Exploring Multimodal Data for Crime Recognition

Aditya Retissin Poozhiyil
Faculty of EEMCS
University of Twente
Enschede, The Netherlands

Dr. Estefanía Talavera
Faculty of EEMCS
University of Twente
Groningen/Enschede, The Netherlands

***Abstract*—With the generation of diverse forms of data being produced exponentially by various forms of devices, researchers have explored exploiting the inherent characteristics of the modality to recognize human actions. The most widely generated data, RGB(color) primarily contributes to the spatial information but lacks the temporal attributes. Conversely, the skeleton modality emphasizes the temporal aspect of the human joints but lacks spatial features. Both these modalities present features that can mutually complement each other. In the context of crime recognition, earlier research focused on capturing and learning temporal patterns by exploring different forms of Transformer architectures with skeleton trajectories. This study extends the work by investigating the fusion of visual context(RGB) with the skeleton to leverage spatial and temporal dynamics of both modalities. The dataset used for this study is the HR-Crime, containing 13 human-related crime categories captured through surveillance cameras. Our experiments show the fusion of both modalities shows improvement compared to the baseline. In addition, we discuss the limitations of our approach and possible ways to tackle them.**

I. Introduction

Closed-circuit television (CCTV) cameras have been widely employed in both public and private settings for surveillance purposes aimed at preventing and resolving criminal activities[1]. However, the manual monitoring of prolonged hours of CCTV footage is a time-consuming and laborious task, thereby reducing the efficiency of CCTV systems. The recognition of criminal activity in CCTV cameras entails the automatic detection and classification of acts such as theft, vandalism, assault, and burglary, among others. The use of such systems allows the authorities to take preparedness measures in advance, enabling them to respond more effectively to various scenarios and potential emergencies.

In recent years, the progress made in novel hardware technologies, capable of capturing and generating diverse forms of data, allowed researchers to employ Deep Learning algorithms for the purpose of recognizing human activities. The accessibility of data, generated by wide range of hardwares such as Kinect, LiDAR and RGB sensors has paved the way for extensive research on exploiting the inherent characteristics of the data[2, 3, 4].

Each modality contains unique inherent characteristics that contribute towards Human Activity Recognition(HAR) tasks. For example, RGB modality is a widely employed color model, that plays a significant role for solving computer vision related tasks[5]. Depth data, generated by LiDAR sensors, provides information about the distance of objects from the camera, hence, encompassing the structural information of the scene[6]. Another modality which is extensively used in context of HAR task is skeleton data, which contributes as the temporal component. The data is generated by the utilisation of Kinect sensors or by employing pose estimation algorithms[7, 8] to extract the poses from video.

Although each modality has its merits, they also exhibit certain limitations. For instance, RGB modality may contain noise and is sensitive to light[9]. Conversely, skeleton data lacks spatial information and may fail to capture subtle variations in human movement, such as nuanced shifts in posture and appearance[10]. Hence, making it highly dependent on the device and pose estimation algorithm employed. Consequently, it becomes evident that each modality presents unique aspects that is absent in the other. Therefore, it is a rational to fuse modalities that exhibit mutual complementary, with the aim of leveraging the distinctive characteristics inherent to each modality.

In the field of Natural Language Processing(NLP), there has been many works combining modalities, primarily text and image, that has proven to be effective, achieving state of the art results in image captioning[11], text to image generation[12], indicating, information when aligned properly, is able to learn relationships between modalities. In context of HAR, it has been observed that employing a single modality may not always yield optimal results due to inherent limitations such as noise, occlusion, and variability in sensor placement[13]. Therefore, researchers have explored the possibility of effectively utilizing multiple modalities, to capture complementary information and improve the performance of HAR systems.

To this end, various fusion techniques have been explored attempting to fuse different modalities[14, 15, 16, 17, 18, 19]. These techniques encompass early, intermediate, and late fusion approaches, where, early fusion refers to combining raw data from different modalities at earlier stages of training, more specifically at the input level, while intermediate fusion involves the fusion of features from both modality at a certain stage of the within the training process. This step allows the model to learn inter-modality relationships. Lastly, late fusion combines the outputs of both modalities at the final stage, thereby enabling the model to only learn relationship between modalities at the decision level.

The primary objective of these fusion methods is to allow the model to exploit the strengths of each modality and overcome their limitations by learning complementary and discriminative representations. By combining the information from multiple modalities, the model can capture a more comprehensive representation of the activity being performed.

In context of crime recognition, earlier works explored different architectures utilising transformers[20]. Whereas, in the study conducted by Joseph[21], which is an extension of the aforementioned work, explore using Tubelet Embedding[22] on skeleton trajectories. However, the presented work solely utilizes skeleton trajectories. This study extends their work by the fusion of the RGB with the skeleton trajectory in order to provide visual context to the skeleton trajectory. Which brings us to our main research question:

1) How does the fusion of visual descriptors affect the classification performance of a skeleton-based model?
    a) How to effectively include the visual information, through full frames or person-centric bounding boxes?

The paper is organized as follows. Section II presents the technical background. In Section III, provides an in-depth exploration is conducted to review the literature on how the modality, specifically, RGB and skeleton are utilised, and the following we review the approaches followed to fuse both the modalities. Section IV explains the approach for the study at hand. In Section V, a detailed description of the experimental setup is presented. This includes pertinent information regarding the dataset used in the study, as well as the training strategies employed to train the model under examination. Section VI is dedicated to the presentation of the experimental results obtained from the study. The outcomes of the conducted experiments are thoroughly elucidated and analyzed in this section. Section VII presents the discussion about the experiments. Finally, SectionVIII concludes the study.

II. Technical background

A. Transformer

A novel work by Vaswani et al., originally introduced transformers to solve wide range of NLP tasks. The model consists of two components. An encoder and decoder shown in Figure 1. The role of the encoder is to extract features, and capture the relationship between words in a sequence. Prior to being fed into the encoder, the input undergoes a transformation into embeddings. Which is essentially a vector representation of the sequence. After which, a positional encoding is added to the input embedding to give a sense of order to the input sequence. The positional encoding contains information of the words in the input sequence, allowing the model to exploit this information to understand the order of the sequence, since the inputs are not processed sequentially. In the encoder and decoder lies module the Multi-Head Self-Attention(MHSA) layer, the core to what drives the transformer architecture[23].

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

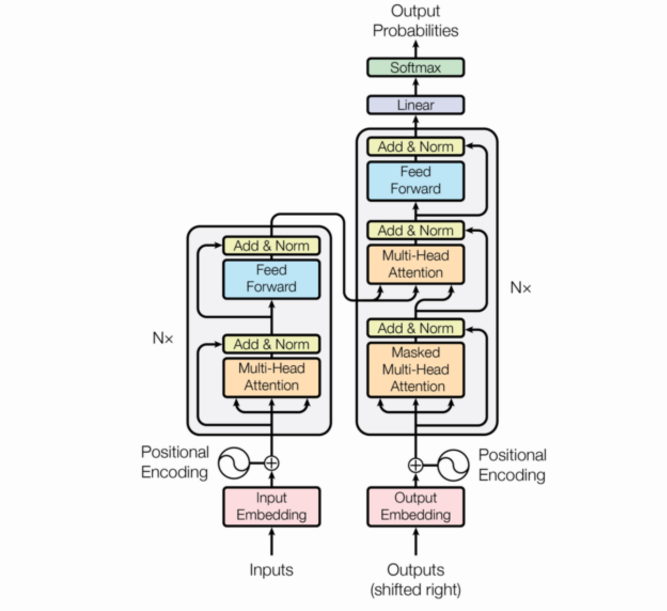


Fig. 1: Transformer architecture[23]

Where each attention head is denoted by Equation 1. Where Q, K, V are the query, key and values respectively, which is obtained by linearly projecting the embedding. As depicted in Equation 1, the calculation involves dot product between the query and the transposed of the key matrix. In this step, every word in the input sequence attends to other words in the sequence to learn their inter-relationships. The term $\sqrt{d_k}$ is used to control the weights to avoid large gradient during back propogation. Subsequently, a softmax operation is applied to the result, which yields the relevancy scores of each word with respect to the other. The resultant factor is multiplied with the associated value vectors to obtain the weights for each word. Similary, in the decoder, contains MHSA block. However the key different lies in the “Masked Multi-Head Attention”. The approach is similar to MHSA, but, here the model tries to decode the next possible input based on the input its already seen. This allows the model to condition its prediction based on the sequence that its already seen and doesn’t allow the model to look ahead.

The self-attention layer allows the model to attend different parts of the input sequence which allows the model in processing and capturing relationships for longer input sequences. This mechanism enables the model to process large amounts of data in parallel, since the tokens are not processed sequentially, as opposed to Long Short Term Memory(LSTM) architecture, where the sequence is processed sequentially[23].

The architecture initiated a paradigm of self-supervised pretraining. Where the model can be pre-trained on large amounts of data by unsupervised learning, allowing it to learn useful representations of text that can be fine-tuned on specific downstream tasks with relatively small amounts of task-specific data. This makes the transformer architecture highly versatile and effective for a wide range of NLP tasks, such as machine translation, language modeling, text classification.

