence domains, such as log analysis [11] and protocol intrusion detection [12]. However, applying Transformers to video file structure anomaly detection is still in its early stages. Relevant literature remains scarce. Some exploratory studies suggest that Transformers excel in capturing internal structural patterns, detecting abnormal frame sequences, and modeling global structural semantics. Given the complexity, diversity, and hidden nature of video file anomalies, more robust and flexible model architectures are needed. These should adapt to various encoding formats and packaging standards. This is the research gap that this paper aims to address.

3. Method

In order to effectively capture and model the complex internal dependencies present within the structural components of video files, this paper introduces a novel anomaly detection method specifically designed for assessing video structural integrity. The proposed approach is grounded in the Transformer architecture, which has demonstrated remarkable success in modeling sequential data across various domains. By leveraging its strong capability in learning long-range dependencies and structural relationships, the method is able to identify subtle and deeply embedded anomalies within video file structures. This design allows the model to move beyond surface-level checks and instead build a comprehensive understanding of the video's internal organization, including temporal sequences, encoding patterns, and metadata consistency. The detailed architecture of the proposed Transformer-based model, which serves as the backbone for this anomaly detection framework, is illustrated in Figure 1.

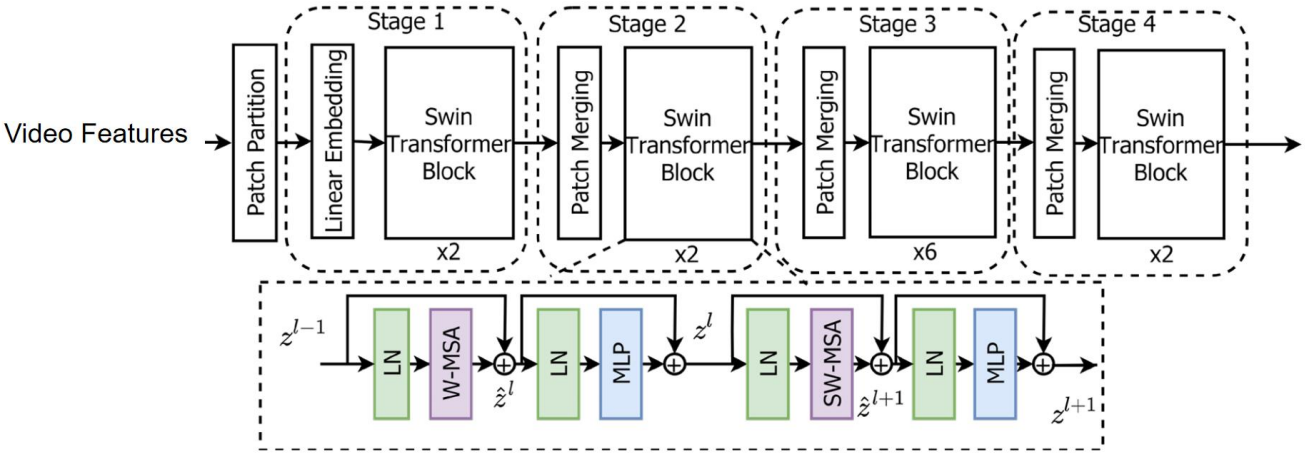

Figure 1. Overall model architecture

Figure 1 shows the overall model architecture for video file structural integrity anomaly detection proposed in this study. The model input is the parsed video file structural sequence features, including multi-dimensional structural information such as frame-level metadata, bitstream index, and encoding parameters. First, the structural features are divided by the Patch Partition module, and the original feature sequence is decomposed into multiple structural blocks to capture local structural patterns; then, the linear embedding layer performs dimensionality enhancement on each structural block to form a high-dimensional representation sequence as the input of the Transformer encoding. The core part consists of

four stages, each of which contains several Swin Transformer modules. The local window self-attention (W-MSA) and shifted window self-attention (SW-MSA) alternating modeling mechanism is adopted to effectively capture the long-range dependencies between structural units while maintaining computational efficiency. The Patch Merging operation is used between each stage to achieve layer-by-layer fusion of structural granularity, enhancing the model's perception of structural anomalies at different levels. Finally, the global structural representation generated by the network is used to evaluate the integrity of video files in the structural dimension, accurately identify abnormal patterns such as metadata damage, frame-level fracture, and encapsulation errors, which is in line with the research goal of video file-level anomaly detection.

