MMTFL: Multi-Timescale Multi-Modal Feature Learning for Weakly-Supervised Anomaly Detection

Erkut Akdag *, Henk Corporaal, Peter H. N. D. With, and Egor Bondarev

Electrical Engineering Department, Eindhoven University of Technology, Eindhoven, The Netherlands
Email: e.akdag@tue.nl (E.A.); h.corporaal@tue.nl (H.C.); p.h.n.de.with@tue.nl (P.H.N.D.W.); e.bondarev@tue.nl (E.B.)
*Corresponding author

Abstract*—Detection of anomalous events is critical for public safety and requires capturing fine-grained motion patterns and contextual information across multiple time-scales. To this end, we propose a Multi-Timescale Feature Learning (MTFL) method to enhance the representation of anomaly features. Short, medium, and long temporal tubelets are employed to extract spatio-temporal video features using a Video Swin Transformer. Experimental results demonstrate that MTFL achieves an anomaly detection performance 87.16% Area Under the Curve (AUC) on the University of Central Florida (UCF)-Crime dataset and 84.57% Average Precision (AP) on the Xi Dian University (XD)-Violence dataset. While MTFL relies solely on spatio-temporal features extracted from a single modality using RGB video, it encounters challenges such as occlusions, ambiguous actions, and limited contextual understanding. To overcome these limitations, we also propose Multi-Modal Multi-Timescale Feature Learning (MMTFL), which integrates spatio-temporal, depth, and text-based features in conjunction with multi-timescale tubelet analysis, rather than focusing only on RGB inputs. Although adding modalities increases feature extraction cost, it remains feasible for real-world purposes. Experimental results demonstrate that the MMTFL outperforms single-modality approaches, achieving 88.29% AUC on the UCF-Crime dataset and 84.96% AP on the XD-Violence dataset. By leveraging complementary information from multiple modalities, the proposed approach achieves more robust and accurate detection of complex and diverse anomalies compared to single-modal methods.

***Keywords*—anomaly detection, surveillance videos, video understanding, multi-modality, feature fusion, attention**

I. Introduction

Nowadays, the smart city concept is a significant research direction that utilizes data collected by a variety of technologies and sensor types to efficiently optimize the delivery of various services and improve life quality for the public. In this setting, anomaly detection in surveillance videos helps to enhance public safety by early detection of threats and incidents. In the current industrial settings, the control-room operators are watching the feeds from multiple surveillance cameras on a display grid. The advances of surveillance technologies have led to a parallel collection of vast amounts of video data, making manual monitoring increasingly impractical. Therefore, automated anomaly detection is a necessity to remove the limitations of manual inspections, while the broad availability of various technologies makes it feasible. There are two popular approaches towards anomaly detection in literature. The first one is the binary classifier method, also referred as 'unsupervised', that learns the dominant data representation, using normal data only. This approach is often inadequate due to vaguely defined borders of normality in the latent space. Alternatively, weakly-supervised techniques have gained significant popularity, reducing the annotation work by employing video labels. However, the existing anomaly detection approaches still face several issues.

One of the major challenges in anomaly detection is the variable temporal duration of anomalies. In the real world, anomalies occur rarely and vary in duration, leading to complications in learning the temporal dynamics of the events. For instance, burglary and arson anomalies last longer, while littering and traffic accidents are short-duration events. The divergence between short, medium, and long-lived anomalies needs exploration in flexible learning over time, e.g. using time scales. A model should learn features from data in multiple temporal scales to detect anomalies, regardless of their durations.

Another limitation is that conventional anomaly detection methods predominantly rely on a single modality, such as RGB video, and often struggle to handle the complexity of real-world environments. The single source of input information makes them vulnerable to challenges like poor lighting, occlusions, or ambiguous actions. These limitations emphasize the need for more encompassing solutions to exploit diverse data sources and to capture the full spectrum of contextual and semantic cues in dynamic environments. Therefore, in addition to standard RGB video data, other data modalities like radar,

Manuscript received May 19, 2025; revised July 18, 2025; accepted September 1, 2025; published February 27, 2026.

depth, acoustic signals, infrared, and LiDAR can be considered to achieve complementary spatial, environmental, and semantic information for scene understanding. For instance, depth data can provide critical spatial context, reducing ambiguities in object interactions, while text features can enhance semantic interpretation, improving comprehension of events.

To overcome the aforementioned challenges, we propose the Multi-Timescale Feature Learning (MTFL) and Multi-Modal Multi-Timescale Feature Learning (MMTFL) methods. The MTFL method extracts features at different temporal scales and fuses motion details from short temporal tubelets with contextual information from long tubelets, enhancing the representation capability of anomaly features. The method employs multi-head cross attention, multi-head self-attention, and 1D convolution kernel to capture the correlations between multi-timescale features, global temporal dependencies across the three integrated feature levels (short, medium, and long), and local temporal dependencies within each tubelet, respectively. These factors are utilized for the scaling and fusion of the features to achieve the final feature representation.

The MMTFL method is proposed to address the limitations of single-modality methods, by integrating RGB, depth, and text features with different feature fusion techniques. By leveraging complementary information from multiple modalities, the proposed approaches achieve more robust and accurate detection of complex and diverse anomalies compared to single-modality methods. To summarize, the main contributions of our work are as follows.

- A multi-timescale anomaly detection model (MTFL) for learning the correlation between different timescale features. The fusion of short, medium, and long temporal features leverages the deployment of variable temporal scales.
- A multi-modal anomaly detection model (MMTFL), offering high performance compared to single-modal approaches, which combines visual, depth, and text features.
- Exploring the individual contributions of video, depth, and text features to the enhancement of anomaly detection performance.

II. Related Work

A. Weakly Supervised Anomaly Detection

A weakly-supervised anomaly detection approach was introduced and extended [1–7]. It uses video-level annotation, which is labeling entire videos as normal or anomalous rather than frame-level labels, significantly reducing annotation costs and enabling models to learn from generalized anomaly patterns.

Weakly-supervised methods often employ Multiple Instance Learning (MIL), where sets of video instances ("bags") are labeled collectively. The model learns to detect anomalies by identifying patterns within bags, even if only a subset of instances is anomalous. Attention mechanisms further improve detection by highlighting relevant spatial or temporal features.

While being annotation-efficient, weak supervision limits precise anomaly localization and may increase false positives due to coarse labeling. Still, it offers a practical trade-off, allowing broader generalization with reduced annotation effort, albeit at the cost of fine-grained accuracy. Recent work has also explored leveraging human pose features for anomaly detection to improve robustness and generalization. Hirschorn *et al*. [8] introduces a method utilizing normalizing flows for human pose anomaly detection, showing strong results under weak supervision without requiring dense frame-level labeling. This work exemplifies the trend towards using structured semantic features, such as pose information, in MIL frameworks for efficient anomaly detection.

B. Video Captioning

Video Captioning (VC) is a key task in video understanding [9]. Early work explored 2D/3D representations [10–13], and later studies focused on object-level features to improve VC [14–16]. With the rise of transformers [17, 18], vision-language models have achieved strong results across various tasks [19–24]. Recent transformer-based VC models further improved performance through end-to-end training [25–27].

Large-scale vision-language models like Contrastive Language-Image Pretraining (CLIP) [12] bridge visual and textual modalities. In video anomaly detection, recent work uses textual prompts to enrich anomaly representations [28–31]. Open-vocabulary anomaly detection [32] and prompt-based scoring with LLMs [33] have emerged, though current methods often skip domain-specific fine-tuning, making performance reliant on base LLM capabilities.