B. Vision Transformer

Motivated by transformers, Dosovitskiy et al., introduced Vision Transformers(ViT). Here, the model takes in images as input, treating each input as words. Initially, the input is first split into patches, following after, the patches are converted to embedding by linearly projecting the values as shown in Appendix A.1. Now, each patches are linearly projected and passed as input to the transformer encoder. In contrast to the transformer model that incorporates both an encoder and decoder block, ViT exclusively employs only encoder block. Finally, a CLS token is prepended at the beginning of the sequence, which is used for classification purposes as in the original transformers. The authors demonstrate that ViT surpasses convolution-based methods in performance.

III. Related Work

This section presents a detailed summary of research conducted on HAR tasks on different modalities. More specifically, RGB and skeleton. We first review the uni-modal approaches and how they are utilised for the task at hand. After which, we explore different fusion approaches of both the modalities.

A. Single Modality

*1) **Color Descriptors(RGB)**:* As stated previously, RGB is the most widely used modality for computer vision tasks. The availability of RGB cameras, sensors promoted the growth of data being generated exponentially on daily basis, allowed researchers to investigate and solve wide range of computer vision related applications. With the help of Convolutional Neural Networks (CNNs), the modality enabled the development of sophisticated models capable of extracting and learning meaningful features from visual data through a series of convolutional blocks, leading to significant advances to solve problems based on image classification, object detection[24, 25], and semantic segmentation[26].

There is currently a significant level of research activity focused on HAR task utilising the foregoing modality. Initial methods involves using hand-crafted feature based approaches[27, 28, 3]. However, these approaches are highly feature dependent and also demands a profound expertise in the domain, thereby posing significant challenges when it comes to deployment. The task has been further investigated by utilising a simple CNN architecture to learn spatio-temporal features from a sequence of frames by first extracting features independently from a sequence of frames, and concurrently stacking them to form a sequence of descriptors, which is then fed to the output layer. Nevertheless, this approach demonstrates relatively lower performance than handcrafted features[29]. Because standard CNN models, such as AlexNet[30], VGG[31], and ResNet[32] learn spatial features from a single frame. Thus, making them ineffective for modelling the temporal information since the previous frames are not taken into consideration until the final layer.

Subsequent approaches aim to leverage temporal information by employing a two-stream CNNs-based architecture, in which one stream is dedicated to extracting spatial features from a sequence of frames, while the other stream focuses on capturing temporal characteristics by providing motion flow images as input. Afterwards, fusion strategies are implemented to fuse the extracted features and obtain the final classification[33].

A notable work by Karpathy et al., further investigated connectivity of CNN in the time domain for video classification, where different fusion methods was investigated. As per the previous work, they adopt a two-stream architecture where the model is trained by feeding low resolution frames by downsampling, namely, the context stream and high resolution frame which is the center cropped portion of the original frame. Through a series of experiments, it was observed that the networks operating on individual frames performed on par with the networks processing the entire spatio-temporal volume of the video[34].

Although the work uptil now is purely based on trying to exploit temporal dependencies through spatial information, due to the aforementioned limitation, it lacks the ability to capture long-term dependencies, which is crucial component when dealing with HAR related tasks[2]. To tackle this, approaches have been made employing different Recurrent Neural Network(RNNs) variants in the pipeline, primarily Long Short Term Memory(LSTM). Vanilla RNNs have been widely used to solve sequential tasks, however, they are prone to vanishing and exploding gradient problems[35]. To tackle this, the former architecture is used.

LSTMS has achieved state of the art performance in sequential tasks, more specifically, NLP tasks. With its ability to capture long-term temporal dependencies, it has been ideal choice to tackle sequential problem. In relation to HAR, LSTMs lacks the ability to learn spatial context in images, since it learns to mainly model temporal information. This issue was addressed by utilising CNN as backbone to extract rich features features[36, 37]. In the study by Donahue et al, [36] demonstrate that the LSTM can effectively capture spatio-temporal patterns in activity recognition tasks by leveraging CNN as the underlying backbone. Further studies have extended this approach by incorporating attention mechanisms to focus on relevant regions[38, 39, 2].

In the more recent work, ViT has been employed in HAR tasks[40, 41, 42, 43] replacing CNN role as the backbone and feeding it to an LSTM as input. By utilising the characteristics of MHSA in ViT, allows for capturing long-range connections and adaptively aggregating spatial information[44]. The preceding work have subsequently been extended to videos[22], where the authors propose Tubelet embedding, which in essence is a 3D convolution that performs a volumetric convolution along the spatio-temporal axis to extract and encode the spatio-temporal tokens as shown in Appendix A.2. Identical to approach in transformers, the output is linearly projected to a dimension D which is then passed as input to the transformer.

*2) **Skeleton**:* Skeleton trajectories contains temporal information about the human body. As discussed in section I skeleton trajectories have the advantage of being invariant to the background and focus more about the character[45]. These characteristics of the modality help focusing on the temporal dynamics of the subject of interest. The data can be obtained by using sensors or pose estimation algorithms[8, 46, 47, 48]. As inputs, they are essentially a pair of coordinates representing the location of the joints in spatial space. The representation of features makes the inputs easily processable and computationally inexpensive[49].

With respect to HAR tasks, studies based on the skeleton modality been focused on trying to capture relationships between joints with respect to the activity[50, 51, 4]. Approaches were followed by grouping skeleton based on semantically similar parts, and having a subnet, more specifically the authors employed BILSTM for each group. The output of each group is later fused with the output of other groups to finally have a combined representation of the skeleton, which will be used for classification[4].

However, since the approaches only the skeleton coordinates as input to the model. Subsequent works utilised Graph Convolutional Networks(GCNs) were used to embed the skeleton points specifically for HAR related tasks[52, 53, 54]. The approach was first introduced by Yan et al., where the authors proposed Spatio Temporal Graph Convolutional Network(ST-GCN). A spatial graph is constructed by utilizing the inherent interconnections of joints within the human body, while incorporating temporal edges that link corresponding joints across consecutive frames[52]. Since the approach involves using a predefined skeleton graph, this restricts its ability to capture important relationships between distant body parts, limiting its effectiveness in recognizing actions that rely on such relationships. To address this issue, Shi et al., proposed a two-stream adaptive graph, where the graphs used in the model are dynamically generated based on the data, thereby mitigating the limitations of the predefined graphs in the original ST-GCN and enhancing the model's ability to capture relevant relationships for action recognition[53].

Subsequent to the release of transformers, numerous studies have been conducted on trying to exploit the characteristics of self-attention mechanism to learn the relationship between joints[55, 56, 57, 58]. In the study by Plizzari et al., they employ a two stream transformer architecture, utilising self-attention mechanisms to capture relationship between the joints individually through time. The two streams are as follows (a) Spatial Self Attention(SSA) (b) Temporal Self Attention(TSA). The former focuses on capturing correlation between each pair of joints independently whereas the latter captures correlation between joints through time[56].

Similarly, the work presented by Wang et al., focused on grouping the joints into a single body part. They use a single transformer encoder block for computing spatial and temporal features of the skeletal joints data. Their proposed method involves computing correlations between joints in one part, across parts in one frame and across frames for same part, using a modified intra-inter part attention mechanism[57]. Similarly, in the work by Boekhoudt, explored different types of transformer architectures, specifically for HAR tasks. One of the proposed architecture involves grouping the body parts and having a transformer block for each body part[20].

Following the aforementioned work, Joseph[21], extended the work of Boekhoudt[20] by incorporating Tubelet embedding to all different architectures presented. The findings of the study indicated that incorporating Tubelet embeddings resulted in comparable performance while reducing the model's complexity compared to the outcomes reported in the previous research.

B. Multimodal

*1) **RGB & Skeleton**:* In the preceding section, the reviewed literature shows the investigation of RGB and skeleton modalities separately. The findings suggest that both modalities possess features that are mutually complementary. Hence, combining both modalities would be a logical approach, as RGB would contribute towards the spatial information and skeleton provides the temporal component. Thereby allowing the model to to exploit their inherent features.

To combine both the modalities, diverse fusion techniques were explored to capture the inter-modal relationships on a feature level and decision level. Generally, approaches consists of two streams architecture, where each stream provides a different modalitiy, and performing a late fusion[14, 15]. The approach involves training two models independently, where one model focuses on learning a unimodal and the outputs of both the modals are concatenated and passed to a fully connected layer as input. However, this approach constraints the model to learn complimentary information between modalities[59] as each model learns features about a unimodal independently.

In the study carried out by Verma et al., they propose a two-stream architecture, where one stream takes in a sequence of RGB frames and the other takes in skeleton sequence. In the RGB stream they convert the sequence of frames into a single image, namely, Motion History Image(MHI) and Motion Energy Image(MEI). MHI, fundamentally represents the motion of the video by increasing the pixels values which have a higher change throughout the video. On the other hand, MEI encodes the whole movement into a single binary image. In the skeleton stream, they condense the sequence of skeletal data into a singular composite image[15].

Subsequent approaches investigated different fusion strategies prior to decision level. At this stage of fusion, information between the modalities are attended on a feature level, allowing the model to learn cross correlation between both the modalities. Works have explored adopting attention mechanisms to learn the information and importance of joints by providing spatial-context[16, 17, 18, 19]. A notable work by Liu et al. [51] proposed a Global Context-Aware Attention LSTM(GCA-LSTM). The method assists the network to focus on the joints in each frame to finally generate an attention representation for the sequence. They utilise a cross attention block to attend features of both the skeleton and the frame. In the study by Zhu et al. combines early and late fusion. Additionally, they utilize a self-attention module to assist the network in directing its attention towards specific body parts.

