Spatio-Temporal Transformers and Semantic Insights: Redefining Video Anomaly Detection

Amara Sai Prasad[a], Nidhi P G[a], Nikita Suresh[a], *Ria R Kulkarni[a], and Ramamoorthy Srinath[a]

[a]PES University, Bangalore, India

ABSTRACT

Traditional video surveillance systems often rely on manual monitoring, resulting in labor-intensive processes that are prone to errors, particularly in detecting complex events that hinder public safety and security. This study addresses these limitations by integrating advanced deep learning models for comprehensive video and text feature extraction, leveraging the power of multi-modal fusion. We employ TimeSformer to capture intricate spatio-temporal patterns within video data and RoBERTa to extract semantic insights from accompanying text captions. Our approach utilizes three multi-modal fusion techniques: Concatenation Fusion, Gated Fusion, and Compact Bilinear Pooling. These methods effectively combine visual and textual representations to enhance anomaly detection capabilities. Experimental results reveal that Concatenation Fusion significantly outperforms the other methods, achieving an accuracy of 85.19% in complex video scenarios. This fusion-based approach not only improves detection accuracy but also reduces the reliance on continuous human oversight, making it a practical solution for applications in public safety and security.

Keywords: anomaly detection, multimodal feature fusion, video surveillance, TimeSformer, RoBERTa, gated fusion, concatenation fusion

1. INTRODUCTION

In recent years, video surveillance has become a crucial component in maintaining public safety, with systems deployed in various settings such as public transportation hubs, commercial spaces, and urban environments. These systems play a vital role in ensuring security, managing crowds, and enabling rapid response to emergencies. However, traditional surveillance systems often depend heavily on human operators to monitor and interpret video feeds, which introduces challenges such as fatigue, delayed response times, and missed incidents. These limitations emphasize the need for automated systems capable of identifying potential threats in real time.

Our research proposes a multimodal anomaly detection system that combines advanced technologies for video and text analysis. TimeSformer is employed for extracting spatio-temporal features from video content, while RoBERTa processes textual captions generated from these videos. This integration of visual and textual data enhances the system's capability to identify complex behaviors such as vandalism, violence, and the presence of abandoned objects, which are typically difficult to detect using traditional approaches. To achieve this, the study explores various fusion strategies designed to combine video and text features effectively, ensuring a comprehensive understanding of the monitored environment. This approach has broad applicability across various domains. In public spaces and urban environments, the system can assist in detecting unusual or suspicious activities, improving safety and security. In transportation hubs, it can help identify abandoned objects or crowd disturbances, while in industrial and commercial spaces, it can monitor for unauthorized access or hazardous incidents. Moreover, its ability to analyze multimodal data makes it a valuable tool in healthcare for detecting emergencies such as falls or unusual behaviors in elder care settings.

By reducing dependence on continuous human monitoring and enabling real-time detection, this system addresses critical gaps in traditional surveillance practices. The integration of multimodal analysis into surveillance systems not only enhances the detection of complex and context-dependent anomalies but also represents a step

*riakulkarni25@gmail.com, +91 8861938336

forward in creating safer, more secure environments. Our experimental results reveal that Concatenation Fusion, with video data as the first modality, achieves the highest accuracy across key metrics, making it an effective approach for real-time anomaly detection in video surveillance applications.

2. RELATED WORK

Anomaly detection in video surveillance has long been an area of active research. The initial methods predominantly used hand-crafted features and statistical models to identify deviations from normal patterns, commonly applied to tasks such as detecting abandoned objects or unusual behavior of the *crowd*[1]. Although effective for specific use cases, these approaches often struggled with scalability and were limited in their ability to simultaneously detect multiple anomalies.

The advent of deep learning brought significant advances in anomaly detection, with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) becoming foundational components of modern approaches. These models are particularly effective in detecting single event anomalies by learning complex spatial and temporal patterns directly from video *data*[2],[3]. However, most of these methods focus on detecting a single type of anomaly, which poses limitations in real-world applications where multiple anomalies can occur concurrently. Moreover, as video data becomes more complex, there is a growing need for models that can process this information more efficiently and accurately.

To address this complexity, attention-based architectures like Vision Transformers (ViT)[11] and $TimeSformer$[12] have recently emerged as powerful tools for video analysis. ViT operates by breaking down images into patches and modeling relationships between these patches using self-attention mechanisms. This allows ViT to effectively capture spatial information in surveillance footage. On the other hand, TimeSformer extends this idea to video data by modeling both spatial and temporal dimensions through divided attention. These models have shown great potential in video-based anomaly detection, offering complex spatio-temporal modeling that traditional CNNs and RNNs often lack.

Building on ViT's role in video analysis, research conducted by Singh et *al.*[15] presents a framework for detecting violence in videos using Video Vision Transformers (ViViT). Addressing the need for real-time, accurate surveillance, the authors focus on datasets like the Hockey Fight and Violent Crowd datasets to train a ViViT model for classifying video clips as violent or nonviolent. By transforming video data into sequences of frames and applying augmentations such as Gaussian blur, the study achieves high accuracy across different types of violent scenarios.

Using a ViViT model and tuning hyper-parameters such as batch size and learning rate, this approach shares similarities with FTP[13] and $VAD - LLaMA$[14] in utilizing ViTs for specialized video tasks. With a precision and recall of 0.99 on violent crowd datasets, the framework demonstrates ViTs' capability in both structured action recognition and high-stakes anomaly detection, making it valuable for scalable, real-time applications in public safety.