C. Video Depth Estimation

Video depth estimation methods are generally classified into feed-forward prediction and test-time optimization. Feed-forward approaches directly predict depth from video frames [34–41]. For instance, DeepV2D integrates depth and camera motion estimation [35]; MAMO uses memory attention [39], and NVDS adds a stabilization module [38]. While efficient, these methods often underperform in open-world scenarios due to limited training data. Test-time optimization approaches refine depth during inference using cues like camera poses or optical flow, offering consistency, but reduced generalization to unconstrained settings [42–45].

D. Multi-Modal Video Learning

Multi-modal video learning integrates modalities such as vision, audio, and text to enhance video understanding. A key task is Temporal Sentence Grounding in Videos (TSGV), which aligns language queries with temporal segments in videos, which is crucial for surveillance event localization [46–49]. Another core task is Video Captioning (VC), which generates textual descriptions of video content, aiding automatic understanding and summarization [50, 51]. Extending VC, Dense Video Captioning (DVC) localizes and describes multiple events

in untrimmed videos, offering detailed surveillance insights [50, 52]. Multi-modal Anomaly Detection (MAD) [53] further combines VC features with visual-temporal data to identify anomalies. Current MAD methods often rely on captions from SwinBERT [25] and are typically trained on open-domain datasets such as VATEX [54].

Although existing methods have advanced the field of video anomaly detection, they often fall short in addressing critical challenges, such as the requirement to operate across multiple temporal scales to effectively capture anomalies of varying durations. To mitigate this limitation, the MTFL method is proposed to learn multi-timescale features for improving the anomaly detection performance. Moreover, despite recent progress in multi-modal learning, video captioning, and depth estimation, significant challenges persist particularly in the efficient fusion of heterogeneous features. Single modality.

Approaches often fail to capture the full scope of contextual cues for reliable anomaly detection. To address these gaps, we propose the MMTFL method that integrates spatio-temporal, depth, and text-based features. This integration is guided by diverse loss functions and inter-modal correlation modelling, which together enhance the model's capacity to learn robust representations. The proposed methods improve anomaly detection accuracy, surpassing single-modality baselines through richer and more discriminative contextual understanding.

III. Multi-Timescale Feature Learning (MTFL)

Considering the variations in the attention focus of feature information across different frame lengths, the Multi-Timescale Feature Learning (MTFL) method aims to integrate essential features from multiple temporal scales, to enhance the discrimination between abnormal and normal snippets. Given the importance of capturing fine-grained motion details and long-term contextual dependencies, feature representations are derived at different temporal resolutions, leveraging short, medium, and long temporal tubelets, to ensure holistic comprehension of motion dynamics and event progression. Temporal tubelets are subsets of video frames categorized by different temporal durations: short, medium, and long. Short tubelets focus on high-frequency motion details and local variations for detecting rapid or subtle actions. Medium tubelets balance detail with a broader temporal scope, allowing the model to analyze interactions and transitions over time. Long tubelets capture extended temporal dependencies, providing a global understanding over time of the scene.

A video snippet refers to a segment of a video, typically consisting of a fixed number of consecutive frames. As illustrated in Fig. 1, the input video is divided into T snippets to ensure a fair comparison with SotA models on public datasets. Each snippet constructs different temporal tubelets based on variable frame lengths, which are used for feature extraction. The Multi-Timescale Feature Generator (MTFG) creates three groups of $T\times D$ features, corresponding to T snippets, namely F_L, F_M, and F_S, long, medium, and short temporal tubelets. Next, the Multi-Timescale Feature Fusion (MTFF) fuses output features obtained from three temporal paths, to create a global feature vector X based on temporal correlations between the snippets. This feature vector is generated by multi-head cross and self-attention blocks. Subsequently, we apply a classifier to vector X and generate corresponding anomaly scores for each snippet.

A. Multi-Timescale Feature Generator (MTFG)

Given T snippets with a frame length of L segmented from an input video, the MTFG generates three feature matrices of size $T\times D$, corresponding to the D dimensional features of these T snippets. Each feature matrix represents a temporal tubelet with the corresponding frame length. As illustrated in Fig. 2, a pre-trained Video Swin Transformer extracts features from a single snippet t within a group of long temporal tubelets with a frame length L_1, medium temporal tubelets with a frame length L_2, and short temporal tubelets with a frame length L_3, resulting in $F'_L(t)$, $F'_M(t)$, $F'_S(t)$ vectors, respectively. The vectors in each feature matrix are averaged along the temporal axis to form single feature vectors $F_L(t)$, $F_M(t)$, $F_S(t)$. With T snippets of the input video, the MTFG outputs three groups of T feature vectors, i.e. $\{F_L(0),...,F_L(T-1)\}$, and the other two with the same formulation with F_M, and F_S for medium and short tubelets, respectively. The T feature vectors in each group are concatenated along the temporal axis, resulting in three feature matrices referring to F_L, F_M, and F_S in Figs. 1 and 3.

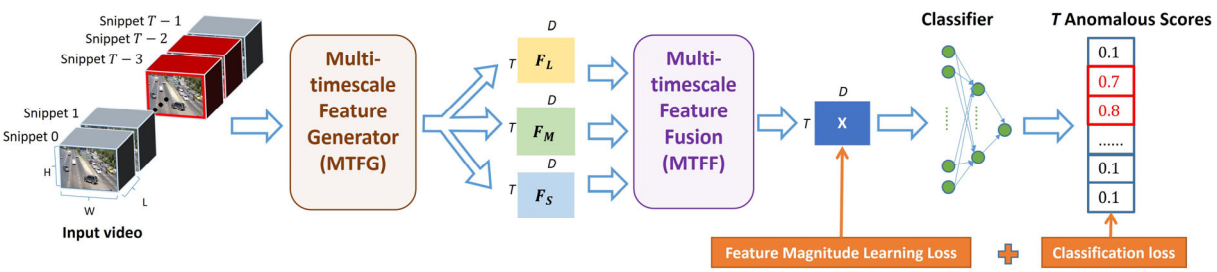

Fig. 1. Workflow of Multi-Timescale Feature Learning (MTFL) model. The input video is segmented into T snippets. The Multi-Timescale Feature Generator (MTFG) creates three sets of features, F_L, F_M, and F_S corresponding to features extracted within long, medium, and short temporal tubelets. Next, the Multi-Timescale Feature Fusion (MTFF) captures the correlations among these three features and the dependencies among different video snippets to fuse the features into the output feature matrix X. The final anomaly scores of T snippets are obtained after a classifier. A loss function involving feature magnitude loss and classification loss is used for training the MTFF and the anomaly classifier.