The studies of HAR tasks have been extended to understanding the activity based on the actions within their surrounding environment[60, 61]. In the study by Faure et al. [61] proposed a "Holistic Interaction Transformer", which composes of RGB and pose subnetwork. The aim of this network is to learn the persons interaction with their surroundings by focusing on key entities that drive most of the actions. The output of both of these network is fused together by an Attention Fusion Module and then to further model how actions evolve in time by looking future and past frames using a temporal interaction unit. Similarly, Li et al. [60] proposes fusion of skeleton and RGB at a feature level where they aim to guide attention on the objects caused by the action using a guided block

Inspired by Slowfast networks[62], researches have explored different frame rates, making a slow and fast pathway using transformer[63, 64]. By having the slow pathway, the model is able to learn the spatial information. Whereas the fast pathway aids in capturing the temporal information[62]. Similarly Jing et al. [63] proposes a dual stream architecture, where the input of the first stream consists of two sub-streams providing the slow and fast image, and the latter provides sequence of skeleton trajectories. Following, they utilise three different fusion strategies(early, halfway and late fusion) to combine the skeleton trajectories, where the outputs are finally concatenated. As per the previous work, Shi et al. [64] utilizes a two-stream transformer named RGBSformer, that takes in heat maps of high frame rate skeleton poses followed with low frame rate RGB images as input.

The reviewed literature shows how different models learn inter-modal relationships using different fusion approaches. Specifically, transformers when used with skeleton trajectories input have demonstrated remarkable proficiency in capturing temporal dependencies and modeling intricate relationships of the temporal dynamics of the skeletal structure. Leveraging the self-attention mechanism characteristics, this approach facilitates the acquisition of meaningful representations from joint sequences. For ViT, is able to extract capture spatial relationships among these patches. Hence by leveraging the features of ViT, we can fuse them with output of the skeleton transformer to enhance the overall representation of the data.

In the work presented by Joseph[21] and Boukedeth's[20], their research specifically explored skeleton modality with transformers on the HR-Crime dataset. Currently, there is no existing work that explores the fusion of both modalities in the context of crime recognition. Building upon their work, our paper extends this investigation by fusing RGB modality to provide visual context to the skeleton trajectories.

.

IV. Methodology

In the subsequent sections, we will discuss about the input representations that will be used throughout our experiments. Followed by the choice and architecture of the model. Lastly, we discuss the fusion methodology that has been chosen for the study at hand.

A. Input Representations

For this study, our attention will be directed towards a pair of distinct input representations. More specifically, our research will focus on the fusion of RGB frames alongside with skeletal structure. For the visual context, we utilize only a single frame which will be the middle frame of the corresponding sequence of skeleton trajectory(please refer Appendix B.1 for a visual illustration).

1) ***Skeleton Representation:*** Skeleton data is usually presented as a series of 2D or 3D coordinates that track keypoints of the human body over time. By extracting the poses of the subject over a sequence of frames, we obtain a sequence of trajectory coordinates $X_{skel} = \{(x_i, y_i) \mid i \in \mathbb{R}^{T \times J}\}$, where x_i and y_i represent the skeleton points of the joints in spatial space. The variable T denotes the segment length, and J represents the total points. It is also to note that, depending on the pose extraction algorithm employed, the number of joints can vary.

2) ***Visual Representation:*** In order to fuse the visual context with the skeleton trajectory, our study entails two different kinds of feature representation, namely, the full frame and bounding-box level representation.

Full Frame. This input representation utilises the entire frame as context. By using the entire frame, we provide a more global representation of the scene. Additionally, we utilise only the middle frame to represent a sequence of trajectories as shown in Appendix B.1.

Bounding-Box Level. For the second feature representation, we expand the coordinates of the skeleton to obtain a Region of Interest(ROI) points. These points are subsequently used to crop a more finely localized visual cue. We scale the coordinates to provide a slightly broader context as opposed to completely isolating the background and focusing on the subject in the video.

To calculate the ROI from the obtained skeleton points, we first determine the bounding box of the skeleton by calculating the width and height of the skeleton coordinates. Given the sequence of trajectory coordinates X_{skel}, the width and height of the skeleton trajectory is calculated as specified in Equation 2,

$$\begin{aligned} w &= max(x_i) - min(x_i), \\ h &= max(y_i) - min(y_i). \end{aligned} \tag{2}$$

Where, w and h is obtained by subtracting the maximum and minimum of the x and y coordinates respectively. Subsequently, we introduce a scaling factor α and multiply it with the terms w and h to obtain the scaled terms E_{width} and E_{height} respectively as specified in Equation 3. Where E_{width} and E_{height} is the expanded width and height of the ROI,

$$\begin{aligned} E_{width} &= \alpha * w, \\ E_{height} &= \alpha * h. \end{aligned} \tag{3}$$

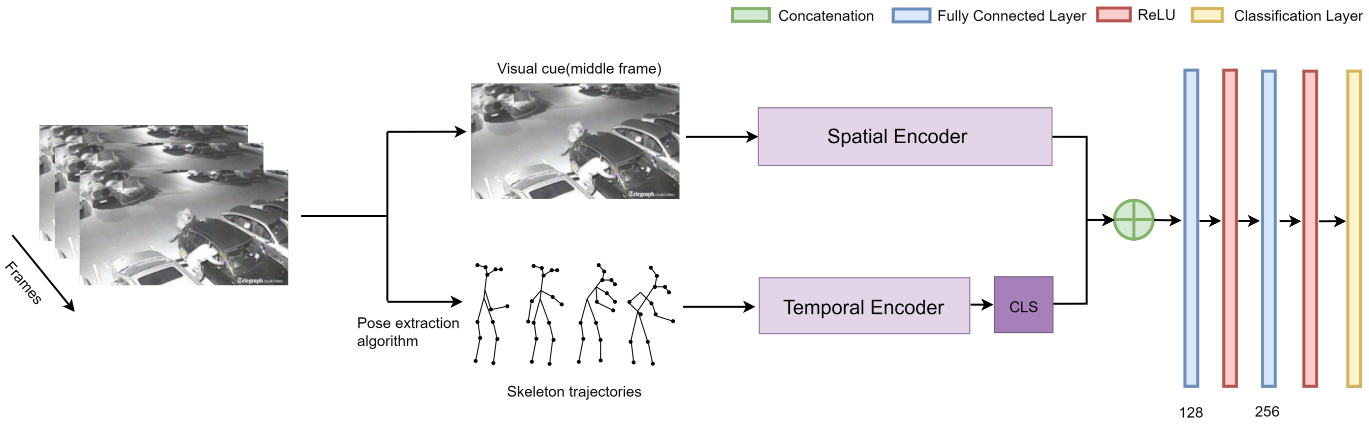


Fig. 2: The Figure illustrates the overall architectural design, **Visual Temporal Transformer(VT-Tran)**. Given a sequence of video frames, AlphaPose[7] is used to estimate the pose representation. Consequently, the middle frame is chosen to represent a sequence. Both the visual cue and skeleton trajectories are then simultaneously passed to the spatial and temporal encoders. Where the outputs of the spatial encoders and the CLS token of the temporal encoder is fused through concatenation. The CLS block in the temporal encoder denotes the *class* token. Subsequently, the features undergo a series of fully connected layers to learn features from both modalities.

We then calculate the expanded region of interest coordinates by finding the minimum and maximum limit with respect to E_{width} and E_{height} to restrict the coordinates from going above or below the original image height or width as shown in Equation 4,

$$\begin{aligned} E_{xmin} &= max(0, min(x_i) - E_{width}), \\ E_{ymin} &= max(0, min(y_i) - E_{height}), \\ E_{xmax} &= min(max(x_i) + E_{width}, 320), \\ E_{ymax} &= min(max(y_i) + E_{height}, 240). \end{aligned} \tag{4}$$

In Equation 4, E_{xmin}, E_{xmax}, represents the lowest and the highest extents of the expanded coordinates along x-axis. Similarly, E_{ymin}, E_{ymax} denotes minimum and maximum of the expanded coordinates along y-axis. The calculation of E_{xmin} and E_{ymin} ensures the values are non-negative, and the terms E_{xmax}, E_{ymax} ensures that values stay within the region of the image. Using these points we crop the ROI to obtain the final context.

Additionally, we resize the image with a width and height of 224 ensuring consistent image size. Furthermore, this step is done to align the input dimensions with the spatial encoder which will be discussed further in Section IV-C. We investigated different values for α and it was observed that having a scaling factor $\alpha = 0.5$ gives satisfactory localized visual context as shown in Appendix B.2.