Assume that the structure of a video file is parsed into a structural feature sequence $X = \{x_1, x_2, ..., x_n\}$ of length n, where each $x_i \in R^d$ represents the vectorized representation of the i-th structural unit (such as frame header, bitstream parameter, timestamp, etc.). First, the structural features are divided into fixed-length structural blocks through the Patch Partition operation, and then embedded into a uniform dimensional space through a linear mapping function f_{emb}, which is defined as follows:

$$z_i = f_{emb}(x_i) = W_e x_i + b_e$$

Where W_e is the learnable embedding matrix and z_i is the initial embedding vector of the i-th building block.

Subsequently, position encoding is introduced to maintain structural order information, so that Transformer has temporal perception when processing structural features. The position encoding p_i is added to the embedding vector z_i to form the input vector $h_i^{(0)}$.

$$h_i^{(0)} = z_i + p_i$$

The position encoding can be a fixed sinusoidal encoding or a learnable encoding method, which is used to enhance the model's perception of structural order anomalies (such as frame misalignment, timestamp inversion, etc.).

In the backbone network, each layer of Swin Transformer Block uses the self-attention mechanism within the local window to extract structural correlations. Within each window, the attention mechanism is calculated in the following form:

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Among them, $Q = hW_Q$, $K = hW_K$, and $V = hW_V$ represent the query, key, and value obtained by linear transformation of the input h, respectively, and d_k is the dimension of the key vector. The shift window mechanism further realizes the cross-window structural association modeling, enabling the model to detect cross-region structural errors (such as frame loss or interval anomalies).

After all Transformer stages are completed, the output sequence $H = \{h_1^{(L)}, h_2^{(L)}, ..., h_n^{(L)}\}$ of the fused structural dependencies is obtained and input into the anomaly scoring module. The final structural integrity anomaly score s is output through weighted aggregation, and the formula is as follows:

$$s = \sigma(\frac{1}{n}\sum_{i=1}^{n} w^T h_i^{(L)} + b)$$

w is a learnable parameter and σ is a Sigmoid function used to map the anomaly score to $[0,1]$, which represents the probability of structural anomaly. The binary cross entropy loss function is used to optimize the model during the training phase so that the predicted score is close to the true label:

$$L = -[y \log(s) + (1-y) \log(1-s)]$$

$y \in \{0,1\}$ is the labeling result of whether the video file structure is abnormal. This loss function makes the model focus on abnormal modeling in the structural dimension rather than the semantic difference of the video content, which is in line with the research goal of this paper on file integrity detection.

4. Experiment

4.1 Datasets

This study utilizes the EVA-7K (Encoded Video Anomalies 7K) dataset as the primary source of experimental data. EVA-7K is a large-scale benchmark specifically curated for evaluating video file-level structural anomaly detection. It contains over 7,000 video samples encoded with diverse codecs (e.g., H.264, H.265, VP9) and container formats (e.g., MP4, MKV, AVI), making it well-suited for tasks such as structural integrity analysis, anomaly detection, and container error recovery. The EVA-7K dataset comprises a comprehensive collection of both normal and anomalous video files. Structural anomalies in the dataset include corrupted headers, missing or inconsistent timestamps, broken frame sequences, misaligned keyframes, and partial or malformed bitstreams. Each anomaly is carefully labeled with metadata that indicates the nature, location, and extent of the structural damage, thereby supporting fine-grained evaluation of model performance.

One of the key advantages of EVA-7K is that each corrupted video is paired with its corresponding clean (original) version. This pairing enables contrastive or comparative learning during model training and evaluation, allowing models to learn discriminative structural features that differentiate normal from anomalous video files. The dataset's diversity and granularity in annotations enhance its utility for training models with strong generalization capabilities.