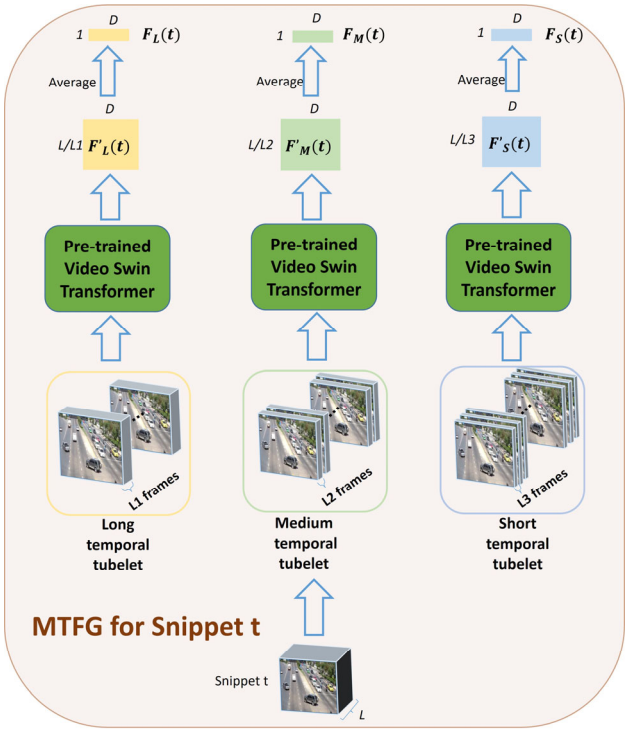

Fig. 2. Multi-Timescale Feature Generator (MTFG). Taking snippet t, a pre-trained Video Swin Transformer extracts features using 3 temporal tubelets with long, medium, and short frame lengths, i.e. L_1, L_2, and L_3, and the obtained features are averaged to 3 single feature vectors.

B. *Multi-Timescale Feature Generator (MTFF)*

The MTFF fuses the resulting feature maps according to their dependencies, as well as the local and global temporal correlation between snippets. As shown in Fig. 3, the architecture consists of 4 fusion modules, mod (i…iv).

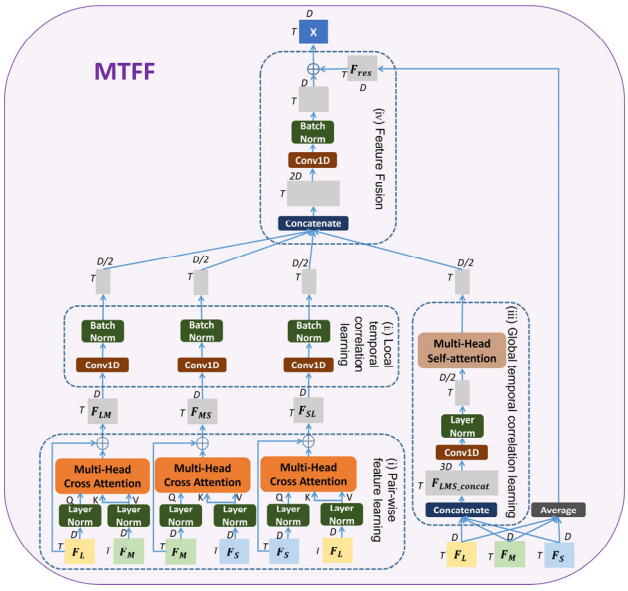

Fig. 3. Multi-Timescale Feature Fusion (MTFF). PFL Mod (i) learns the correlations between different path features for pairwise integration. Mod (ii) captures local temporal dependencies and scales the features accordingly. Mod (iii) learns the global temporal dependencies of the integrated three features among T snippets. Mod (iv) obtains the final feature X by fusing the output features from mod's (ii, iii).

- **Mod (i)** Pairwise Feature Learning (PFL) employs three Multi-head Cross-Attention (MCA) functions to perform pairwise fusion of 3 feature matrices F_L, F_M, and F_S generated by MTFG, and outputs F_{LM}, F_{MS}, and F_{SL}.
- **Mod (ii)** Local Temporal correlation Learning (LTL) scales pairwise fusion matrices by their local temporal correlation through 1D convolutional kernels.
- **Mod (iii)** Global Temporal correlation Learning (GTL), F_L, F_M, and F_S are concatenated to form a feature matrix F_{LMS_concat}. After the dimensional reduction of F_{LMS_concat} by 1D convolution, we use the Multi-head Self-Attention (MSA) block to scale the features based on the global temporal correlations across different snippets.
- **Mod (iv)** Feature Fusion (FF) concatenates the output feature maps of size $T\times D/2$ from mod's (ii) and (iii) and applies a 1D convolutional kernel to reduce the dimension. In the end, the MTFF outputs the final fused feature X as shown in Fig. 3 with a residual connection.

C. *Anomaly Detection*

After the MTFF module, the fused feature matrix X is supplied into a 3-layer fully connected classifier to output the anomaly probability scores for T snippets, as shown in Fig. 1. The loss function proposed by Tian *et al*. [3] is used for training the anomaly detection model, including loss functions for feature magnitude learning and classifier training. A brief overview of anomaly detection loss $\mathcal{L}_{AD}$ is provided as follows.

$$\mathcal{L}_{AD} = \mathcal{L}_{BCE} + \lambda_{fm}\mathcal{L}_{FM} + \lambda_1 \sum_{t=0}^{T-1} \left| s_t^+ \right| + \lambda_2 \sum_{t=1}^{T-1} \left| s_t^+ - s_{t-1}^+ \right|^2 \quad (1)$$

where $\mathcal{L}_{FM}$ is the feature magnitude learning loss function [3] that maximizes the separability between the top-k features from normal and abnormal videos, the Binary Cross-Entropy (BCE) loss $\mathcal{L}_{BCE}$ is used for training the classifier, λ_{fm}, λ_1, and λ_2 are the weighting factors of each loss component. The components s_{t-1}^+ and s_t^+ represent the scores of two consecutive snippets t-1 and t in an anomalous input, respectively. The summations $\sum_{t=0}^{T-1} \left| s_t^+ \right|$ and $\sum_{t=1}^{T-1} \left| s_t^+ - s_{t-1}^+ \right|^2$ are added for sparsity and temporal smoothness constraints of scores in anomalous videos.

IV. Multi-Timescale Feature Learning (MMTFL) Method

Exploiting multiple modalities is essential for achieving reliable anomaly detection, as each modality type provides unique and complementary insights into an event. Spatio-temporal features capture motion and structural patterns over time, depth features enhance spatial awareness by distinguishing object positioning and enhancing

background separation, while text-based features offer semantic context that aids in event interpretation. To harness the strengths of these diverse modalities, the MMTFL is introduced by extending the MTFL model. The MTFG component of MTFL model is employed for extracting spatio-temporal, depth, and text-based features from video sequences. In feature fusion step, the MTFF module is examined together with concatenation and mean functions in different combinations, considering the same classifier and loss function.

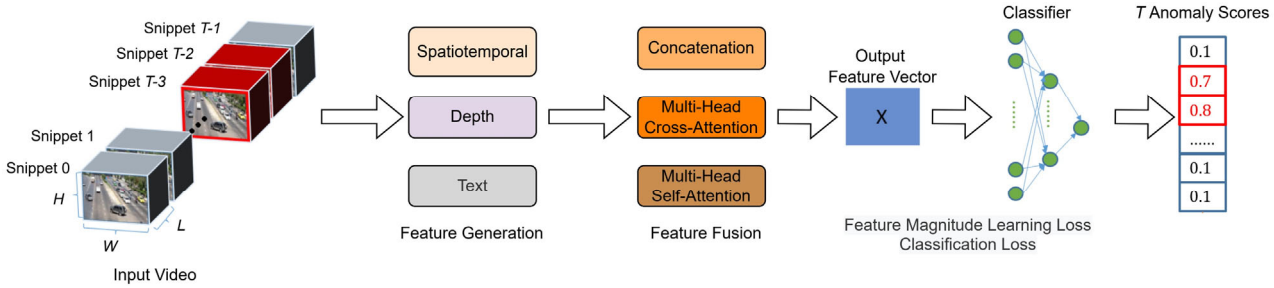

Fig. 4. Workflow of the MMTFL model. Spatio-temporal, depth, and text features are extracted at varying temporal scales. After obtaining these diverse features, the feature fusion stage captures the correlations among them and the dependencies across different video segments, which outputs the feature matrix X. Subsequently, anomaly scores for video snippets are obtained through a fully-connected classifier that considers both feature magnitude and classification losses.