B. Temporal Encoder

For the temporal stream, we will be using the Temporal Transformer presented by Boekhoudt, with the aim of capturing mainly the temporal patterns[20]. Given the sequence of trajectories, $S \in \mathbb{R}^{T \times J}$, the sequence is linearly projected to a higher dimension D by using a fully connected layer to obtain $S_{proj} \in \mathbb{R}^{T \times D}$. Additionally, we utilise a *class* token, where $X_{class} \in \mathbb{R}^{1 \times D}$ which is prepended along the temporal axis of the sequence, to obtain the dimension $S \in \mathbb{R}^{(T+1) \times D}$(Similar to BERT). The inclusion of *class* token summarizes the sequence by capturing the global temporal information of all the trajectory in a sequence by utilising the characteristics of self-attention mechanism. Furthermore, positional embedding is added to the embedding of the sequence $S_{pos} \in \mathbb{R}^{(T+1) \times D}$ to learn the positional information of the sequence. Where $(T+1)$ accounts for the *class* token(please refer to Appendix C.1 for visual illustration). The final input representation is as follows:

$$O = [x_{class}; s^1 S_{proj}; s^2 S_{proj}; ...; s^t S_{proj}] + S_{pos} \tag{5}$$

The input O is subsequently fed into transformer encoder layers to learn relationships between trajectories. Concurrently, the *class* token is obtained, which will be the final representation of the temporal encoder which will be further fused with output of the spatial encoder which will be discussed in Section IV-D.

C. Spatial Encoder

For the spatial stream we will adopt ViT as backbone pre-trained on ImageNet[65]. The image $x \in \mathbb{R}^{H \times W \times C}$ is first decomposed into non-overlapping patches of the shape $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$. Where (H, W) represents the height and width, and C represents the number of channels in the image. After the following, the patches are linearly projected to a higher dimension $x_p \in \mathbb{R}^{N \times D}$. In order utilise the pre-trained weights, gradient computations are disabled across the entire network. Concurrently, the classification head is removed to attain image descriptors where $x_{desc} \in \mathbb{R}^{N \times 768}$. Simultaneously, the classification head is replaced by adding two sets of fully connected layers, where the second layer is

T Class	8		14		20		24		32	
	R	D	R	D	R	D	R	D	R	D
Abuse	0.052	58.4	0.052	57.6	0.052	56.8	0.052	93.0	0.048	40.3
Arrest	0.046	52.1	0.046	50.9	0.046	49.7	0.045	48.9	0.059	49.1
Arson	0.029	32.8	0.029	32.1	0.029	31.4	0.029	31.0	0.011	9.6
Assault	0.062	70.0	0.062	68.7	0.062	67.4	0.062	66.5	0.055	12.6
Burglary	0.074	83.7	0.075	82.3	0.075	80.9	0.075	80.0	0.043	35.4
Explosion	0.029	32.5	0.029	31.9	0.029	31.4	0.029	31.0	0.015	12.6
Fighting	0.067	75.0	0.066	73.3	0.066	71.5	0.066	70.4	0.140	116.3
Road Accidents	0.064	72.1	0.064	70.8	0.064	69.5	0.064	68.7	0.028	23.1
Robbery	0.087	98.3	0.087	96.3	0.087	94.3	0.087	93.0	0.133	109.8
Shooting	0.074	83.6	0.074	82.1	0.074	80.6	0.074	79.6	0.018	15.3
Shoplifting	0.287	321.5	0.289	317.6	0.291	313.7	0.292	311.1	0.372	307.4
Stealing	0.071	78.6	0.069	76.4	0.068	74.3	0.068	72.95	0.059	48.7
Vandalism	0.052	58.6	0.052	57.1	0.051	55.6	0.051	54.64	0.013	11
Total	1	1117.1	1	1097.1	1	1077.1	1	1100.7	1	791.2

TABLE I: Number of trajectories after dividing fixed segment lengths (**T**) on the test set. Where **D** depicts the total number of trajectories(expressed in thousands) and **R** denotes the probability of random guess for each class.

the spatial encoder output as shown in Figure 11. Making these two layers the only trainable layers in the spatial encoder.

It is also important to note that the spatial encoder output does not correspond to the classification scores, but a layer to align the dimensions with the temporal stream which will be fused in the later stage. From here on, we will use ViT and spatial encoder interchangeably.

D. Fusion Layer

To combine both the modalities, we train both the unimodals independently as illustrated in Figure 2. Before fusion, the output of the temporal and spatial encoder are aligned with a dimension D to ensure same dimensionality. Note the alignment is done at the initial stages, where we ensure both the encoders outputs the same dimension D. Subsequently, the resulting output of both the encoders are concatenated along the feature axis where $F_{late} \in \mathbb{R}^{N\times(D*2)}$ to obtain the combined input representation $x_i = [x_i^{skel}; x_i^{rgb}]$. Where x^{skel} represents the output of the temporal encoder and x^{rgb} signifies the spatial encoder output and **;** denotes concatenation. The combined representation then undergoes a series of fully connected layers, facilitating the learning of features from both modalities. Finally, we employ a log-softmax function to convert the scores for each class on a logarithmic scale, yielding the final prediction scores.

V. Experimental setup

A. Dataset

The dataset we employ for our experiments is the Human Related Crime Recognition(HR-Crime) dataset[66], which is a subset of UCF-Crime[67] dataset, consists of 950 anomaly videos. The dataset is categorized into 13 classes, namely, Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shooting, Shoplifting, Stealing and Vandalism(please refer Appendix D.2 for class visualization). The dataset contains skeleton trajectory which was obtained by first detecting the human body proposals using YOLOv3-spp and then Alphapose[7] estimate the pose which is tracked using PoseFlow[68]. In our experiments, we sample the data using the Stratified split in the proportion of 80:20 ratio. This step is taken to ensure a balanced proportion of each class since we are dealing with class imbalance problems. A comprehensive overview of the distribution of videos & trajectories is present in Appendix D.1.

Baseline Approach In the results presented by [20], the train/test/val was split based on trajectory level, essentially having subjects in the same video appear across all the splits. This approach, however, exhibits a limitation when fusing visual cues with trajectory. Given that multiple subjects within a scene share identical visual cues, it can result in data leakage, leading the model to overfit the entire dataset. Hence, to align the baseline with our experiments, we reproduce the results by splitting the data on video-level, after which, we divide the trajectories into fixed segment trajectory. The split ensures the absence of data leakage within any of the splits. Table II presents the results of T-Tran[20] before and after the new split.

Model	Accuracy(w)	F1 score	Segment Length
T-Tran-V1-6	0.476 +/- 0.009	0.611 +/- 0.004	24
T-Tran-V1-6*	0.121 +/- 0.030	0.205 +/- 0.030	24

TABLE II: Reported baseline accuracy before and after the split where * denotes the the new split

B. Implementation Details

Input preprocessing. The trajectories are divided into fixed segments T, ensuring same segment length for each trajectory. Table I presents the distribution after dividing fixed segments along with the corresponding random guess values(calculated by $\frac{c}{n}$, where c denotes the total number of samples per class, and n denotes the total number of samples in the test set) for each class on the test set. Whereas, for the spatial encoder, the dimensions of the image are resized to 224.

Temporal Encoder. We fix the same hyper parameters presented by Boekhoudt[20], having $N_{depth} = 4$ and $N_{heads} = 8$. The weights of the network are uniformly initialized from the range of -0.1 to 0.1.

Spatial Encoder. As for the spatial stream, the RGB frames are divided into non overlapping patches with $N_{patches} = 16$. Subsequently, the input is normalized according to the mean and standard deviation of the ImageNet dataset. Table III presents the mean and standard deviation of the entire ImageNet dataset over the 3 channels of the images.

Channel	Mean	Standard Deviation
1	0.485	0.229
2	0.456	0.224
3	0.406	0.225

TABLE III: Mean and standard deviation of ImageNet dataset.

Training. To train the model, we fix a learning rate of 0.0001, and a batch size of 500. Adam optimizer was used throughout our experiments in order to update weights. Training strategies such as early stopping was implemented to avoid over-fitting. To ensure robustness in the performance, the video-level data was split over 5 K-folds, followed by dividing the segment into fixed length.

C. Evaluation Metrics

To evaluate the performance of the model, we validate using balanced accuracy and F_1 score as the primary metric as shown in Equation 7 and 6 respectively.

$$Accuracy = \frac{sensitivity + specifity}{2}, \tag{6}$$

$$F_1 = 2 \times \frac{precision * recall}{precision + recall}. \tag{7}$$

Where $precision$ defines as the ratio of true positive predictions to the total number of instances identified as positive, incorporating both true positives and false positives as denoted by Equation 8. Whereas $recall$ measures the ability of a model to correctly identify and classify all instances of a positive class within a given dataset. It is computed as the ratio of true positives to the sum of true positives and false negatives as shown in Equation 9

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}, \tag{8}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}. \tag{9}$$

Additionally, we assess the misclassification by visualizing and interpreting the confusion matrix. Furthermore, we qualitatively analyze the top prediction under different background scenarios to understand the visual contribution towards classes. Finally, we examine the attention weights of the temporal encoder.

VI. Results

In this section we present the results of the experiments conducted on VT-Tran(Visual Temporal Transformer). Table IV provides a comprehensive comparison of VT-Tran utilising the two different types of visual context, namely, the entire frame and bounding-box level context while diversifying the parameter T and D to understand its implication on the model's performance.

A. Quantitative Analysis

Ablation study on Segment Length(T) & Alignment Dimension(D). The results of VT-Tran is presented in Table IV. The table encompasses distinct sections, each corresponding to specific segment length (T), which is further divided into sub-sections in order to make comparison between both the input representations.