In our experiments, approximately 80% of the video samples from EVA-7K were used for model training, with the remaining 20% reserved for validation and testing. All features were extracted directly from the binary video files rather than image-level data. These extracted features include frame index sequences, timestamp intervals, segment size distributions, and header field structures. This ensures that the model input strictly represents low-level structural characteristics, fully aligning with this study's emphasis on file-level integrity verification. The comprehensiveness and high-quality annotations of the EVA-7K dataset provide a robust and reliable foundation for evaluating the effectiveness of the proposed methodology.

4.2 Experimental Results

First, this paper gives the comparative experimental results, and the experimental results are shown in Table 1.

Table 1: Experimental results

Method	Accuracy (%)	Precision	Recall	F1-Score
LSTM [13]	84.62	81.15	79.48	80.30
GRU [14]	85.27	82.44	80.03	81.22

Gated-GCN[15]	86.93	84.08	82.65	83.36
Temporal GCN [16]	88.12	85.97	84.51	85.23
Ours	91.48	89.35	87.22	88.27

As shown in Table 1, the proposed Transformer-based method for video structural integrity anomaly detection outperforms existing mainstream deep learning models across all evaluation metrics. Specifically, our model achieves an Accuracy of 91.48%, significantly higher than LSTM (84.62%) and GRU (85.27%). This indicates that the Transformer architecture provides stronger global perception when modeling video structural features. It captures long-term dependencies and latent anomaly patterns in file structures more effectively.

For Precision and Recall, our method reaches 89.35% and 87.22%, respectively. These results surpass those of Temporal GCN (85.97% and 84.51%) and Gated-GCN (84.08% and 82.65%). This demonstrates the model's advantage in reducing false positives and improving anomaly detection coverage. It also shows stronger capability in modeling boundary errors and complex dependencies within video container structures. The window-based attention mechanism used in the Transformer enhances local and cross-segment structural anomaly modeling while maintaining computational efficiency.

Moreover, the F1-Score, which balances Precision and Recall, reaches 88.27% with our model—again outperforming all baselines. This further confirms the stability and robustness of the proposed method in comprehensive structural anomaly detection. Compared to traditional sequential networks (e.g., LSTM, GRU) and graph-based models (e.g., Gated-GCN), our method adapts better to diverse and complex anomaly types. These results highlight the effectiveness and scalability of the Transformer architecture for video file-level structural analysis tasks.

Secondly, this paper presents an experiment on the generalization ability of video coding formats, and the experimental results are shown in Figure 2.