As depicted in Fig. 4, the MMTFL is designed based on video snippets, multi-timescale feature levels, and multiple modalities to enhance feature representation. The multi-timescale spatio-temporal features, extracted from short, medium, and long temporal tubelets, are incorporated with depth, and text information. This enables to obtain both localized details and broader contextual patterns, thereby improving the detection of anomalies in complex video environments. The details of each modality are summarized as follows:

- *Spatio-temporal RGB features* are extracted utilizing a VST [55], a hierarchical transformer-based model designed for video understanding. The VST model employs shifted window attention to efficiently capture long-range dependencies while maintaining computational efficiency. The selection of this model has been justified in the ablation study through comparative analysis with other feature extraction techniques. For a video containing T snippets, the feature generation step produces three groups of feature vectors: short-term, medium-term, and long-term representations, represented by F_S^{rgb}, F_M^{rgb}, and F_L^{rgb}, each having a dimensionality of $T\times D$.
- *Depth features* provide valuable cues for distinguishing objects and background elements, improving motion segmentation, and identifying anomalies based on depth variations. To incorporate depth features into the MMTFL model, the Depth AnythingV2 model is adopted for dense depth estimation, since it delivers SotA performance in extracting fine-grained depth representations [35]. Depth AnythingV2 employs a transformer-based architecture to predict per-pixel depth values for each frame in T snippets, which enhances scene understanding beyond RGB-based representations. The generated three groups of depth feature vectors are short-term, medium-term, and long-term features, denoted as F_S^{depth}, F_M^{depth}, and F_L^{depth} respectively, each with a size of $T\times D$.
- *Textual features* are integrated into MMTFL employing a vision-language model to extract semantic representations from each associated video snippet. The CLIP model aligns image and text in a shared embedding space through contrastive learning, leveraging large-scale image-text pairs to enhance semantic understanding and improve multi-modal feature fusion [12]. This model is selected to ensure a fair comparison with existing methods in literature. The CLIP model generates three text feature vectors from a video of T snippets, capturing short-term, medium-term, and long-term representations. These feature groups are F_S^{text}, F_M^{text}, and F_L^{text}, each with a dimensionality of $T\times D$.

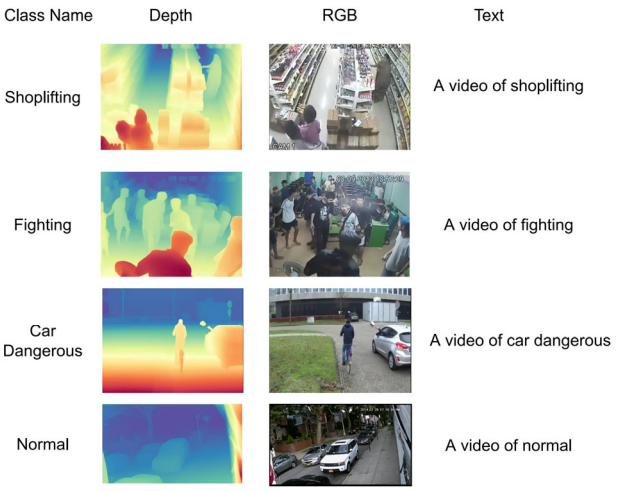

Fig. 5. Example frames of four anomaly classes shoplifting, fighting, dangerous throwing, and normal for three different modalities. The first column displays the depth map generated from the corresponding RGB input, by employing the Depth AnythingV2 model. The second column presents the original RGB image. The third column contains the text prompt applied to obtain the text embeddings using the CLIP model.

For short tubelets, depth features capture fine-grained depth changes, such as sudden object movements, while text-based features provide concise descriptions of

immediate actions. At the level of medium tubelets, depth features facilitate tracking of dynamic depth shifts, while text-based features describe evolving scene semantics, such as object interactions or behavioural changes. In long tubelets, depth features emphasize long-term structural consistency, while textual features contextualize event sequences, enhancing anomaly detection by aligning motion patterns with semantic interpretations. Therefore, the proposed MMTFL model improves the ability to differentiate between normal and anomalous events in complex video scenarios, by combining multi-modal feature fusion with multi-timescale tubelet analysis. Some example frames from four anomaly classes, each corresponding to various modalities, are shown in Fig. 5.

A. Feature Fusion Strategies

Different feature fusion strategies, three designs (Feature Fusion 1, 2, and 3), are proposed for the MMTFL model (Table I). The details of these feature fusion techniques are provided in Figs. 6–8. To further clarify the implementation, pseudocode descriptions of each feature fusion strategy are also included, making the explanations more accessible and transparent. Feature Fusion 2 is selected for the proposed architecture as it achieves the highest performance with 88.29% AUC, demonstrating that averaging depth and text features prior to concatenation, while processing spatio-temporal features separately with MTFF feature fusion, leads to better anomaly detection.

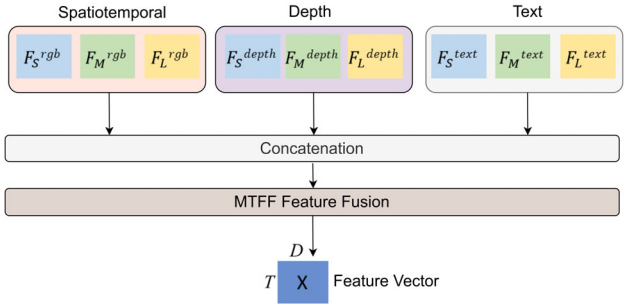

Fig. 6. Feature Fusion 1: Overview of the multi-modal feature fusion process in the MMTFL model. The extracted spatio-temporal, depth, and text features are concatenated and passed through the MTFF module, producing the final feature vector X for anomaly detection.

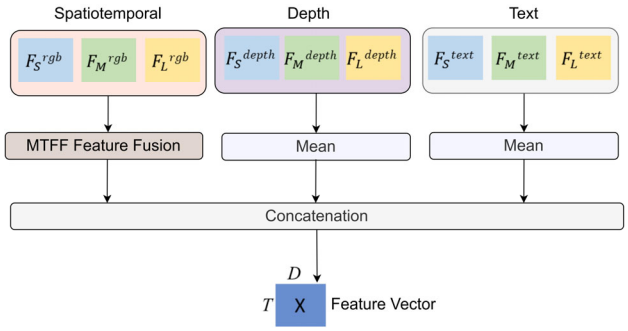

Fig. 7. Feature Fusion 2: Depth and text features are first averaged across temporal tubelets prior to fusion, while spatio-temporal features are processed through the MTFF module. The resulting feature representations are then concatenated to generate the final feature vector X for anomaly detection.

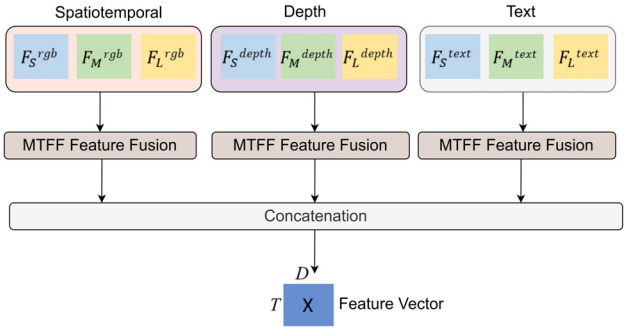

Fig. 8. Feature Fusion 3: Spatio-temporal, depth, and text features are independently processed through the MTFF module. The fused representations from each modality are then concatenated to form the final feature vector X for anomaly detection.