Table IV indicates an increasing trend across all metrics when increasing T, specifically when fusing the skeleton trajectories with the entire-frame input representation. At $T = 8$, utilising the entire frame has achieved highest accuracy, outperforming the bounding-box representation at $D = 256$, where the model reported an accuracy of 0.228 coupled with an F_1 score of 0.308. Whereas, the use of the bounding reported a balanced accuracy of 0.184 with an F_1 score of 0.237 for $D = 64$, highlighting a 4% difference in the balanced accuracy. The same was observed at $T = 14$, where using the entire frame reported the highest accuracy at $D = 64$, where, the resulting balanced accuracy is 0.232, which again, showing a 5% difference comparing to the highest accuracy when using the bounding box representation.

As we extend our analysis to longer temporal segments($T = 20, 24, 32$), the pattern remains consistent, demonstrating the fusion of entire frame consistently surpasses the bounding-box level representation. At $T = 20$, the model achieved the highest accuracy of 0.172 when $D = 64$ when using the entire frame. In contrast, the reported balanced accuracy for the bounding box of the foregoing segment length is 0.122. Similarly, for $T = 24$, we observe that utilization of full-frame context obtained the highest accuracy of 0.235 with $D = 256$. Lastly, for $T = 32$, the model attains an accuracy level of 0.280 for $D = 256$, which, notably reported the highest performance across all metrics among all T values.

The findings show that increasing segment length increases the overall performance. However, the standard deviation of the performance indicates low variance for higher D values at shorter segment lengths, indicating stable performance at different folds. Whereas for longer segment lengths(*T=24,32*), the standard deviation indicates higher variance for larger D values, indicating longer lengths require.

Confusion Matrix. To further assess the misclassification instances made by the model, we visualize the confusion matrix shown in Figure 3. The Figures 3a and 3b depict the confusion matrices of T-Tran and VT-Tran respectively.

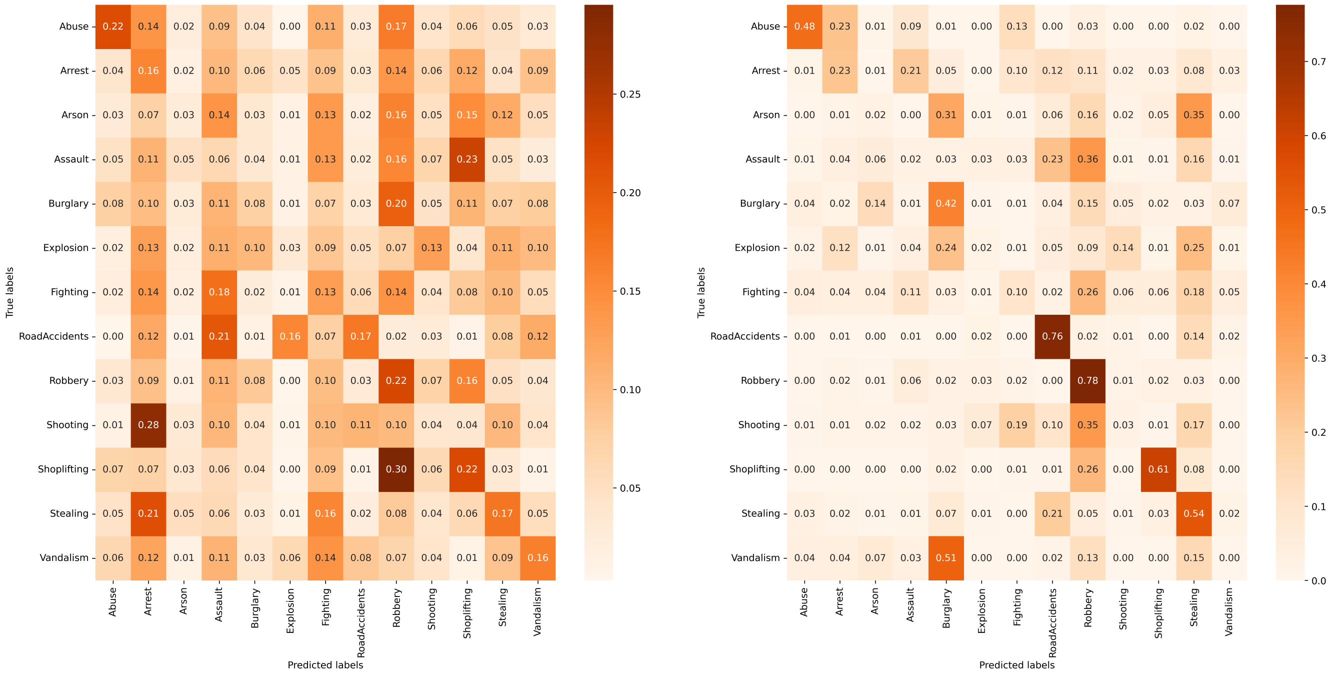
Assessing the predictions made by T-Tran(Figure 3a), the classes that achieved the highest accuracy is *Abuse*, *Robbery* and *Shoplifting* reporting an accuracy of 0.22 for all classes.

Frame Type	D	Balanced Accuracy	F1-score	Top-3 Accuracy	Top 5 Accuracy
		(a) Segment Length = 8			
Full Frame	64	0.216 ± 0.014	0.283 ± 0.016	0.536 ± 0.022	0.689 ± 0.018
	128	0.181 ± 0.006	0.237 ± 0.013	0.540 ± 0.015	0.704 ± 0.012
	256	**0.228 ± 0.008**	0.308 ± 0.021	0.555 ± 0.021	0.705 ± 0.016
	512	0.224 ± 0.004	0.306 ± 0.029	0.556 ± 0.039	0.711 ± 0.026
Bounding Box	64	0.184 ± 0.011	0.237 ± 0.026	0.546 ± 0.015	0.705 ± 0.014
	128	0.103 ± 0.011	0.153 ± 0.078	0.434 ± 0.004	0.567 ± 0.001
	256	0.177 ± 0.007	0.226 ± 0.010	0.534 ± 0.010	0.697 ± 0.009
	512	0.174 ± 0.009	0.227 ± 0.009	0.526 ± 0.009	0.691 ± 0.010
		(b) Segment Length = 14			
Full Frame	64	**0.232 ± 0.009**	0.308 ± 0.009	0.548 ± 0.035	0.697 ± 0.015
	128	0.220 ± 0.013	0.300 ± 0.016	0.557 ± 0.032	0.706 ± 0.017
	256	0.225 ± 0.009	0.298 ± 0.020	0.537 ± 0.018	0.701 ± 0.019
	512	0.151 ± 0.007	0.263 ± 0.030	0.531 ± 0.035	0.668 ± 0.016
Bounding Box	64	0.184 ± 0.012	0.246 ± 0.028	0.542 ± 0.012	0.707 ± 0.012
	128	0.181 ± 0.010	0.238 ± 0.016	0.542 ± 0.019	0.701 ± 0.015
	256	0.178 ± 0.009	0.233 ± 0.018	0.531 ± 0.018	0.696 ± 0.010
	512	0.136 ± 0.014	0.238 ± 0.010	0.474 ± 0.032	0.625 ± 0.021
		(c) Segment Length = 20			
Full Frame	**64**	**0.172 ± 0.008**	0.313 ± 0.020	0.539 ± 0.020	0.688 ± 0.019
	128	0.163 ± 0.016	0.287 ± 0.028	0.533 ± 0.012	0.678 ± 0.015
	256	0.169 ± 0.026	0.298 ± 0.016	0.547 ± 0.030	0.607 ± 0.029
	512	0.155 ± 0.013	0.258 ± 0.027	0.515 ± 0.012	0.673 ± 0.011
Bounding Box	64	0.111 ± 0.008	0.198 ± 0.027	0.403 ± 0.017	0.550 ± 0.028
	128	0.103 ± 0.030	0.287 ± 0.013	0.585 ± 0.001	0.678 ± 0.006
	256	0.122 ± 0.002	0.219 ± 0.014	0.410 ± 0.021	0.551 ± 0.029
	512	0.098 ± 0.020	0.140 ± 0.073	0.456 ± 0.029	0.603 - 0.013
		(d) Segment Length = 24			
Full Frame	64	0.225 ± 0.013	0.259 ± 0.019	0.548 ± 0.017	0.708 ± 0.016
	128	0.234 ± 0.014	0.307 ± 0.020	0.551 ± 0.034	0.698 ± 0.024
	256	**0.235 ± 0.021**	0.308 ± 0.036	0.535 ± 0.032	0.695 ± 0.018
	512	0.225 ± 0.008	0.304 ± 0.013	0.570 ± 0.018	0.724 ± 0.014
Bounding Box	64	0.097 ± 0.006	0.153 ± 0.061	0.431 ± 0.007	0.565 ± 0.011
	128	0.182 ± 0.010	0.241 ± 0.023	0.541 ± 0.015	0.700 ± 0.012
	256	0.174 ± 0.008	0.230 ± 0.013	0.528 ± 0.012	0.694 ± 0.013
	512	0.175 ± 0.009	0.222 ± 0.014	0.522 ± 0.021	0.688 ± 0.017
		(e) Segment Length = 32			
Full Frame	64	0.277 ± 0.020	0.383 ± 0.056	0.606 ± 0.055	0.728 ± 0.044
	128	0.267 ± 0.018	0.364 ± 0.035	0.605 ± 0.037	0.747 ± 0.020
	256	**0.280 ± 0.018**	0.396 ± 0.045	0.633 ± 0.044	0.760 ± 0.025
	512	0.272 ± 0.012	0.369 ± 0.033	0.609 ± 0.026	0.745 ± 0.011
Bounding Box	64	0.211 ± 0.019	0.295 ± 0.053	0.610 ± 0.026	0.755 ± 0.015
	128	0.205 ± 0.015	0.283 ± 0.038	0.604 ± 0.020	0.753 ± 0.016
	256	0.205 ± 0.017	0.282 ± 0.040	0.599 ± 0.018	0.745 ± 0.008
	512	0.201 ± 0.018	0.268 ± 0.035	0.580 ± 0.023	0.727 ± 0.014

TABLE IV: Ablation study on alignment dimension D and segment length T based on two kinds of visual input representation, namely, (a) Full Frame (b) Bounding Box. Where the **bold** text indicates the highest performance between both the input representations.