Lu, H et. *al*[13] tackle the challenge of generalization in Vision Transformers (ViTs) for video action understanding, particularly when applied across diverse datasets with distinct action characteristics, such as object usage and environmental context. To address these challenges, the authors propose the Four-Tiered Prompts (FTP) framework, which leverages the contextual richness of Visual Language Models (VLMs) to improve ViT performance. The FTP framework introduces prompts based on four aspects of human actions—action category, components, description, and context—to refine spatio-temporal embeddings in the ViT model.

The FTP framework includes an augmented ViT architecture with feature processors that align visual embeddings with textual descriptions generated by VLMs. The training process consists of two stages: aligning visual and textual features through the feature processors, followed by fine-tuning for action classification. Benchmarked on datasets like Kinetics-400 and AVA V2.2, FTP significantly outperforms prior models in action recognition, achieving a top-1 accuracy of 93.8% on Kinetics-400. This indicates the power of combining ViTs with VLM-based prompts to overcome dataset bias and improve generalization.

Parallel to advancements in computer vision, the development of large language models (LLMs) such as GPT-3 and BERT has revolutionized semantic understanding in various domains. These models exhibit remarkable capabilities in understanding and generating human-like text, excelling at tasks like sentiment analysis, text $summarization^4$ and question $answering^5$.

The combination of LLMs and video analysis introduces an additional semantic layer, enhancing the interpretability of detected anomalies. For example, LLMs can generate textual descriptions of detected anomalies, offering a more detailed and context-aware interpretation of events captured in video $footage^6$. This is particularly valuable when paired with vision models like ViT and TimeSformer, as LLMs can provide a semantic understanding that complements the spatial and temporal insights extracted from video data. This integration helps bridge the gap between raw visual data and human-readable interpretations, addressing the shortcomings of purely visual approaches that often struggle with context-based anomaly detection.

H. Lv et. al's $study^{14}$ builds on the idea of augmenting video-based tasks with large language models, specifically for Video Anomaly Detection (VAD). The proposed VAD-LLaMA framework integrates a Video-based Large Language Model (VLLM) to improve anomaly detection and provide contextual explanations, which traditional threshold-based methods lack. To capture both normal and abnormal contexts effectively, VAD-LLaMA employs a Long-Term Context (LTC) module that preserves temporal information across video clips, enhancing the ability to distinguish anomalies over time.

Like the FTP $framework^{13}$ in action understanding, VAD-LLaMA involves a multi-stage training approach, where an initial VAD generates anomaly scores, then co-trains with the LTC module to capture long-term contextual relationships. Evaluated on benchmark datasets such as UCF-Crime, the model achieved a 3% increase in AUC, significantly improving anomaly detection accuracy and localization, demonstrating that integrating VLMs with video models not only enhances detection performance but also provides robust explanatory power for anomaly identification.

Expanding on VLM-augmented models, the $LAVILA^{16}$ framework demonstrates how large language models can enhance video-language representations through multi-stage narration. By training a GPT-2 based model (NARRATOR) to produce video narrations and a T5 model (REPHRASER) for paraphrasing, LAVILA enhances data diversity in video-text pairs, supporting models like FTP^{13} and $VAD - LLaMA^{14}$ that rely on rich, contextual text alignments.

The study combines a video encoder and text encoder using an InfoNCE loss, creating a robust video-language representation. Benchmarks indicate improved performance on action classification tasks, and the model's effectiveness with multimodal data aligns with findings in anomaly detection, where VLLMs help improve contextual understanding and classification accuracy.

The fusion of multimodal data, such as text, video and audio, has gained significant traction in recent years as a means of enhancing model robustness and accuracy. In video surveillance, multimodal fusion allows systems to leverage both visual and textual information, improving the precision of anomaly detection.

Various fusion techniques exist, including early fusion, where features from different modalities are combined before feeding into a model; along with late fusion, where predictions from each modality are $merged^8$. Concatenation facilitates the preservation of the original feature space, enabling the model to learn from the raw characteristics of each modality without the potential loss of information that might occur in other fusion techniques. In the Multimodal Machine Learning $Survey^{18}$, concatenation is described as one of the methods of early fusion, where features from different modalities are combined into a single representation.

More advanced approaches, such as attention mechanisms and transformers, have been particularly effective in capturing complex interactions between modalities, enhancing both the accuracy and robustness of anomaly detection $systems^9$. In the realm of multimodal learning, Arevalo et $al.^{19}$ present a significant advancement through the introduction of the Gated Multimodal Unit (GMU). This model is designed to effectively integrate information from multiple modalities, such as text and images, by employing gated neural networks that learn to modulate the influence of each modality on the unit's activation. The GMU utilizes multiplicative gates to dynamically determine how much each input modality contributes to the overall output. This adaptive mechanism allows for a more nuanced representation of multimodal data compared to traditional fixed fusion strategies. The effectiveness of the GMU was demonstrated in a multi-label movie genre classification task, where it outperformed both single-modality and other fusion-based approaches when considering both plot and poster information.

This multimodal fusion has the potential to create more comprehensive surveillance systems, capable of detecting anomalies that may not be apparent through a single modality. For example, textual descriptions from LLMs could be combined with the visual insights provided by ViT and TimeSformer to offer a more holistic view of surveillance footage, increasing the system's ability to detect complex, context-dependent anomalies.

To tackle feature integration challenges in neural networks, Dai et *al.*[17] in Attentional Feature Fusion introduced a framework that leverages attention mechanisms to address inconsistencies in scale and semantic meaning among input features. AFF employs multi-scale channel attention mechanisms that unify local and global contexts for feature integration, significantly improving performance across various tasks, including image classification and semantic segmentation.