TABLE I. PSEUDOCODE FOR FEATURE FUSION TECHNIQUES

Classification	Feature Fusion Strategies
Feature Fusion 1	**Concatenate** spatio-temporal features from short, medium, and long temporal tubelets: F_{con}^{rgb} **Concatenate** depth features from short, medium, and long temporal tubelets: F_{con}^{depth} Concatenate text features from short, medium, and long temporal tubelets: F_{con}^{text}
	Apply MTFF feature fusion on the obtained concatenated vectors (F_{con}^{rgb}, F_{con}^{depth}, F_{con}^{text}) Output: X
Feature Fusion 2	**Apply** MTFF feature fusion for the spatio-temporal feature: F_{MTFF}^{rgb} **Compute** average of depth features from short, medium, and long temporal tubelet: F_{mean}^{depth} Compute average of text features from short, medium, and long temporal tubelets: F_{mean}^{text}
	Concatenate obtained spatio-temporal, depth, and text features (F_{MTFF}^{rgb}, F_{mean}^{depth}, F_{mean}^{text}) Output: X
Feature Fusion 3	**Apply** MTFF feature fusion for the spatio-temporal features: F_{MTFF}^{rgb} **Apply** MTFF feature fusion for the depth features: F_{MTFF}^{depth} Apply MTFF feature fusion for the text features: F_{MTFF}^{text}
	Concatenate obtained spatio-temporal, depth, and text features (F_{MTFF}^{rgb}, F_{MTFF}^{depth}, F_{MTFF}^{text}) Output: X

Input: Spatio-temporal features: F_S^{rgb}, F_M^{rgb}, and F_L^{rgb}
Depth features: F_S^{depth}, F_M^{depth}, and F_L^{depth}
Text features: F_S^{text}, F_M^{text}, and F_L^{text}.

V. EXPERIMENTS

A. Implementation Details

The MTFG module of the MTFL method extracts video features from different frame-length tubelets. We apply frame lengths of $L_1 = 8$, $L_2 = 32$, $L_3 = 64$ (motivated by optimal combination of tubelets ablation studies) for short, medium, and long temporal tubelets, respectively. The sampling interval for all tubelets is 1 frame, and the output feature map of an input video comprises 32 feature vectors, corresponding to the total number of snippets $T = 32$.

For feature extraction, we utilize pre-trained Video Swin Transformer model pre-trained on the Kinetics-400 dataset, named as VST-RGB. Moreover, we use pre-trained feature extraction models of CLIP, VST, and DepthAnythingV2 without any additional fine-tuning.

In the MTFF module, we utilize three 3×1 Conv1D kernels for local temporal correlation learning, with mixed temporal reception achieved by various dilation rates (1, 2, and 4). The number of heads in the attention blocks is set to 4. The network is trained by the Adam optimizer with a weight decay of 0.0005 and a batch size of 128 for 1000 epochs. During training, each mini-batch consists of 64 randomly selected normal and 64 abnormal videos, with a learning rate of 0.0001.

Following the literature, the Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) are employed for evaluation on the UCF-Crime dataset [1]. Besides this, the Precision-Recall curve and its corresponding Average Precision (AP) are utilized to evaluate the performance of the proposed models on the XD-Violence dataset [5].

B. Computational Costs

The execution performance of the MTFL video anomaly detection model is measured with a GTX-2080Ti GPU. The feature extraction in video anomaly detection takes 104.96 ms for a video segment of a length of 2.56 s (64 frames) for each temporal granularity (short, medium, long). This results in 104.96 × 3 = 314.88 ms for generating the intermediate feature vector across all temporal granularities, by considering only one GTX-2080Ti GPU. The multiscale feature fusion step, which employs the discussed attention mechanisms, adds an additional 2.43 ms, bringing the total computational time to 317.31 ms. Therefore, the buffer delay (the minimum time required to process a video segment) is 317.31 ms, meaning that the system needs at least this total time to proceed with each 2.56 s video segment before moving to the next one.

In contrast, when comparing the MMTFL model that incorporates depth and textual feature extraction alongside spatio-temporal computational requirements increase. For depth features extracted using the DepthAnythingv2 model, the inference time for 64 frames is approximately 640 ms, while textual features extracted via CLIP require around 320 ms for the same number of frames. When combined with spatiotemporal feature extraction (314.88 ms) and the fusion step, the overall processing time for one video segment rises to 1.27 s. While this represents an increase in per-segment processing time compared to MTFL, the enriched multi-modal information provided by MMTFL leads to significantly improved anomaly detection performance, justifying the additional computational cost in many practical scenarios.

This processing time of MTFL is more reasonable for a real-time surveillance application by recursively refreshing each segment's output. Since feature extraction takes 317.31 ms for a video segment (64 frames), the system processes video at an approximate rate of 201.70 frames per second (fps), which is significantly faster than real-time (30 fps or 25 fps). This also means that the system can handle the incoming video stream without a processing bottleneck.

Similarly, although MMTFL introduces additional computational overhead due to the inclusion of depth and textual features, its total processing time of approximately 1277 ms per segment remains within a near real-time range. Considering that anomaly detection applications often tolerate a latency of couple of seconds without compromising effectiveness, MMTFL's enriched multi-modal approach offers a practical balance between better anomaly detection performance and feasible inference speed in real-world surveillance scenarios.

C. Anomaly Detection Results

Table II compares the proposed MTFL and MMTFL models to SotA anomaly detection methods on the UCF-Crime and XD-Violence datasets. Existing approaches primarily rely on I3D-RGB and recent approaches include CLIP-based features. RGB-only methods extract spatio-temporal information, while CLIP-based models leverage both visual and textual features for enhanced semantic understanding. The method of Sultani *et al.* [1] is added because it is the first paper publishing the UCF-Crime dataset and providing the results by the C3D feature extraction model.

TABLE II. ANOMALY DETECTION PERFORMANCE COMPARISON WITH THE SOTA ON UCF-CRIME (AUC) AND XD-VIOLENCE (AP) DATASETS

Method	Feature	UCF Crime /%	XD Violence /%
Sultani *et al*. [1]	C3D-RGB	75.41	73.20
Wu *et al*. [5]	I3D-RGB	82.44	78.64
RTFM [3]	I3D-RGB	84.30	77.81
Wu *et al*. [56]	I3D-RGB	84.89	75.90
TEVAD [53]	I3D-RGB	84.90	79.80
MGFN [57]	I3D-RGB	86.67	80.11
UR-DMU [58]	I3D-RGB	86.97	81.66
PEL [59]	I3D-RGB	86.76	**85.59**
MSL [60]	VST-RGB	85.62	78.59
MGFN [57]	VST-RGB	86.67	80.11
MTFL	VST-RGB	**87.16 (+0.19)**	84.57
UMIL [61]	CLIP	86.75	-
CLIP-TSA [29]	CLIP	87.58	82.19
TPWNG [62]	CLIP	87.79	83.68
VadCLIP [63]	CLIP	88.02	84.51
MMTFL	CLIP, DEPTH	**88.29 (+0.27)**	84.96

The MTFL model achieves 87.16% AUC on UCF-Crime dataset, notably surpassing the SotA anomaly detection methods, and 84.57% AP on the XD-Violence dataset. The results indicate that the MTFL method demonstrates better capability for identifying complex anomalies with various real-world scenarios presented in the UCF-Crime and XD-Violence datasets. Compared to the methods utilizing the same VST-RGB features, such as MSL [60] and MGFN [57], the MTFL model exhibits a substantial superiority on the UCF-Crime and XD-Violence datasets, manifesting its efficiency regardless of the feature extraction approach. Moreover, contrasted with other weakly-supervised methods using the same loss function, the MTFL method outperforms RTFM [3] by 2.86% AUC on the UCF-Crime and 6.76% AP on the XD-Violence.