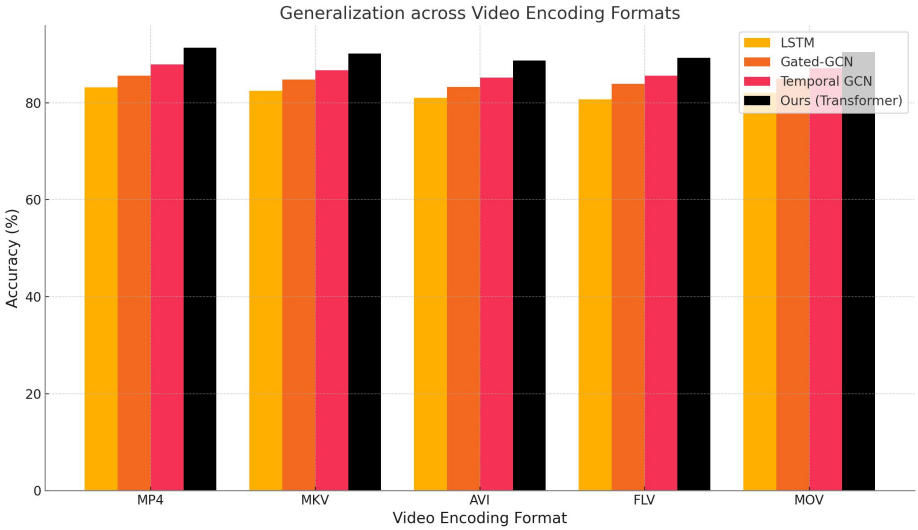

Figure 2. Video Coding Format Generalization Experiment

As shown in Figure 2, the proposed Transformer-based model for video structural integrity detection demonstrates excellent stability and strong generalization ability across a variety of mainstream video encoding formats. Specifically, when evaluated on MP4, MKV, AVI, FLV, and MOV formats, the model consistently achieves an accuracy level close to 90%, with minimal fluctuation. This performance

significantly exceeds that of other baseline models, underscoring the robustness and adaptability of the proposed approach. The results suggest that the model is capable of effectively capturing semantic variations and structural distinctions introduced by different encoding mechanisms and container standards. This cross-format generalizability is crucial for real-world applications, where video data are often heterogeneous in format and structure.

In contrast, baseline models such as LSTM and Gated-GCN exhibit evident performance degradation when confronted with video formats that exhibit large structural differences, particularly AVI and FLV. Their accuracy metrics tend to fluctuate by approximately 2% to 4% across different formats, highlighting their limitations in modeling complex or irregular structural features. Although the Temporal GCN demonstrates improved performance over traditional sequence-based models in most formats, it still lags behind the proposed Transformer-based model in overall accuracy and consistency. Notably, its detection capability on MOV and MKV formats remains relatively weak, indicating insufficient generalization to non-standard or richly structured container formats.

The superior and consistent performance of the proposed method across diverse encoding formats can be attributed to the intrinsic strengths of the Transformer architecture, particularly its capacity for hierarchical modeling of video file structure. The model's use of a window-based self-attention mechanism enables it to simultaneously capture fine-grained local patterns and integrate them into a coherent global representation. This dual capability enhances the model's understanding of both micro-level anomalies and macro-level structural consistency. The experiment thus verifies the method's high degree of transferability and its potential for practical deployment in diverse, heterogeneous, and large-scale video data systems. It also provides a solid foundation for future extension to more complex multimodal or multi-format video analysis scenarios.

Furthermore, this paper also presents an experiment to analyze the detection capability of different anomaly types, and the experimental results are shown in Figure 3.

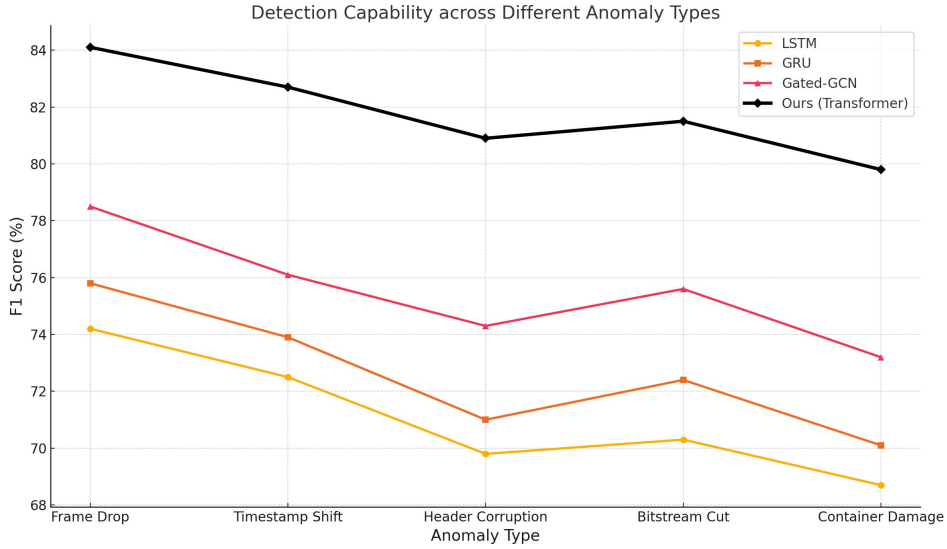

Figure 3. Detection Capability across Different Anomaly Types

As shown in Figure 3, the proposed Transformer-based model exhibits consistently stable and superior performance when applied to a variety of video structural anomaly detection tasks. Notably, in the detection of common temporal anomalies such as "Frame Drop" and "Timestamp Shift," the model achieves impressive F1 scores of 84.1% and 82.7%, respectively. These results significantly outperform those of other baseline models, including LSTM, GRU, and Gated-GCN. The performance gap illustrates