This study complements the discussions in previous studies by demonstrating the efficacy of attention mechanisms in enhancing feature fusion, applicable across different architectures, from CNNs to ViTs. By optimizing feature fusion, AFF provides a pathway to more effective multimodal integration, as seen in models like $VAD-LLaMA$[14], where enhanced feature alignment is critical for anomaly detection. This work underscores the importance of attention mechanisms, offering insights into how future anomaly detection and action recognition frameworks can further optimize feature integration for enhanced performance.

The literature reveals a growing need for an integrated, multimodal approach to anomaly detection in video surveillance—one that combines the strengths of computer vision, LLMs and attention-based models like ViTs and TimeSformer. While significant progress has been made in detecting single-event anomalies, the simultaneous detection of multiple events remains a critical challenge. The proposed framework aims to address this challenge by combining the semantic understanding of LLMs with the spatial and temporal capabilities of the TimeSformer model. This approach has the potential to enhance the accuracy and robustness of surveillance systems, thereby contributing to the safety and security of public spaces.

3. DATASET

In this study, we introduce a new dataset specifically designed for multimodal anomaly detection in video surveillance. Unlike existing datasets, which often lack comprehensive annotations, our dataset includes both video and corresponding textual captions across four categories: vandalism, violence, abandoned objects and normal scenarios. This dataset was curated from a blend of publicly available sources and original video recordings, ensuring a broad and balanced class representation of real-world scenarios that present both visually and contextually distinct patterns.

VIDEX (Video Dataset for Anomaly Exploration), our novel dataset, consists of two key components: video data and corresponding textual data in the form of captions.

Vandalism Class: The data for the vandalism category is derived from the DCSASS *dataset*[20], which includes video samples that capture incidents of burglary, shoplifting, and vandalism.

Violence Class: For the violence class, videos were selected from the SCVD (Street Crime Video Database)[21] and RLVS (Real-Life Violence Situations)[22] datasets. These datasets encompass numerous real-life street fight scenarios filmed across diverse environments and conditions, allowing for a comprehensive examination of violent behavior in public settings.

Normal Scenarios Class: The normal scenarios are represented by video samples extracted from the $UCF-Crime$[20] dataset, specifically focusing on non-anomalous activities. This inclusion is essential for establishing a baseline against which anomalous behaviors can be identified.

Abandoned Object Class: The abandoned object category presented unique challenges due to the scarcity of publicly available labeled data. To address this, data was compiled from the ABODA *dataset*[23], supplemented by samples sourced from YouTube. Additionally, original footage was recorded by simulating abandoned object situations in various locations. Each of these videos lasts between 25-30 seconds, adhering to the research findings that suggest an object must remain unattended for at least 30 seconds to be classified as abandoned.

The captions for the videos in this dataset were generated using the Pegasus model from Twelve Labs, a state-of-the-art video-to-text generation framework. This model integrates visual and auditory information to produce detailed descriptive text that accurately reflects the content of the videos, including any anomalies present. Custom prompts were crafted based on the specific class of each video—abandoned objects, vandalism, violence and normal scenarios. These prompts guided the Pegasus model in generating captions that focus on providing a comprehensive description of the video content. Emphasis was placed on detailing the actions depicted in the videos and identifying any anomalous behaviors or events.

4. PROPOSED METHODOLOGY

The proposed methodology is depicted in Figure 1, highlighting the main components and the overall workflow. Figure 2 provides a representative diagram illustrating an example that demonstrates the functioning of the entire pipeline.

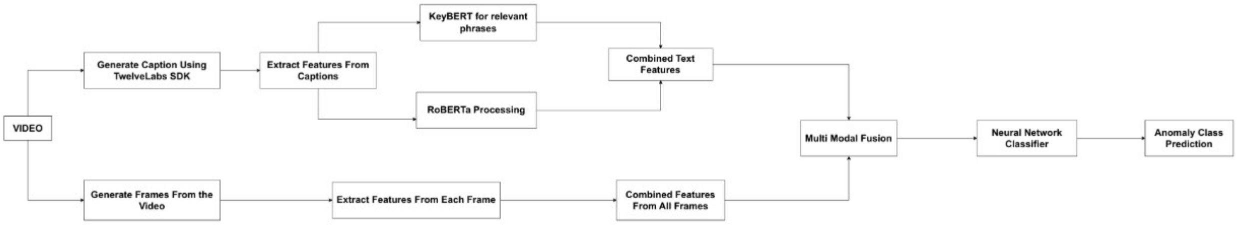
VIDEO
Generate Caption Using TwelveLabs SDK
Extract Features From Captions
KeyBERT for relevant phrases
RoBERTa Processing
Combined Text Features
Generate Frames From the Video
Extract Features From Each Frame
Combined Features From All Frames
Multi Modal Fusion
Neural Network Classifier
Anomaly Class Prediction