The proposed MMTFL model extends beyond the dominant prior methods (I3D, VST, and CLIP) by incorporating depth features alongside CLIP-based text

and RGB representations, which enables a more comprehensive multi-modal feature learning. As a result, the MMTFL model achieves 88.29% AUC on the UCF-Crime dataset, surpassing previous methods, including VadCLIP [63], CLIP-TSA [29], and TPWNG [62]. This improvement highlights the effectiveness of multi-modal feature fusion, particularly the depth modality, which enhances structural understanding beyond text and RGB features. Depth information plays a key role in real-world surveillance, aiding anomaly detection by capturing background elements, object positioning, and motion dynamics.

Although MMTFL remains competitive on XD-Violence with 84.96% AP, it does not surpass the SotA result of 85.59% AP achieved by PEL [59]. A key factor influencing this outcome is the nature of the XD-Violence dataset, which primarily consists of movie clips, differing significantly from real-world surveillance videos in the UCF-Crime dataset. Since movie scenes are well-structured, artificially lit, and edited, depth estimation may not provide as much discriminative information as real-world settings, where depth features help distinguish objects and movement patterns in uncontrolled environments.

These results underscore the importance of multi-modal fusion, demonstrating that integrating spatio-temporal, depth, and text features enhances anomaly detection in complex video scenarios. Although depth features significantly improve performance in real-world surveillance videos (the UCF-Crime dataset), their impact is less pronounced in controlled cinematic settings (the XD-Violence dataset).

The proposed MMTFL advances beyond recent multi-modal fusion methods such as VadCLIP, CLIP-TSA, and PEL in several key aspects. First, MMTFL integrates spatio-temporal, depth, and textual features using a unified and efficient fusion strategy, supported by formalized pseudocode and computational cost analysis to enhance clarity and reproducibility. Unlike prior methods that primarily leverage textual and visual modalities, our architecture explicitly incorporates depth features, leading to richer representation and improved anomaly detection performance. Furthermore, we provide a detailed exploration of different feature fusion designs and empirically demonstrate their impact, offering robust justification for our chosen strategy and contributing practical enhancements for future multi-modal frameworks.

D. Qualitative Results

Fig. 9 presents the anomaly scores generated by the MTFL model for randomly selected anomaly videos, consisting of 1 normal and 7 anomalous videos. The anomalous videos include six from the UCF-Crime dataset, such as Arson016, Shooting004, Robbery102, Burglary037, Fighting033, and Shoplifting004, each accompanied by corresponding anomaly scores and frame numbers. According to the anomaly scores, MTFL efficiently distinguishes between normal and anomalous segments in videos. It performs well even in scenarios where multiple events are occurring in a single video, such as Burglary. Apart from the detection of anomalies with obvious motions, the MTFL demonstrates the ability to discern subtle anomalies that require an understanding of contextual information and behavior, like Burglary and Shoplifting. For more challenging cases, such as frequent throwing actions in CarDangerous024, while MTFL may not isolate each individual event due to limited visibility and distance (it often occurs at a greater distance in car-related dangerous activities), the proposed model successfully identifies periods containing these activities. These qualitative results highlight the model's capability to generalize across diverse anomaly types.

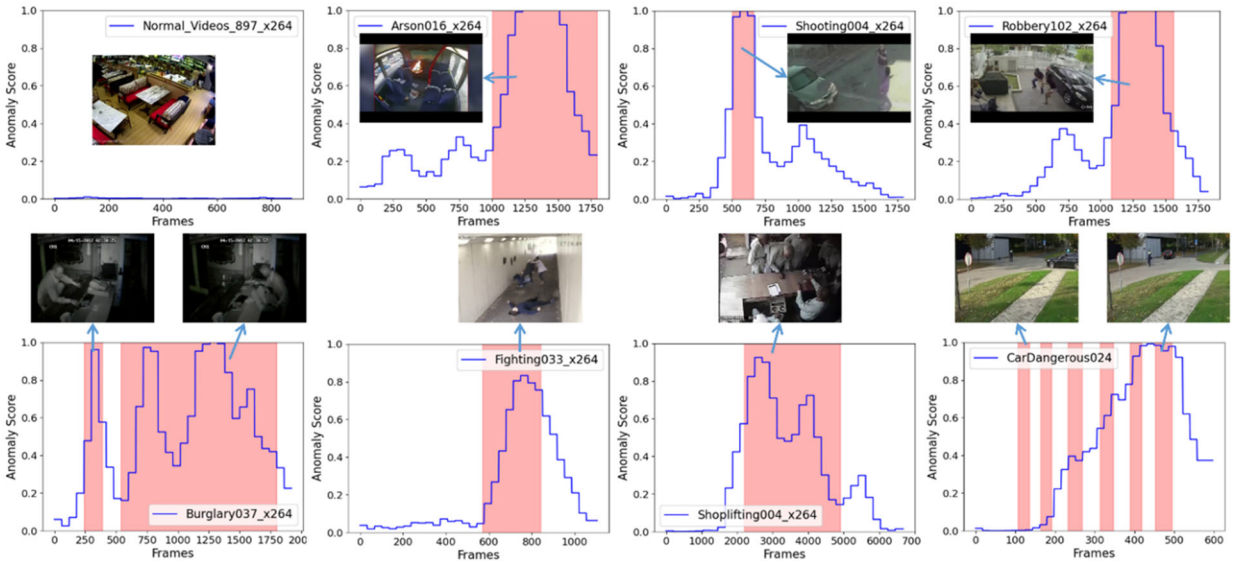

Fig. 9. Anomaly scores of the proposed MTFL model on the UCF-Crime data (Normal897, Arson016, Shooting004, Robbery102, Burglary037, Fighting033, and Shoplifting004) and a dangerous throwing video from the CarDangerous024. The red-colored regions indicate the manually labeled occurrences of anomalous events.

VI. Ablation Studies

A. Best Feature Extraction for Anomaly Detection

To explore the impact of different feature extraction models on the performance of anomaly detection, we examine five feature extraction models: C3D [64], I3D [65], SlowFast [9], VST [55], and MViTv2 [66] with RTFM method [3]. Table Ⅲ shows the AUC results of anomaly detection using RTFM with these five video feature extractors on the UCF-Crime dataset. It should be noted that the AUC score of MTFL model does not exist in this table because these experiments are conducted with RTFM method to find out the better feature extraction model. Based on the results, the VST model outperforms other feature extractors. When trained on the UCF-Crime dataset, the RTFM with VST results in 85.99% AUC on the UCF-Crime test set.