Whereas the categories *Arrest*, *Fighting*, *Road Accidents*, *Stealing* and *Vandalism* shows an accuracy more than 10%. Lastly, the categories that exhibited the lowest performance is *Shooting*, *Explosion*, *Burglary*, *Assault* and *Arson* resulting in accuracy less than 10%, where, *Arson* and *Explosion* showed the reported the lowest accuracy of 3% in both classes.

Observing the misclassified instances, the confusion matrix shows that T-Tran predicts *Arrest*, *Assault*, *Fighting* and *Robbery* for the majority of the classes, indicating a bias towards these specific classes. The category *Shoplifting* reported the highest number of misclassifications with a score of 0.30. Followed by the class *Shooting* for *Arrest* reporting an accuracy of 0.28 respectively. Interestingly, instances are observed where semantically similar classes undergo misclassifications. The category *Fighting* experiences misclassification towards the category *Assault* with an accuracy of 0.18 and vice versa. Similarly, for the class *Abuse* and *Assault*, where, 9% of instances were misclassified as *Assault*, and 11% have been



(a) Confusion matrix of T-Tran baseline results at $T = 24$ and $D = 256$

(b) Confusion matrix of VT-Tran utilising full frame input representation at $T = 32$ and $D = 64$

Fig. 3: Confusion matrix of T-Tran and VT-Tran.

classified as *Fighting*. This observation is also evident in theft-related categories, where, the classes *Robbery* and *Shoplifting* were mutually misclassified.

Upon investigating the confusion matrix of VT-Tran from Figure 3b, it becomes evident that a notable reduction in misclassifications and improvements have been observed in numerous instances. The categories *Robbery* and *Road Accidents* reported the highest accuracy of 0.78 and 0.76 respectively. In contrast to T-Tran, demonstrates an improvement of 60% and 56%. The categories *Shoplifting* and *Stealing*, *Abuse*, *Burglary* reported accuracies in the range of 40%-60%, where, *Shoplifting* yielded the highest accuracy of 0.61, whereas *Burglary* reported the relatively lowest score of 0.42. Notably, the performance of these categories significantly outperforms that of T-Tran. In contrast, the remaining categories, namely *Arson*, *Assault*, *Explosion*, *Shooting* and *Vandalism* exhibited significantly lower accuracy levels in comparison to the rest of the classes, all falling below the 10%. Where, the case of *Vandalism*, reported the lowest accuracy, where no instances being accurately predicted.

By assessing the misclassification instances of VT-Tran, the category *Abuse* is often predicted as *Arrest*, *Assault*, and *Fighting* with accuracies of 0.23, 0.09, and 0.13, respectively. Within the *Assault* category, misclassifications are evident for *Robbery* and *Road Accidents* with accuracies of 0.36 and 0.23.

In the *Fighting* category, misclassifications occur for *Robbery*, *Stealing*, and *Assault*, yielding accuracy scores of 0.26, 0.18, and 0.10. This can be attributed to the fact that these categories share visual and temporal similarities, particularly the challenges encountered by the model in distinguishing between *Assault* and *Fighting*. Unlike, *Shoplifting*, the category *Robbery* and *Stealing*, where the former includes videos depicting theft incidents and, at times, physical altercations involving the perpetrators and victims, typically occurring in both retail and outdoor environments. Hence, contributing to the overall complexity in differentiating between categories.

Additionally, the *Arrest* categories, shows analogous misclassifications, where *Assault*, *Road Accidents* and *Robbery* is associated with accuracy scores of 0.23, 0.12 and 0.11 respectively. The misclassifications for *Assault*, stem from situations where authorities are seen grappling with or engaging in combat against the perpetrators. Furthermore, the consideration of *Road Accidents* is prompted by the visual similarities shared with other categories, primarily due to the presence of vehicles in both classes. This common element presents a challenge for the model in differentiating between them.

Concerning theft-related categories, a similar observation for the categories *Shoplifting* and *Stealing*, where, the misclassification for the former is seen primarily seen towards the *Robbery* class, where approximately, 26% of cases were inaccurately classified. Additionally, 8% for the class *Stealing*. Regarding the category *Stealing*, a substantial portion of 21% has been misclassified as *Road Accidents*. For the justification of misclassification in this category, the perpetrators were seen stealing vehicles. The presence of vehicles in the scenes leads

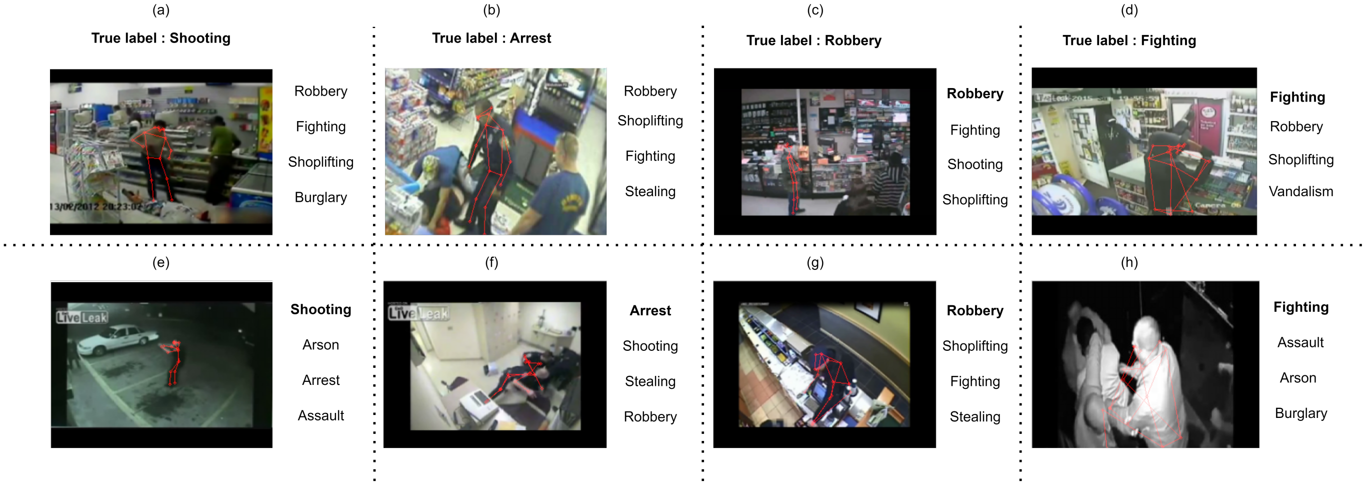


Fig. 4: Predictions made by VT-Tran($T = 32$ and $D = 64$) for different classes containing similar background. In the context where the highlighted subject(in the color red) serves to denote the trajectory for which the model is making predictions on.

the model to be overly confident in assigning them to this class.

Comparison of VT-Tran with random guessing. The confusion matrix shows in many cases the model exhibited accuracy greater than random guessing(please refer Table I). Conversely, substantial categories are seen to perform lower than random guess. Notably, these categories include *Vandalism*, *Assault*, *Fighting* and *Explosion*. Where, the confusion matrix indicates that no accurate predictions were made for the former, which by default is considered worse than random guessing. The category *Assault* reports an accuracy of 0.02, where the random guessing probably for this class is 0.055, implying the model's performance for this class is significantly subpar. Similarly, the class *Fighting*, shows an accuracy of 0.10, where the probability of random guessing for this class is 0.14, highlighting 4% difference. Finally, the *Explosion* class reported an accuracy 0.02, which is marginally higher than its corresponding random guessing probability of 0.015.

Comparison with Baseline & Best performing model. Table V presents comparative results of VT-Tran and T-Tran. In addition, we include the performance of just the spatial encoder to understand the significance of visual contribution. The table indicates that VT-Tran surpassed T-Tran by 16%. However, results show the performance mainly comes from the spatial encoder, showing a 4% difference compared to VT-Tran. Nevertheless, the outcomes reveal that the primary source of performance mainly comes from the spatial encoder, showing a 4% difference compared to VT-Tran.

Model	Balanced Accuracy	F1-Score
T-Tran	0.121 ± 0.030	0.205 ± 0.030
Spatial Encoder	0.240 ± 0.033	0.351 ± 0.031
VT-Tran	**0.280 ± 0.018**	0.396 ± 0.045

TABLE V: Performance comparison with the baseline approaches.

B. *Qualitative Analysis*

To qualitatively assess the model, we investigate how the background affects the predictions made. Next, we compute the attention score of the temporal encoder of VT-Tran.