the Transformer model's strong ability to capture subtle temporal disruptions and sequence-level inconsistencies, such as missing or misaligned keyframes. Its effectiveness in modeling temporal dependencies and understanding structural context highlights the advantages of attention-based mechanisms over traditional sequential or graph-based approaches, which often struggle with long-range temporal modeling and context comprehension.

When facing low-level structural corruption types such as "header corruption" and "bitstream cut," the limitations of traditional models become more apparent. For example, the performance of LSTM drops sharply, with F1 scores falling to as low as 69.8%. This decline indicates that models relying on recurrent mechanisms or fixed graph structures have difficulty identifying irregularities in deeply embedded or non-sequential file components. In contrast, the Transformer model maintains detection performance above 80% in these scenarios, demonstrating strong robustness and resilience. The improvement can be attributed to the multi-layer hierarchical attention mechanism, which empowers the model to extract and integrate features from multiple semantic levels. This capacity allows it to accurately recognize complex structural anomalies that span different parts of the file and prevents errors such as false positives or overlooked anomalies, which are common in models with limited receptive fields or shallow representation capabilities.

Of particular significance is the model's performance in detecting "Container Damage" anomalies—structural faults that impact the overall integrity and packaging of the video file. In this challenging category, the Transformer model still achieves the highest F1 score of 79.8%, whereas other models exhibit a marked performance degradation. This result demonstrates the proposed method's ability to go beyond detecting localized or frame-level defects; it can also assess global structural coherence and packaging integrity. Such comprehensive detection capability underscores the model's adaptability to diverse anomaly types and its practical value in real-world applications, where multiple forms of structural inconsistencies may coexist. The Transformer's balanced performance across all anomaly categories confirms its potential as a unified solution for multi-type and multi-level structural anomaly detection in complex video systems. Furthermore, this paper presents the experimental results of model inference efficiency under low-resource settings, as shown in Figure 4.

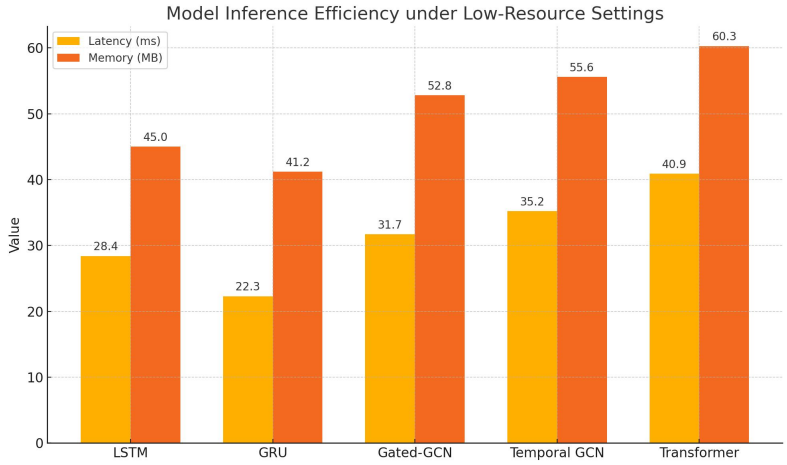

Figure 4. Model Inference Efficiency under Low-Resource Settings

As shown in Figure 4, there are clear differences in inference efficiency among the models under low-resource conditions. LSTM and GRU show low resource consumption in both latency and memory usage. GRU, in particular, achieves the lowest latency at 22.3 ms, making it the most efficient among all models. This makes it suitable for deployment on edge devices or in scenarios with limited computational

capacity. Gated-GCN and Temporal GCN have slightly higher overhead. They offer a balanced trade-off between performance and efficiency.