TABLE III. Comparison of the RTFM Method, Employing Five Feature Extractors Trained on the UCF-Crime (AUC) Dataset

Training Data	Architecture	UCF-Crime/%
UCF-Crime	C3D + RTFM	72.27
	I3D + RTFM	82.63
	SlowFast + RTFM	81.75
	MViTv2 + RTFM	84.13
	VST + RTFM	**85.99**

B. Efficiency of Internal Multi-Timescale Feature Fusion (MTFF) Stages

Referring to Fig. 3, the architecture of the designed MTFF module comprises four stages: Pairwise Feature Learning (PFL), Local Temporal correlation Learning (LTL), Global Temporal correlation Learning (GTL), and Feature Fusion (FF). To investigate the effects of these stages, we conduct experiments between different MTFF designs by removing different stages, as shown in Table IV. The results reveal that the most significant performance degradation is due to the removal of PFL and FF stages. This leads to noticeable drops in results on the UCF-Crime. The LTL and GTL stage removals result in relatively minor deterioration. These comparisons show that mutual learning of features extracted from spatio-temporal tubes of different frame lengths, coupled with feature fusion, enhances feature representation capability with the help of PFL and FF. Further enhancement in performance is achieved by incorporating local and global correlation learning to scale features within LTL and GTL. Overall, these experiments validate the impact of each stage in the MTFF module.

TABLE IV. Ablation Studies on Four Modules of MTFF

PFL	LTL	GTL	FF	UCF-Crime/%
	✓	✓	✓	83.72
✓		✓	✓	86.23
✓	✓		✓	87.08
✓	✓	✓		85.43
✓	✓	✓	✓	**87.16**

C. Optimal Combination of Tubelets

We extract features using 4 different tubelets with frame lengths of 8, 16, 32, and 64, respectively (Table V). By selecting three tubelet lengths as inputs to the MTFF module, we compare detection results for various length combinations as shown in Table V. Notably, the best detection results are achieved when the short tubelet length is 8 frames, the medium tubelet length is 32 frames, and the long tubelet length is 64 frames. This observation suggests that combining the 8-frame feature for capturing rapid motion details, the 64-frame feature for contextual information, and the 32-frame feature for correlation information improves the understanding of anomalous events within the MTFL method.

TABLE V. Ablation Studies on Different Tubelet Lengths

8	16	32	64	UCF-Crime/%
	✓	✓	✓	86.87
✓		✓	✓	**87.16**
✓	✓		✓	86.74
✓	✓	✓		86.71

D. Feature Modalities Contribution for Anomaly Detection

To assess the impact of different feature modalities on anomaly detection, an ablation study is conducted on the UCF-Crime dataset, evaluating the individual contributions of RGB, depth, and text features. Different feature combinations are evaluated to assess their individual and combined impact, of which the results are shown in Table VI.

TABLE VI. Ablation Study for the Obtained AUC Value on the UCF-Crime Dataset, Using up to Three Different Feature Types (Spatio-Temporal RGB, Depth, and Text) in the MMTFL Model

RGB	Depth	Text	UCF-Crime/%
✓	–	–	87.16
✓	✓	–	87.69
✓	–	✓	87.58
✓	✓	✓	**88.29**

The baseline MTFL model achieves 87.16% AUC, utilizing only RGB features. Introducing depth features alone improves the performance to 87.69% AUC, indicating that structural scene information captured by depth plays an important role in detecting anomalies. Similarly, replacing depth with text-based features results in 87.58% AUC, suggesting that semantic context extracted from textual descriptions also contributes to the performance. The highest 88.29% AUC is achieved by integrating all three modalities (RGB, depth, and text features), which demonstrates that multi-modal fusion enhances feature representation and anomaly differentiation. These results highlight the complementary nature of depth and text information, where depth facilitates in understanding spatial relationships, and text provides semantic meaning.

E. Comparison of Feature Fusion Strategies

To evaluate the impact of three feature fusion strategies

given in Figs. 6–8 (Section IV) is compared on the UCF-Crime dataset. Table VII presents the AUC performance of these feature fusion strategies on the UCF-Crime dataset. Among the evaluated designs, Feature Fusion 2 achieves the highest performance with 88.29% AUC, demonstrating that averaging depth and text features prior to concatenation, while processing spatio-temporal features separately with MTFF feature fusion, leads to better anomaly detection. Feature Fusion 3 follows closely with 87.92% AUC, where all three modalities are independently fused via MTFF module before concatenation. This indicates that direct fusion still captures useful cross-modal relationships but may not be as effective as pre-averaging certain modalities.

TABLE VII. PERFORMANCE COMPARISON OF DIFFERENT FEATURE FUSION STRATEGIES, MEASURED BY AUC METRIC, IN THE PROPOSED MMTFL MODEL ON THE UCF-CRIME DATASET

Experiment	Strategy	UCF-Crime/%
Feature Fusion1	MTFF (Concat (rgb, depth, text))	87.66
Feature Fusion2	Concat (MTFF (rgb), Avg (depth), Avg (text))	**88.29**
Feature Fusion3	Concat (MTFF (rgb), MTFF (depth), MTFF (text))	87.92

Moreover, Feature Fusion 1 records the lowest performance at 87.66% AUC, suggesting that early concatenation of all modalities before MTFF fusion may not optimally exploit complementary information across temporal scales. These results highlight the importance of evaluating different strategies for feature integration and show that selective pre-processing and adaptive weighting of different modalities can significantly impact anomaly detection performance.

VII. CONCLUSION

This paper proposes the MTFL model that leverages multiple temporal scales to understand behavior anomalies in videos, enabling a powerful fusion of motion details and event feature information for anomaly detection. The MTFL model achieves leading results on the UCF-Crime with 87.16% AUC and on the XD-Violence dataset with 84.57% AP. It shows high performance in detecting anomalies without distinct motion patterns due to the advanced time-scale partitioning. The MTFL model captures fine-grained motion details and long-term contextual dependencies through multi-timescale tubelet analysis, extracting features at different short, medium, and long temporal resolutions.

Moreover, the MMTFL model has been proposed to overcome the limitations of single-modal feature approaches, integrating spatio-temporal, depth, and text-based features. MMTFL achieves 88.29% AUC on the UCF-Crime dataset, surpassing previous methods, and 84.96% AP on the XD-Violence dataset. This approach improves structural understanding in real-world surveillance and ensures a complete representation of motion dynamics, enhancing the model's ability to distinguish normal and anomalous events in complex video scenarios.

The experimental results emphasize the importance of multi-modal fusion, integrating spatio-temporal, depth, and text-based features. The proposed models employ concatenation, averaging, multi-cross attention and self-attention mechanisms to jointly exploit complementary information across modalities and align multiple temporal domains. Moreover, it is important to note that incorporating additional modalities enhances the performance of the anomaly detection model. While the feature extraction cost increases slightly, it remains within a range suitable for real-world applications.

Limitations and Future Work: While our multi-modal anomaly detection framework demonstrates promising results, several key limitations remain. The current solution leverages pre-trained feature extraction models, which may not optimally capture domain-specific details without targeted fine-tuning. The availability and quality of diverse, well-annotated datasets are vital for further improving generalization and robustness. For future work, we plan to explore student–teacher learning mechanisms to transfer knowledge efficiently between large visual models. Additionally, we aim to address privacy concerns and optimize the approach for deployment on the resource-constrained platforms (e.g., edge devices), to ensure safe and scalable use in real-world applications

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Erkut Akdag has implemented the methods, conducted the experiments and evaluations, and contributed to the underlying concepts as well as the manuscript preparation. Henk Corporaal, Peter H. N. De With, and Egor Bondarev have devised and supervised the research, provided suggestions and recommendations, coordinated the writing process, and ensured the coherence and clarity of the final manuscript. All authors have approved the final version.

FUNDING

This work is supported by the European ITEA project SMART on intelligent traffic flow systems and the NWO Efficient Deep Learning (EDL) RMR project.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 6479–6488.