Feature & Context Overlap. This section conducts a qualitative evaluation of the model's performance under varying background image conditions. To streamline this analysis, we investigate the behavior of the model in the context of crimes pertaining to retail and non-retail environments. Figure 4 illustrates the top predictions made by the model, where each sub-figure corresponds to the visual context and the trajectory(marked in red) on which the predictions are. Figures 4a,4b,4c,4d depicts scenarios of actions taking place within a retail-based environment, while Figures 4e,4f,4g,4h illustrates the same in a non-retail environment.

From Figure 4a, the subject is seen engaging in crime under a retail environment, where the subject is wielding a firearm and pointing towards the victim under a retail environment. Whereas for Figure 4e, the subject is seen wielding a firearm more specifically in an outdoor environment. Although the confusion matrix shows the performance for this class is marginally higher than random guessing, it is interesting to observe the top predictions made by the model. The predictions indicate the influence of the background, as we see *Robbery*, *Shoplifting* among the top predictions, indicating visual overlap among classes.

The observation remains consistent for Figure 4b, 4c and 4d where we see *Shoplifting* and *Robbery* in the retail environments. Interestingly, in Figure 4c, the subject is observed wielding a firearm in a retail environment, and the true label for the corresponding figure is *Robbery*. While the model successfully predicts the class for this instance, it is also important to recognize that the top predictions *Fighting*,*Shooting*, and *Shoplifting*, indicate the complexity in deciding the crime.

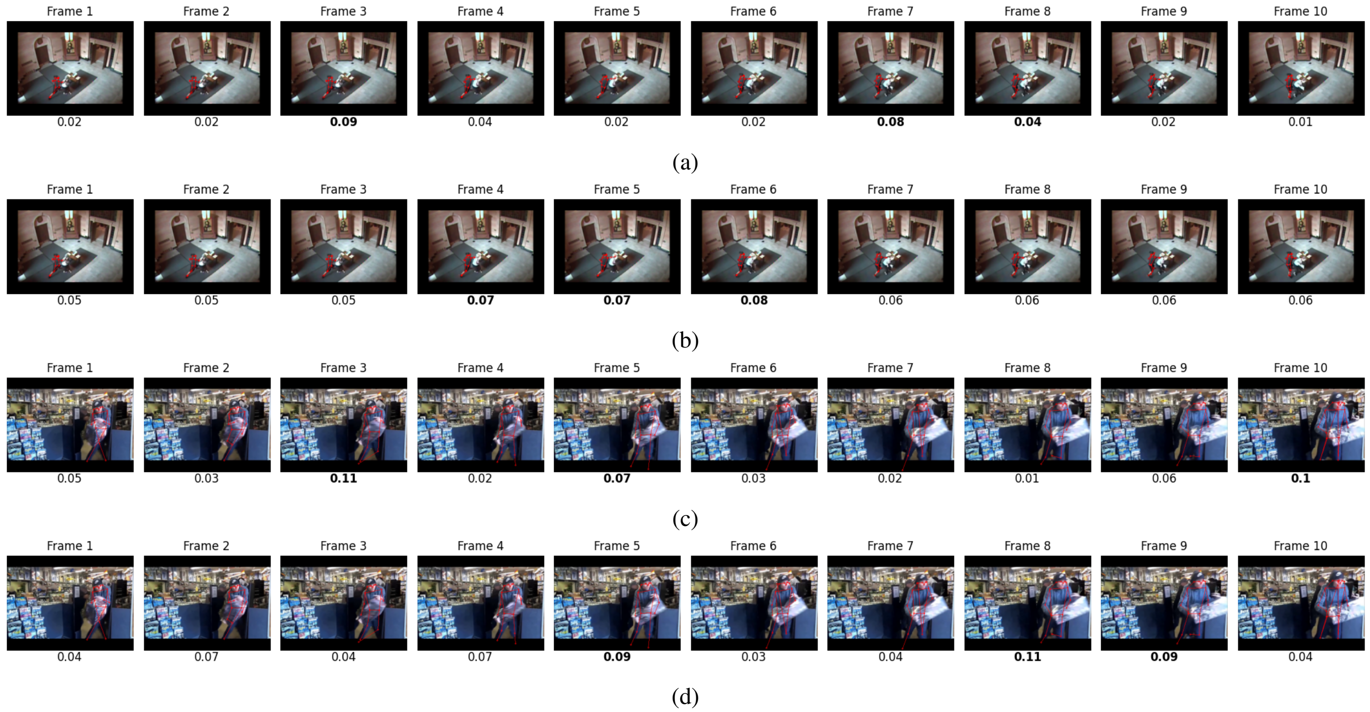


Fig. 5: Mean attention score of all the heads for the T-Tran and the VT-Tran. Figure 5a and 5c are the attention scores of the T-Tran and Figure 5b and 5d depicts the attention score of VT-Tran. For the visualization purpose, we select the classes *Assault* and *Shoplifting*. The drawn skeleton trajectory indicates the subject at which the model is making prediction to.

Influence of visual context on the attention heads within the temporal encoder. We further analyze the internal representation of VT-Tran, namely the attention heads of the temporal encoder. The attention score is determined by averaging the values across all individual heads. Subsequently, the *class* token is used as the final score. We compare the attention scores of the baseline(T-Tran) in parallel to empirically observe significance when visual cues are fused. For this analysis, we examine the attention of the categories *Abuse* and *Stealing*. Figure 5 illustrates the first 10 frames alongside its corresponding attention score. Additionally, Figure 5a and 5c displays the attention scores of T-Tran, while Figure 5b and 5d depicts the attention score of VT-Tran.

Analyzing Figure 5a and Figure 5b for the class *Abuse*, it becomes evident that the attention score of the temporal encoder achieves peak attention within the temporal span, spanning frames from 4 to 7 with an attention score of 0.07, 0.07, 0.08 and 0.06 respectively. Interestingly, the temporal window aligns with the action taking place. Whereas for T-Tran, the attention score for the foregoing temporal span is 0.04, 0.02, 0.02, and 0.08 respectively.

Upon examining the *Stealing* category, Figure 5c and 5d depicts a scenario wherein an individual is seen appropriating an item from a retail store. Unlike the previous observation where increased attention is seen over a span of multiple frames. In this specific instance, peak attention is observed only at certain frames. Frames 5, 8, and 9 attained the highest attention, where the obtained attention score is 0.09, 0.11, and 0.09 respectively. Conversely, in the context of T-Tran, the highest attention is particularly concentrated on frames 3 and 1, exhibiting attention scores of 0.11 and 0.1, respectively.

VII. Discussion

Effect of Visual Context. In this study, we have explored two different kinds of input representation. The results from Table IV indicate that fusion of the entire frame as visual context reported the highest balanced accuracy of 0.280 with an F_1 score of 0.396 and Top-3 and Top-5 accuracy of 0.633 and 0.760 respectively. Evidently, it surpasses the baseline performance. By considering the entirety of the frame, the model learns a broader range of visual information and potentially extracts more relevant and nuanced details for its learning process. However, a problem arises because of the noise the background can contain, causing the model to learn features that do not represent the class. Conversely, using a more localized visual cue eliminates a substantial portion of the background, resulting in fewer features to learn from. Hence, resulting in a lower performance.

The analysis of the data presented in Table V reveals that the spatial encoder alone reports a balanced accuracy of 0.240, indicating over-fitting towards the RGB modality. To address this imbalance, utilizing a weighting procedure within the loss function to add more weight toward the output of the temporal stream may aid the training process.

In a comprehensive analysis of misclassifications, VT-Tran indeed demonstrates significant improvements compared to

T-Tran. Reduced misclassifications and improved accuracy are evident in many instances, which is a positive outcome. However, it's worth noting that both approaches still struggle with differentiating between semantically similar classes, highlighting the inherent complexity of the problem.

To address these challenges and further improve the overall classification performance, one potential approach is to utilise Motion-flow images. Motion-flow images can capture temporal information and movement patterns, which may aid in distinguishing between classes that share visual similarities. By adding this additional modality, the model could gain a better understanding of the dynamics within the video data, potentially leading to more accurate and reliable predictions

Ablation study of Segment Length & Alignment Dimension. Apart from the input representation, two tuneable parameters were investigated. Segment length T and alignment dimension D. Empirically, it is observed that as T increases, the performance increases across all metrics. We speculate the reason for this behavior is that as we increase T, trajectories shorter than the specified length are removed, essentially discarding any important features shorter than the specified length(please refer to Table I), making the model learn fewer features compared to shorter segment lengths, thereby increasing accuracy. Conversely, using shorter segment lengths increases the number of data points, but they may not fully represent the class itself, hence, introducing noise to the model.

However, it is to be noted that the standard deviation in performance shows an increasing trend when varying the segment length, indicating high sensitivity towards the dataset. It can plausibly be attributed to the fact that certain videos benefit from shorter or longer segment lengths, indicates a trade-off in choosing the right segment length for a specific D. To accommodate a larger number of samples in shorter segment lengths, it may be necessary to employ higher values of the variable D, and a similar requirement arises for longer segment lengths, although the number of samples decreases.