In contrast, the proposed Transformer-based model uses 60.3 MB of memory and has a latency of 40.9 ms. These are higher than the lightweight models. This is mainly due to the multi-layer window attention mechanism and feature fusion modules. While these components improve detection accuracy, they also introduce additional inference cost. However, the resource demands remain within a controllable range. In systems with moderate hardware, the model can still support real-time processing.

Overall, the experiment confirms the Transformer model's adaptability in resource-constrained environments. Although it shows slightly higher latency and memory usage, its performance advantage remains significant. It is especially suitable for industrial video file scenarios where high accuracy in structural integrity detection is required. In the future, techniques like model compression and distillation may help reduce resource usage further and improve deployment flexibility.

Finally, this paper gives a loss function drop graph, as shown in Figure 5.

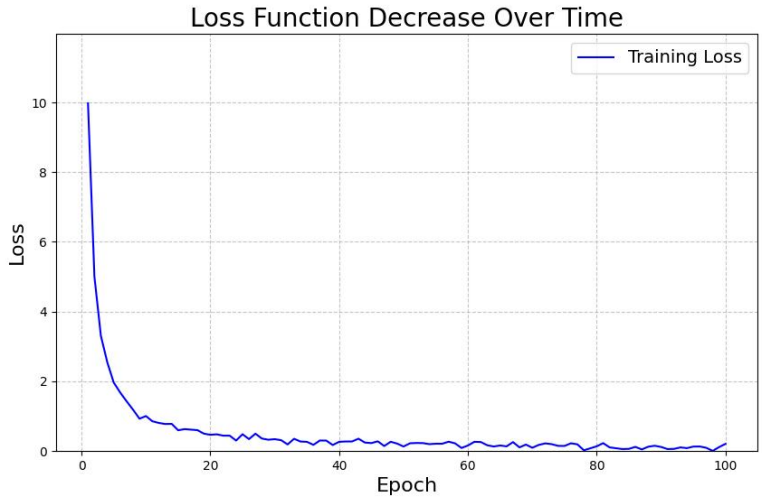

Figure 5. Loss function drop graph

Figure 5 shows the trend of the loss function (Loss) changing with the number of training rounds (Epoch) during training. It can be observed from the figure that in the early stage of training (within the first 10 epochs), the loss function decreases very quickly, indicating that the model has a significant learning effect on structural features in the initial stage, quickly grasps the basic pattern of the video file structure and corrects most of the wrong predictions.

As the training progresses, the loss value gradually stabilizes, especially after 20 epochs, the decline becomes smaller and tends to converge, indicating that the model has basically completed the feature learning of structural integrity anomalies. The slight fluctuation of loss at this stage is normal, which may be related to the complexity or diversity of abnormal samples in the training data, but it remains at a low level overall, showing the stability and robustness of model training.

In the final stage (after 60 epochs), the loss value fluctuates less and basically remains below 0.2, indicating that the model has a strong generalization ability and the learning effect of video structural anomalies is stable and reliable. This result verifies the effectiveness of the proposed Transformer structure in modeling video file structure sequences, and also provides a good training basis for the actual deployment of subsequent models and structural anomaly detection.

5. Conclusion

This paper addresses the limitations of current video file structural integrity detection methods, including insufficient accuracy and limited modeling capacity. A Transformer-based structural anomaly detection approach is proposed. The method constructs structural feature sequences from video files and leverages the multi-scale modeling capabilities of the Swin Transformer. It enables deep representation of container formats, frame sequences, and structural dependencies. This provides a modeling scheme that combines global awareness with local sensitivity for file-level anomaly detection.

In the experimental section, the model is systematically evaluated in terms of overall performance, comparative analysis, anomaly type recognition, cross-format generalization, and inference efficiency under resource constraints. The results show that the proposed method significantly outperforms mainstream baselines across multiple metrics. It maintains strong detection performance even in complex and subtle structural corruption scenarios, demonstrating both effectiveness and generalizability.