[2] H. Lv *et al*., "Localizing anomalies from weakly-labeled videos," *IEEE Trans. Image Process*., vol. 30, pp. 4505–4515, 2021.

[3] Y. Tian *et al*., "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2021, pp. 4975–4986.

[4] M. A. Carbonneau, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit*., vol. 77, pp. 329–353, 2018.

[5] P. Wu *et al*., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Comput. Vis.—ECCV*, 2020, pp. 322–339.

[6] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly

detection," in *Proc. IEEE Int. Conf. Image Process*., 2019, pp. 4030–4034.
[7] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," arXiv Print, arXiv:1907.10211, 2019. doi.org/10.48550/arXiv.1907.10211
[8] O. Hirschorn and S. Avidan, "Normalizing flows for human pose anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2023, pp. 13545–13554.
[9] V. Sharma *et al*., "Video processing using deep learning techniques: A systematic literature review," *IEEE Access*, vol. 9, pp. 139489–139507, 2021.
[10] C. Feichtenhofer *et al*., "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2019, pp. 6202–6211.
[11] K. He *et al*., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 770–778.
[12] A. Radford *et al*., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn*., 2021, pp. 8748–8763.
[13] C. Szegedy *et al*., "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell*., vol. 31, no. 1, 2017.
[14] Y. Hu *et al*., "Hierarchical global-local temporal modeling for video captioning," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 774–783.
[15] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 8327–8336.
[16] Z. Zhang *et al*., "Object relational graph with teacher-recommended learning for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 13278–13288.
[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[18] A. Vaswani *et al*., "Attention is all you need," in *Adv. Neural Inf. Process. Syst*., 2017.
[19] N. Carion *et al*., "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis*., 2020, pp. 213–229.
[20] R. Girdhar *et al*., "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 244–253.
[21] J. Lei *et al*., "TVQA: Localized, compositional video question answering," arXiv Print, arXiv:1809.01696, 2018. doi.org/10.48550/arXiv.1809.01696
[22] J. Lei *et al*., "TVR: A large-scale dataset for video-subtitle moment retrieval," in *Comput. Vis.—ECCV*, 2020, pp. 447–463.
[23] Z. Liu *et al*., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2021, pp. 10012–10022.
[24] S. Zheng *et al*., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2021, pp. 6881–6890.
[25] K. Lin *et al*., "SwinBERT: End-to-end transformers with sparse attention for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 17949–17958.
[26] P. H. Seo *et al*., "End-to-end generative pretraining for multimodal video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 17959–17968.
[27] H. Ye *et al*., "Hierarchical modular network for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 17939–17948.
[28] Z. Qing *et al*., "Learning from untrimmed videos: Self-supervised video representation learning with hierarchical consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 13821–13831.
[29] H. K. Joo *et al*., "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection," in *Proc. IEEE Int. Conf. Image Process*., 2023, pp. 3230–3234.
[30] X. Xu *et al*., "Towards robust video object segmentation with adaptive object calibration," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 2709–2718.
[31] T. Yuan *et al*., "Towards surveillance video-and-language understanding: New dataset baselines and challenges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2024, pp. 22052–22061.
[32] D. Xu, Y. Ricci, Y. Yan, J. Song, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand*., vol. 156, pp. 117–127, 2017.
[33] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," arXiv Print, arXiv:2306.02858, 2023. doi.org/10.48550/arXiv.2306.02858
[34] Z. Li *et al*., "Temporally consistent online depth estimation in dynamic scenes," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis*., 2023, pp. 3018–3027.
[35] Z. Teed and J. Deng, "DeepV2D: Video to depth with differentiable structure from motion," arXiv Print, arXiv:1812.04605, 2018. doi.org/10.48550/arXiv.1812.04605
[36] C. Wang *et al*., "Web stereo video supervision for depth prediction from dynamic scenes," in *Proc. Int. Conf. 3D Vis*., 2019, pp. 348–357.
[37] Y. Wang *et al*., "Less is more: Consistent video depth estimation with masked frames modeling," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 6347–6358.
[38] Y. Wang *et al*., "Neural video depth stabilizer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2023, pp. 9466–9476.
[39] R. Yasarla *et al*., "MAMO: Leveraging memory and attention for monocular video depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2023, pp. 8754–8764.
[40] R. Yasarla *et al*., "FutureDepth: Learning to predict the future improves video depth estimation," in *Proc. Eur. Conf. Comput. Vis*., 2024, pp. 440–458.
[41] H. Zhang *et al*., "Exploiting temporal consistency for real-time video depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2019, pp. 1725–1734.
[42] Y. Chen, C. Schmid, and C. Sminchisescu, "Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2019, pp. 7063–7072.
[43] J. Kopf, X. Rong, and J. B. Huang, "Robust consistent video depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2021, pp. 1611–1621.
[44] X. Luo *et al*., "Consistent video depth estimation," *ACM Trans. Graph*., vol. 39, no. 4, pp. 71:1–71:13, 2020.
[45] Z. Zhang *et al*., "Consistent depth of moving objects in video," *ACM Trans. Graph*., vol. 40, no. 4, pp. 1–12, 2021.
[46] W. Barrios *et al*., "Localizing moments in long video via multimodal guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2023, pp. 13667–13678.
[47] X. Lan *et al*., "A survey on temporal sentence grounding in videos," *ACM Trans. Multimedia Comput., Commun., Appl*., vol. 19, no. 2, pp. 1–33, 2023.
[48] M. Liu *et al*., "A survey on video moment localization," *ACM Comput. Surv*., vol. 55, no. 9, pp. 1–37, 2023.
[49] H. Zhang *et al*., "Temporal sentence grounding in videos: A survey and future directions," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 45, no. 8, pp. 10443–10465, 2023.
[50] M. Abdar *et al*., "A review of deep learning for video captioning," *IEEE Trans. Pattern Anal. Mach. Intell*., 2024.
[51] S. Li *et al*., "Visual to text: Survey of image and video captioning," *IEEE Trans. Emerg. Topics Comput. Intell*., vol. 3, no. 4, pp. 297–312, 2019.
[52] R. Krishna *et al*., "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis*., 2017, pp. 706–715.
[53] W. Chen *et al*., "TEVAD: Improved video anomaly detection with captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2023, pp. 5549–5559.
[54] X. Wang *et al*., "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2019, pp. 4581–4591.
[55] Z. Liu *et al*., "Video swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2022, pp. 3202–3211.
[56] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3513–3527, 2021.
[57] Y. Chen *et al*., "Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell*., vol. 37, no. 1, 2023, pp. 387–395.

[58] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell*., vol. 37, no. 3, 2023, pp. 3769–3777.
[59] Y. Pu, X. Wu, and S. Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," arXiv Print, arXiv:2306.14451, 2023. doi.org/10.48550/arXiv.2306.14451
[60] S. Li *et al*., "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell*., vol. 36, no. 2, 2022, pp. 1395–1403.
[61] H. Lv *et al*., "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2023, pp. 8022–8031.
[62] Z. Yang, J. Liu, and P. Wu, "Text prompt with normality guidance for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2024, pp. 18899–18908.
[63] P. Wu *et al*., "Vadclip: Adapting vision language models for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell*., vol. 38, no. 6, 2024, pp. 6074–6082.
[64] D. Tran *et al*., "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis*., 2015, pp. 4489–4497.
[65] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2017, pp. 6299–6308.
[66] Y. Li *et al*., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4804–4814.