Figure 1. Proposed Methodology

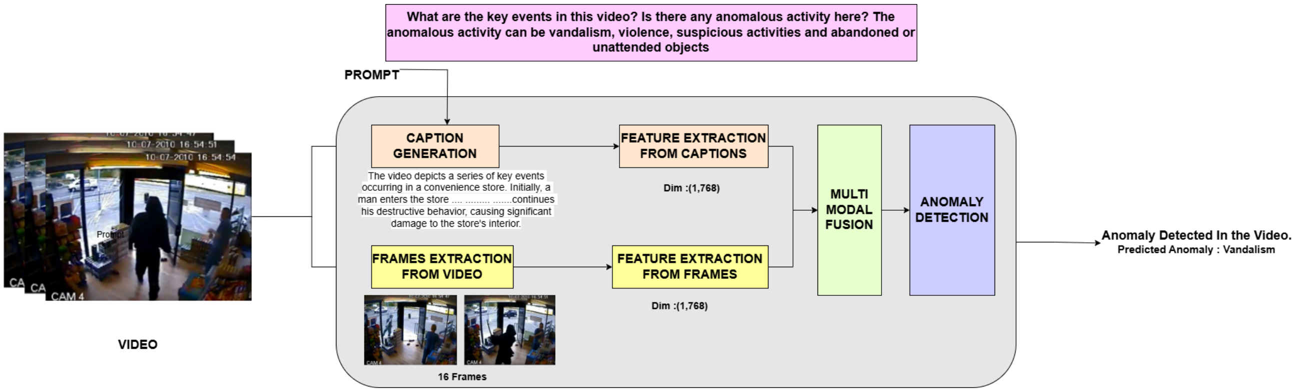
What are the key events in this video? Is there any anomalous activity here? The anomalous activity can be vandalism, violence, suspicious activities and abandoned or unattended objects
PROMPT
CAPTION GENERATION
The video depicts a series of key events occurring in a convenience store. Initially, a man enters the store ....continues his destructive behavior, causing significant damage to the store's interior.
FEATURE EXTRACTION FROM CAPTIONS
Dim :(1,768)
FRAMES EXTRACTION FROM VIDEO
16 Frames
FEATURE EXTRACTION FROM FRAMES
Dim :(1,768)
MULTI MODAL FUSION
ANOMALY DETECTION
Anomaly Detected In the Video.
Predicted Anomaly : Vandalism
VIDEO

Figure 2. An illustrative example demonstrating the functioning of the proposed pipeline, showcasing the flow and interaction between the key components.

4.1 Feature Extraction From Videos

Video feature extraction is the most crucial and usually, the first step for understanding and analyzing video content, particularly in applications like anomaly detection. By extracting meaningful features from videos, models can identify patterns and behavior that deviate from the norm, enabling the detection of events. Effective feature extraction allows the model to focus on relevant information while reducing noise, thereby improving the accuracy of classification tasks. This process involves capturing both spatial and temporal dynamics within the video, which is essential for distinguishing between normal scenarios and anomalies.

TimeSformer, a specialized transformer model for video understanding, effectively captures complex interactions within video sequences by leveraging its self-attention mechanism across space and time. This enables it to learn rich spatiotemporal representations directly from video data, making it particularly suited for tasks like action recognition and anomaly detection. By handling longer video clips than traditional methods, TimeSformer plays a crucial role in extracting meaningful features from video data, especially for recognizing complex activities that unfold over time.

The TimeSformer model treats videos as sequences of frames, which are further divided into patches. The specific variant being utilized here is the 'timesformer-hr-finetuned-k400', pre-trained on the Kinetics-400 dataset. This model processes 16 frames from each video, necessitating a method to ensure uniform frame extraction across varying video lengths.

To achieve this, we developed a structured methodology for frame extraction. Each video is read to access its individual frames, during which we compute the frames per second (FPS) and the total number of frames (frame count) to ascertain the video's duration. Frames are then extracted at regular intervals based on the video's length, allowing us to collect exactly 16 frames per video.

Once extracted, these frames undergo resizing and normalization transformations to meet the input specifications of the TimeSformer model. This approach guarantees uniform feature extraction across videos of different lengths, enabling robust representation generation for each video in the dataset.

The extracted frames are subsequently divided into non-overlapping patches of a specified size (16x16 pixels). Each extracted patch is then flattened and linearly transformed into a vector representation. This transformation includes positional encoding to retain spatial information about where each patch originates within the frame.

The model employs a divided space-time attention mechanism, applying spatial and temporal attention separately to learn dependencies across both dimensions. The spatial attention focuses on relationships between patches within the same frame, allowing the model to understand how different parts of an image interact with one another. Conversely, the temporal attention captures dependencies across frames, enabling the model to learn how actions evolve over time.

The output from the model consists of hidden states from the TimeSformer encoder, which represent feature-rich embeddings of the video frames. After passing through the TimeSformer model, we collect these hidden states corresponding to each frame. To generate a single feature vector for the entire video, we average these hidden states across the temporal dimension (i.e., across all 16 frames). This aggregation effectively captures essential characteristics of the video over time.

4.2 Feature Extraction From Captions

The text feature extraction module is designed to process, analyze and embed textual data from video captions to provide contextual cues in anomaly detection. The process begins with text preprocessing, where captions are tokenized and cleaned to remove stop words using NLTK's stopword library. This ensures that only the most relevant and meaningful tokens are retained, enhancing the quality of extracted features. This preprocessing pipeline standardizes the text data by converting it to lowercase, tokenizing and removing noise, which is particularly useful for reducing dimensionality and improving model interpretability.

Given a caption as a sequence of words, $T = \{w_1, w_2, \ldots, w_n\}$, we filter out common words or "stop words" to retain only meaningful terms. For embedding, the module leverages the RoBERTa model, a transformer-based language model known for its effective contextual representation of text. The RoBERTa tokenizer converts each processed caption into tokenized input suitable for the model. Textual features are extracted as embeddings by feeding these tokenized inputs through RoBERTa's model layers, producing a last hidden state output for each token in the text.