In addition, real-world video file structural data is used for model training and evaluation, ensuring the credibility and practical relevance of the results. The observed differences in modeling various anomaly types also provide insights for future development of fine-grained detection models. Although the proposed method requires more computational resources than lightweight models, the improvement in detection accuracy justifies the cost in high-demand application scenarios. In future work, this approach can be extended to multimodal file structure modeling, such as jointly analyzing the consistency of video, audio, and subtitle tracks. Model compression and distillation techniques can also be introduced to enhance deployment efficiency in edge computing environments. Moreover, combining the method with secure training strategies such as federated learning may promote real-world applications in areas like forensic analysis, secure video transmission, and content compliance review.

References

[1] Xiang, Z., Horváth, J., Baireddy, S., Bestagini, P., Tubaro, S., & Delp, E. J. (2021). Forensic analysis of video files using metadata. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1042-1051).

[2] Yang, P., Baracchi, D., Iuliani, M., Shullani, D., Ni, R., Zhao, Y., & Piva, A. (2020). Efficient video integrity analysis through container characterization. IEEE Journal of Selected Topics in Signal Processing, 14(5), 947-954.

[3] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. Advances in neural information processing systems, 34, 15908-15919.

[4] Neimark, D., Bar, O., Zohar, M., & Asselmann, D. (2021). Video transformer network. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3163-3172).

[5] Ranasinghe, K., Naseer, M., Khan, S., Khan, F. S., & Ryoo, M. S. (2022). Self-supervised video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2874-2884).

[6] Kim, T. H., Sajjadi, M. S., Hirsch, M., & Scholkopf, B. (2018). Spatio-temporal transformer network for video restoration. In Proceedings of the European conference on computer vision (ECCV) (pp. 106-122).

[7] Gupta, M., Shah, S., & Salmani, S. (2021, June). Improving whatsapp Video Statuses using FFMPEG and Software based encoding. In 2021 International Conference on Communication information and Computing Technology (ICCICT) (pp. 1-6). IEEE.

[8] Chandru, R., & Priscilla, R. (2024, March). Video Integrity Detection with Deep Learning. In 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1288-1292). IEEE.

[9] Hu, X., Pan, Y., Wang, Y., Zhang, L., & Shirmohammadi, S. (2021). Multiple description coding for best-effort delivery of light field video using GNN-base

[10] Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., & Lin, W. (2023). Discovqa: Temporal distortion-content transformers for video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 33(9), 4840-4854.

[11] Li, Y., Ye, J., Zeng, L., Liang, R., Zheng, X., Sun, W., & Wang, N. (2024). Learning hierarchical fingerprints via multi-level fusion for video integrity and source analysis. IEEE Transactions on Consumer Electronics, 70(1), 3414-3424.

[12] Capasso, P., Cattaneo, G., & De Marsico, M. (2024). A comprehensive survey on methods for image integrity. ACM Transactions on Multimedia Computing, Communications and Applications, 20(11), 1-34.

[13] Ullah, W., Ullah, A., Hussain, T., Khan, Z. A., & Baik, S. W. (2021). An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. Sensors, 21(8), 2811.

[14] Kwong, N. W., Chan, Y. L., Tsang, S. H., & Lun, D. P. K. (2023). Quality feature learning via multi-channel CNN and GRU for no-reference video quality assessment. IEEE access, 11, 28060-28075.

[15] Zhang, Z., Huang, L., Tang, B. H., Wang, Q., Ge, Z., & Jiang, L. (2025). Non-Euclidean Spectral-Spatial feature mining network with Gated GCN-CNN for hyperspectral image classification. Expert Systems with Applications, 126811.

[16] Wu, S., Chen, J., Xu, T., Chen, L., Wu, L., Hu, Y., & Chen, E. (2021, October). Linking the characters: Video-oriented social graph generation via hierarchical-cumulative GCN. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 4716-4724).