By passing T through the RoBERTa model, we obtain a sequence of hidden states, $H = \{h_1, h_2, \ldots, h_n\}$, for each token w_i in T, where each hidden state h_i has a dimensionality of 768. These embeddings capture nuanced, context-aware information about each word in the caption. To create a single embedding vector representing the entire caption, we average the hidden states across all tokens:

$$e_{\text{RoBERTa}} = \frac{1}{n} \sum_{i=1}^{n} h_i$$

This embedding captures semantic nuances and syntactic structure, enhancing the model's ability to differentiate between captions based on context. The module further applies KeyBERT, a keyword extraction method based on the RoBERTa model, to extract the most salient phrases within each caption. The extract_keywords function in KeyBERT is configured to retrieve up to 5 top keywords or key phrases, limited to unigrams and bigrams, providing a concise representation of essential terms. These keywords are instrumental in capturing context-specific concepts within the video.

Using RoBERTa embeddings, KeyBERT ranks these phrases based on their relevance score, selecting keywords $\{p_1, p_2, \ldots, p_k\}$ that best summarize the input text.

$$T : p_1, p_2, \ldots, p_k = KeyBERT(T)$$

Topic modeling is then applied using LDA (Latent Dirichlet Allocation)[24] which organizes these processed captions into a set number of topics (in this case, four, corresponding to 'violence,' 'vandalism,' 'abandoned object' and 'normal'). LDA utilizes a document-term matrix constructed by the CountVectorizer with max and min document frequency thresholds to balance generalizability and specificity of terms. The output is a probability distribution of topics for each caption, providing high-level thematic insights into the video context.

$$e_{LDA} = p(topic_1|T), p(topic_2|T), \ldots, p(topic_K|T)$$

After obtaining RoBERTa embeddings, keyword phrases, and LDA topic distributions, these features are combined into a consolidated feature vector. This fusion process leverages np.hstack to horizontally stack the RoBERTa embeddings and LDA topic distributions, creating a multi-dimensional vector that serves as the final representation of the text content.

$$e_{combined} = e_{RoBERTa} \oplus e_{LDA}$$

where $\oplus$ denotes concatenation. This combined feature set is saved for each video caption, allowing for efficient retrieval during model inference. The embeddings are stored individually for each video in a directory structure to facilitate organization, while global embeddings are saved as a single array, ensuring that local and batch-based model inferences are possible.

4.3 Multimodal Fusion

Multimodal fusion is the process of integrating information from multiple modalities to enhance understanding and analysis of complex phenomena. This approach capitalizes on the unique strengths of each modality resulting in a more comprehensive representation of the data. There are various strategies for implementing multi-modal fusion, including feature fusion or early fusion , where raw data from different modalities is combined at the input level before processing; and decision fusion or late fusion, which involves processing each modality independently through separate models and then combining their outputs at a later stage.

4.3.1 Concatenation Multimodal Fusion

The concatenation-based fusion method is a straightforward yet powerful approach for combining multi-modal data, particularly useful for integrating visual and textual features in tasks like anomaly detection. This approach captures both visual and textual representations within a single, unified feature vector, retaining the complete information from each modality without applying weighting or gating constraints. In concatenation-based fusion, let x_v represent the visual feature vector extracted from video content, and x_t denote the textual feature vector derived from captions. The concatenated feature vector, h, is formulated as:

$$h = \{x_v \cdot x_t\}$$

This operation stacks the two vectors side-by-side, forming a combined feature vector with a dimension equal to the sum of the individual dimensions of x_v and x_t.

This simple fusion method is both computationally efficient and easy to implement. It allows machine learning models to learn multi-modal representations directly from the fused vector h without applying modality-specific weighting, enabling the model to capture correlations between visual and textual data during training.Concatenation thus provides a balanced and all-encompassing representation that is well-suited for initial or baseline multi-modal fusion stages, as it includes all available information for downstream processing without needing additional parameters or gating mechanisms.

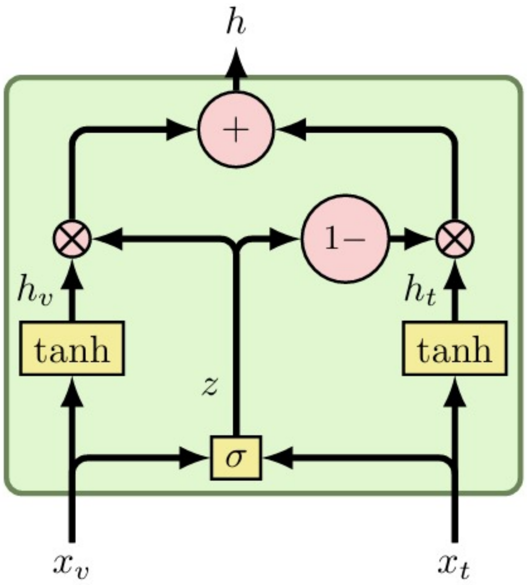

Figure 3. Representation Of GMU

4.3.2 Gated Multimodal Fusion

The Gated Multimodal Unit (GMU) model is an integral component of the anomaly detection architecture, specifically designed to synthesize intermediate representations by integrating visual features from videos and textual features from captions.

The GMU's design draws inspiration from recurrent neural networks (RNNs), particularly Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, both of which utilize gating mechanisms to facilitate effective information flow. At its core, the GMU employs multiplicative gates that dynamically evaluate the importance of various input features. This gating mechanism allows the unit to prioritize data aspects that are more likely to enhance the accuracy of output generation.

In the GMU framework, each modality is represented by a feature vector x_i. These vectors are processed by neurons using a 'tanh' activation function, which helps derive internal representations specific to each modality. For every input modality, a corresponding gate neuron assesses how much influence that modality should have on the overall output. When new data is introduced into the network, the gate neuron evaluates the feature vectors from all modalities and determines whether to incorporate the current modality's contribution into the internal representation of the input sample.

Figure 3 gives the illustration of GMU. The equations governing the GMU are as follows.

$$(i) \quad h_v = \tanh(W_v \cdot x_v)$$

$$(ii) \quad h_t = \tanh(W_t \cdot x_t)$$

$$(iii) \quad z = \sigma(W_z \cdot [x_v, x_t])$$

$$(iv) \quad h = z \cdot h_v + (1 - z) \cdot h_t$$

$$(v) \quad \Theta = \{W_v, W_t, W_z\}$$

$$(vi) \quad \tanh(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}}$$

Here,

x_v represents visual features obtained from the videos ; x_t represents the textual features obtained from the captions. Θ represents the parameters to be learned and $[\cdot, \cdot]$ indicates the concatenation operation.

This approach enables the GMU to generate a rich multimodal representation without requiring manual adjustments or tuning. Instead, it learns directly from training data, effectively capturing complex relationships between visual and textual inputs. By leveraging this architecture, the GMU can adaptively learn and optimize its performance based on the data it processes, enhancing its capability in anomaly detection tasks.

4.4 Anomaly Detection

Following the multi-modal fusion process, the fused representation, which encapsulates the relevant features from both modalities, is forwarded to the classification network. This network is structured as a series of linear layers, each designed to reduce dimensionality while applying ReLU activation functions. These layers work collaboratively to model complex patterns that distinguish between anomalous and normal events.

Incorporated between these layers, dropout is employed to randomly disable a portion of neurons during training. This technique prevents the model from becoming overly reliant on specific neurons and enhances its ability to generalize to new data. By introducing this regularization method, the model can better adapt to variations in input, reducing the risk of overfitting.

The final layer of the classification head employs a sigmoid activation function, which outputs a probability for each class. This output effectively represents the model's confidence level regarding each potential anomaly. The resulting probabilities enable the system to determine whether an event belongs to one of the anomalous categories or falls within the normal scenario. A higher probability assigned to any of the anomalous classes indicates the presence of an anomaly, facilitating differentiation from routine behaviors.

The model is trained using a cross-entropy loss function, which helps optimize classification by minimizing the difference between predicted and actual class probabilities. This loss function further refines the model's capability to distinguish between anomaly classes, enhancing its accuracy and robustness in anomaly detection.

Once the model generates probabilities for each class, the class with the highest probability is selected as the final prediction. If the model predicts any class other than ”Normal” it is flagged as an anomaly. The specific class identified provides valuable insights into the nature of the detected anomaly, aiding in further analysis and interpretation.

5. EXPERIMENTS

During the experimentation phase, various configurations of dropout rates, epoch counts and learning rates were explored to determine the optimal parameters for model performance. Higher dropout rates, such as 0.5 and 0.6, were initially tested to enhance regularization and mitigate overfitting. However, these values resulted in overall accuracy around 60%, which was suboptimal. A dropout rate of 0.3 ultimately provided a balanced level of regularization, yielding improved stability and higher accuracy and was therefore selected for the final model.

Similarly, experimentation with reducing the number of epochs to 10 showed that the model's validation accuracy was limited to approximately 55%, indicating insufficient generalization. An epoch count of 20 was found to be more effective, as accuracy stabilized around this value, providing an optimal trade-off between underfitting and overfitting.

Learning rate tuning was also critical to ensure effective convergence. A learning rate of 5e-5 yielded a low accuracy of approximately 44%, suggesting that it was too slow for effective model training. In contrast, a learning rate of 1e-3 was identified as the most effective, allowing for a faster and more stable convergence while achieving the highest accuracy among tested values.

To further ensure model robustness, early stopping was applied during trials with alternative loss functions, such as Focal Loss and Binary Cross-Entropy, to halt training if validation accuracy began to dip. This measure helped prevent overfitting and conserve computational resources. After evaluating various options, Cross-Entropy Loss demonstrated the best performance and was retained in the final configuration due to its superior accuracy.

Along with concatenation and gated fusion, we experimented with another multimodal fusion technique called Compact Bilinear *Pooling*[25] wherein the model architecture integrates multiple components designed to capture interactions between video and text modalities effectively. First, it employs count sketch-based projections to reduce feature dimensionality and combines input features in the Fourier domain, enabling high-order interactions between video and text features.

Given input video feature $\mathbf{v} \in \mathbb{R}^{T_v \times d_v}$ and text feature $\mathbf{t} \in \mathbb{R}^{T_t \times d_t}$,

$$\mathbf{v}_{\text{proj}} = \mathbf{v}\mathbf{W}_v + \mathbf{b}_v, \quad \mathbf{t}_{\text{proj}} = \mathbf{t}\mathbf{W}_t + \mathbf{b}_t$$

where W is the projection matrix, b is the bias vector and d is the common fusion dimension. For projected features v_{proj} and t_{proj},

$$\mathbf{v}' = \text{CountSketch}(\mathbf{v}_{\text{proj}}, h_v, s_v), \mathbf{t}' = \text{CountSketch}(\mathbf{t}_{\text{proj}}, h_t, s_t)$$

where h stands for hash functions for index mapping and s stands for sign functions for feature transformation. Fourier and inverse Fourier transforms are then applied to obtain the fused feature as below,

$$\mathbf{z}_{\text{fft}} = \mathcal{F}(\mathbf{v}') \odot \mathcal{F}(\mathbf{t}')$$
$$\mathbf{z}_{\text{fused}} = \mathcal{F}^{-1}(\mathbf{z}_{\text{fft}})$$

The fused representation is then enriched by multi-head attention, which, akin to the attention mechanism in transformer models, allows the model to focus on various parts of the input sequence, while positional encoding embeds positional information to help retain the order of elements in the sequence.

This combined feature set is processed through a transformer encoder layer, which applies multi-head self-attention and a feedforward network with residual connections and layer normalization to enhance the representation.

$$\mathbf{z}_{\text{final}} = \mathbf{z}_{\text{enc}}^{(L)}\mathbf{W}_{\text{final}} + \mathbf{b}_{\text{final}}$$

where L represents the number of encoder layers.

The core of the architecture, the multimodal fusion model, projects video and text features into a common space, applies compact bilinear pooling, adds positional encoding, and passes them through a transformer encoder, followed by a linear projection.

6. RESULTS

To assess the effectiveness of various fusion strategies in multimodal anomaly detection, three primary methods—Concatenation Fusion, Compact Bilinear Pooling (CBP), and Gated Fusion—were evaluated on metrics including accuracy, precision, recall and F1-score which have been summarized in Table 1. The performance of each fusion strategy was analyzed based on the fusion order [video+text] and [text+video] to capture any directional dependencies in combining modalities.

Using the Concatenation Fusion method, the model achieved the highest performance, with an accuracy of 85.19% for [video+text], outperforming the reverse order, [text+video], which achieved 55.56% accuracy. Corresponding metrics further support this trend, with precision, recall, and F1-score of 0.85 across these measures for [video+text], compared to 0.55 for [text+video]. This substantial gap indicates that the [video+text] combination effectively leverages video data as a primary source, which adds contextual depth when textual descriptions are appended.

Gated Fusion showed moderate performance, with an accuracy of 77.78% for [video+text] and 70.37% for [text+video]. Precision, recall, and F1-scores for these configurations were 0.77, 0.67, and 0.71 for [video+text] and 0.73, 0.70, and 0.71 for [text+video], respectively. This method offered an improved understanding of cross-modal interactions compared to CBP but did not reach the level of effectiveness observed with concatenation. The performance discrepancy between the [video+text] and [text+video] orders suggests a mild directional dependency in feature integration, although less pronounced than in Concatenation Fusion.

Compact Bilinear Pooling (CBP), which theoretically enhances representation by preserving higher-order interactions between visual and textual features, yielded the lowest results, with an accuracy of 34%. Precision, recall, and F1-scores for CBP were 0.34, 0.41, and 0.37, respectively, indicating an overall weaker performance in combining video and text modalities effectively. The relatively low effectiveness of CBP suggests that, in this context, higher-order feature interactions may introduce noise rather than capturing meaningful cross-modal relationships.

Method	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
Compact Bilinear Pooling	34.00	0.34	0.41	0.37
Gated Fusion (Video + Text)	77.78	0.77	0.67	0.71
Gated Fusion (Text + Video)	70.37	0.73	0.70	0.71
Concatenation (Video + Text)	85.19	0.85	0.85	0.85
Concatenation (Text + Video)	55.56	0.55	0.55	0.55

Table 1. Performance Metrics for Different Methods

7. CONCLUSION

Our research demonstrates the effectiveness of a multimodal anomaly detection system that integrates visual and textual data to classify anomalies in video content accurately. Through the combined use of the TwelveLabs SDK for caption generation, TimeSformer for video feature extraction and RoBERTa, KeyBERT and LDA for text analysis, the model was able to capture complex interactions within videos, leading to robust feature representations. The choice of multimodal fusion strategy played a crucial role in performance, with simple concatenation outperforming other complex methods like gated fusion and compact bilinear pooling.

We evaluated three fusion strategies—Concatenation, Gated Fusion and Compact Bilinear Pooling (CBP)—to integrate video and text data for multimodal anomaly detection. The findings demonstrated that Concatenation Fusion achieved the highest accuracy and performance across all metrics, particularly when video data was prioritized as the primary modality with textual data appended. Gated Fusion offered moderate performance and displayed greater flexibility across fusion orders, providing an adaptable alternative when moderate performance is sufficient. In contrast, CBP struggled to deliver effective results, suggesting that higher-order feature interactions may not be as advantageous for this specific task. The results indicate that concatenation-based fusion approaches are highly effective in this scenario of video-based anomaly detection tasks, emphasizing the importance of modality prioritization in fusion design.

The system achieved high precision and recall for detecting 'Violence' and 'Vandalism' anomalies, while 'Abandoned Object' cases posed a greater challenge due to the visual similarity between certain static anomalies and regular scenes. The model's high accuracy for 'Normal' classifications indicates its robustness against false positives, which is a critical requirement for deployment in real-world surveillance or security systems. The system's performance across diverse anomaly types demonstrates the potential of leveraging multimodal fusion to capture richer contextual cues and improve anomaly classification accuracy in video analysis applications.

Building on these insights, future work will explore several avenues to further enhance multimodal anomaly detection. First, fine-tuning the sequence and weight distribution in Concatenation Fusion could further improve performance, potentially optimizing the impact of secondary modalities in different scenarios. Additionally, dynamic fusion methods that adapt modality prioritization based on the type or complexity of the scene could be developed to improve adaptability in real-world applications.

Addressing the lower performance on 'Abandoned Object' detection is another area for improvement. Advanced object-tracking techniques or the integration of scene-aware models, which focus on spatial consistency and context within scenes, could yield more reliable results for static anomalies. Another promising direction includes dynamic thresholding mechanisms that adjust detection sensitivity based on environmental conditions, improving model adaptability in varied scenarios, such as night-time or inclement weather. Lastly, a more extensive dataset with greater variety in anomaly types and environmental contexts could further refine model performance and generalizability. This continued exploration of multimodal techniques and adaptive methods will enhance the framework's real-world applicability across surveillance and security.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Jarvis Labs for generously providing GPU resources, which were critical in meeting the computational requirements of this work. We are grateful to PES University for offering the necessary resources, facilities, and opportunities that greatly contributed to the successful execution of this research. Finally, we are deeply thankful to Dr. Ramamoorthy Srinath for his invaluable guidance, support, and encouragement throughout the course of this project.

REFERENCES

[1] Mehran, R., Oyama, A., and Shah, M., ”Abnormal crowd behavior detection using social force model,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 935-942 (2009)

[2] Sultani, W., Chen, C., and Shah, M., ”Real-world anomaly detection in surveillance videos,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 6479-6488 (2018).

[3] Wu, T., Suen, C. Y., and Kruse, D., ”Multiple abnormal event detection in videos using deep learning,” *Pattern Recognit.* 93, 53-68 (2019).

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., ”BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics*, 4171-4186 (2019).

[5] Brown, T., Mann, B., Ryder, N., et al., ”Language models are few-shot learners,” *Adv. Neural Inf. Process. Syst.* 33, 1877-1901 (2020).

[6] Xu, L., Liu, C., Feng, Z., and Yu, M., ”Enhancing video surveillance with natural language processing: Integrating LLMs for semantic anomaly detection,” *IEEE Trans. Multimedia* 24(5), 1273-1282 (2022).

[7] Snoek, C. G., van de Sande, K. E., de Rooij, O., and Worring, M., ”The challenge problem for automated detection of anomalous events in video,” *Proc. ACM Int. Conf. Multimedia Retrieval*, 1-7 (2020).

[8] Chen, M., Fang, W., Tang, Y. Y., and Zhang, J., ”Multi-modal anomaly detection in videos via attention mechanism,” *IEEE Trans. Image Process.* 29(2), 5573-5582 (2020).

[9] Vaswani, A., Shazeer, N., Parmar, N., et al., ”Attention is all you need,” *Adv. Neural Inf. Process. Syst.* 31, 5998-6008 (2017).

[10] Wu, P., Cheng, F., and Liu, Z., ”Smart surveillance systems in public spaces: A case study on real-time anomaly detection,” *IEEE Access* 8, 129841-129852 (2020).

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., ”An image is worth 16x16 words: Transformers for image recognition at scale,” *Proc. Int. Conf. Learn. Represent.* (2021).

[12] Bertasius, G., Wang, H., and Torresani, L., ”Is space-time attention all you need for video understanding?” *Proc. Int. Conf. Mach. Learn.* (2021).

[13] Lu, H., Jian, H., Poppe, R., and Salah, A. A., ”Enhancing video transformers for action understanding with VLM-aided training,” *arXiv preprint* arXiv:2403.16128 (2024).

[14] Lv, H., and Sun, Q., ”Video anomaly detection and explanation via large language models,” *arXiv preprint* arXiv:2401.05702 (2024).

[15] Singh, S., Dewangan, S., Krishna, G. S., Tyagi, V., Reddy, S., and Medi, P. R., ”Video vision transformers for violence detection,” *arXiv preprint* arXiv:2209.03561v2 (2022).

[16] Zhao, Y., Misra, I., Krähenbühl, P., and Girdhar, R., ”Learning video representations from large language models,” *arXiv preprint* arXiv:2212.04501 (2022).

[17] Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., and Barnard, K., ”Attentional feature fusion,” *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 3559-3568 (2021).

[18] Baltrušaitis, T., Ahuja, C., and Morency, L.-P., ”Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.* 41(2), 423-443 (2019).

[19] Arevalo, J., Solorio, T., Montes-y-Gómez, M., and González, F. A., ”Gated multimodal units for information fusion,” *arXiv preprint* arXiv:1702.01992 (2017).

[20] Sultani, W., Chen, C., and Shah, M., ”Real-world anomaly detection in surveillance videos,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4085-4093 (2018).

[21] Aremu, T., Zhiyuan, L., Alameeri, R., Khan, M., and El Saddik, A., ”SSIVD-Net: A novel salient super image classification & detection technique for weaponized violence,” *IEEE Trans. Inf. Forensics Secur.* (2022).

[22] Soliman, M., Kamal, M., Nashed, M., Mostafa, Y., Chawky, B., and Khattab, D., ”Violence recognition from videos using deep learning techniques,” *Proc. Int. Conf. Intell. Comput. Inf. Syst.*, 79-84 (2019).

[23] Lin, K., Chen, S. C., Chen, C. S., Lin, D. T., and Hung, Y. P., ”Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance,” *IEEE Trans. Inf. Forensics Secur.* 10(6), 1359-1370 (2015).

[24] Blei, D. M., Ng, A. Y., and Jordan, M. I., ”Latent Dirichlet Allocation,” *J. Mach. Learn. Res.* 3, 993–1022 (2003).

[25] Gao, Y., Beijbom, O., Zhang, N., and Darrell, T., ”Compact bilinear pooling,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 317-326 (2016).

AUTHORS' BACKGROUND

Your Name	Title	Research Field	Personal Website
Nidhi P Gururaj	Undergraduate Student	Deep Learning & Computer Vision	Nidhi P Gururaj
Nikita Suresh	Undergraduate Student	Deep Learning & Computer Vision	Nikita Suresh
Ria R Kulkarni	Undergraduate Student	Deep Learning & Computer Vision	Ria R Kulkarni
Sai Amara Prasad	Undergraduate Student	Deep Learning & Computer Vision	Sai Amara Prasad