Limitations. During this study, several limitations have been encountered, including:

1) The dataset used in this study is weakly labeled, specifically based on video-level annotations. This labeling approach assumes that every individual within the scene is engaged in the action being categorized. Consequently, the accuracy of such annotations may be compromised.
2) The data collection process relies on a pose extraction algorithm to obtain key information. However, the accuracy of this extraction is contingent upon the quality of the input video. Variations in video quality may introduce inaccuracies in the extracted pose data.
3) In this study, only the middle frame is taken into consideration. This approach may not always capture the full contextual information required for precise action classification, potentially leading to misclassifications.
4) Since our experiments are based on a new split, it makes it difficult to have fair comparisons with the existing state of the art.

Deployment Readiness. Although the fusion of visual context with skeleton trajectory has shown to have an overall improvement compared to the baseline approach. A comprehensive assessment of the balanced accuracy and the confusion matrix reveals the model's performance is less than optimal for categorizing violent categories but also in several other categories, which is an essential need for such a system. Consequently, this suggests that further refinements or alternative strategies such as exploring different architectures and feature engineering strategies are necessary to ensure the readiness of the model for deployment.

VIII. Conclusion

For this study, we investigated the fusion of RGB with skeleton modality for crime recognition. Additionally, we explore two kinds visual representations, namely, full-frame and bounding box level. Notably, the highest performance is achieved when using full-frame visual context. Additionally, we observed the fusion of visual context improved the performance on the existing state-of-the-art approach reporting an accuracy of 28%, indicating the benefits of introducing a visual modality. While the performance indicates the model is not ready for deployment, further experimentation by deriving new features from the existing modality or exploration of different model architecture is required for further improvement.

IX. Future Work

1) Employing motion-flow techniques like MHI and MEI to convert the skeleton trajectory to an image[15], and subsequently fusing this image representation with RGB frames.
2) Diverse fusion approaches should be explored, including early fusion and the utilization of a cross-attention mechanism.
3) By visualizing the saliency map of the spatial encoder, we've observed the model focusing on the boundaries of the image. In the future work, center crop during the preprocessing phase would be essential.
4) In numerous study, cosine annealing learning rate scheduler has been employed. This scheduler initially employs a significantly elevated learning rate, which is subsequently diminished at a relatively swift pace until reaching a minimum threshold before undergoing a rapid increment once more.
5) In the context of class imbalance, future work may benefit from the consideration of class weighting strategies in the loss function, where the assignment of weights to classes is determined by their respective frequencies, as this approach has the potential to improve the performance, as we are penalizing the classes with high distribution.

References

[1] Amanda L. Thomas et al. "The internationalisation of cctv surveillance: Effects on crime and implications for emerging technologies". In: *International Journal of Comparative and Applied Criminal Justice* 46.1 (2022), pp. 81–102. DOI: 10.1080/01924036.2021.1879885. eprint: https://doi.org/10.1080/01924036.2021.1879885. URL: https://doi.org/10.1080/01924036.2021.1879885.

[2] Khan Muhammad et al. "Human Action Recognition Using Attention Based LSTM Network with Dilated CNN Features". In: *Future Gener. Comput. Syst.* 125.C (Dec. 2021), pp. 820–830. ISSN: 0167-739X. DOI: 10.1016/j.future.2021.06.045. URL: https://doi.org/10.1016/j.future.2021.06.045.

[3] Matteo Bregonzio, Shaogang Gong, and Tao Xiang. "Recognising action as clouds of space-time interest points". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1948–1955. DOI: 10.1109/CVPR.2009.5206779.

[4] Yong Du, Wei Wang, and Liang Wang. "Hierarchical recurrent neural network for skeleton based action recognition". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1110–1118. DOI: 10.1109/CVPR.2015.7298714.

[5] Zhou Shuchang. *A Survey on Human Action Recognition*. 2022. arXiv: 2301.06082 [cs.CV].

[6] Xiangyu Li et al. *Trear: Transformer-based RGB-D Egocentric Action Recognition*. 2021. arXiv: 2101.03904 [cs.CV].

[7] Hao-Shu Fang et al. *AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time*. 2022. arXiv: 2211.03375 [cs.CV].

[8] Zhe Cao et al. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*. 2019. arXiv: 1812.08008 [cs.CV].

[9] Xu Weiyao et al. "Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition". In: *IEEE Sensors Journal* 21.17 (2021), pp. 19157–19164. DOI: 10.1109/JSEN.2021.3089705.

[10] Haodong Duan et al. *Revisiting Skeleton-based Action Recognition*. 2022. arXiv: 2104.13586 [cs.CV].

[11] Oriol Vinyals et al. *Show and Tell: A Neural Image Caption Generator*. 2015. arXiv: 1411.4555 [cs.CV].

[12] Chitwan Saharia et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: 2205.11487 [cs.CV].

[13] Santosh Kumar Yadav et al. "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions". In: *Knowledge-Based Systems* 223 (2021), p. 106970. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2021.106970. URL: https://www.sciencedirect.com/science/article/pii/S0950705121002331.

[14] Pushpajit Khaire, Praveen Kumar, and Javed Imran. "Combining CNN streams of RGB-D and skeletal data for human activity recognition". In: *Pattern Recognition Letters* 115 (2018). Multimodal Fusion for Pattern Recognition, pp. 107–116. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2018.04.035. URL: https://www.sciencedirect.com/science/article/pii/S0167865518301636.

[15] Pratishtha Verma, Animesh Sah, and Rajeev Srivastava. "Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition". In: *Multimedia Systems* 26.6 (Dec. 2020), pp. 671–685. ISSN: 1432-1882. DOI: 10.1007/s00530-020-00677-2. URL: https://doi.org/10.1007/s00530-020-00677-2.

[16] Xiaoguang Zhu et al. "Skeleton Sequence and RGB Frame Based Multi-Modality Feature Fusion Network for Action Recognition". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 18.3 (Mar. 2022). ISSN: 1551-6857. DOI: 10.1145/3491228. URL: https://doi.org/10.1145/3491228.

[17] Guiyu Liu et al. "Action Recognition Based on 3D Skeleton and RGB Frame Fusion". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019, pp. 258–264. DOI: 10.1109/IROS40897.2019.8967570.

[18] Sijie Song et al. "Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection". In: *IEEE Transactions on Image Processing* 27.7 (2018), pp. 3459–3471. DOI: 10.1109/TIP.2018.2818328.

[19] Fabien Baradel, Christian Wolf, and Julien Mille. "Human Activity Recognition with Pose-driven Attention to RGB". In: *BMVC 2018 - 29th British Machine Vision Conference*. Newcastle, United Kingdom, Sept. 2018, pp. 1–14. URL: https://inria.hal.science/hal-01828083.

[20] Kayleigh Boekhoudt et al. *HR-Crime: Human-Related Anomaly Detection in Surveillance Videos*. Version V1. 2021. DOI: 10.34894/IRRDJE. URL: https://doi.org/10.34894/IRRDJE.

[21] A.M. Joseph. *Investigating vision transformers for human activity recognition from skeletal data*. Jan. 2023. URL: http://essay.utwente.nl/94291/.

[22] Anurag Arnab et al. *ViViT: A Video Vision Transformer*. 2021. arXiv: 2103.15691 [cs.CV].

[23] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[24] Shaoqing Ren et al. *Object Detection Networks on Convolutional Feature Maps*. 2016. arXiv: 1504.06066 [cs.CV].

[25] Ren Wu et al. *Deep Image: Scaling up Image Recognition*. 2015. arXiv: 1501.02876 [cs.CV].

[26] Kaiming He et al. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV].

[27] Lena Gorelick et al. "Actions as Space-Time Shapes". In: *IEEE Transactions on Pattern Analysis and Machine*

Intelligence 29.12 (2007), pp. 2247–2253. DOI: 10.1109/TPAMI.2007.70711.

[28] Jingen Liu and Mubarak Shah. “Learning human actions via information maximization”. In: June 2008. DOI: 10.1109/CVPR.2008.4587723.

[29] Reshma Khemchandani and Sweta Sharma. “Robust least squares twin support vector machine for human activity recognition”. In: *Applied Soft Computing* 47 (2016), pp. 33–46. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2016.05.025. URL: https://www.sciencedirect.com/science/article/pii/S1568494616302265.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[31] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[32] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[33] Karen Simonyan and Andrew Zisserman. *Two-Stream Convolutional Networks for Action Recognition in Videos*. 2014. arXiv: 1406.2199 [cs.CV].

[34] Andrej Karpathy et al. “Large-scale Video Classification with Convolutional Neural Networks”. In: *CVPR*. 2014.

[35] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[36] Jeff Donahue et al. *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. 2016. arXiv: 1411.4389 [cs.CV].

[37] Lin Sun et al. *Lattice Long Short-Term Memory for Human Action Recognition*. 2017. arXiv: 1708.03958 [cs.CV].

[38] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. *Action Recognition using Visual Attention*. 2016. arXiv: 1511.04119 [cs.LG].

[39] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. *LSTA: Long Short-Term Attention for Egocentric Action Recognition*. 2019. arXiv: 1811.10698 [cs.CV].

[40] Chenlin Zhang, Jianxin Wu, and Yin Li. *ActionFormer: Localizing Moments of Actions with Transformers*. 2022. arXiv: 2202.07925 [cs.CV].

[41] Jiewen Yang et al. “Recurring the Transformer for Video Action Recognition”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 14043–14053. DOI: 10.1109/CVPR52688.2022.01367.

[42] James Wensel, Hayat Ullah, and Arslan Munir. “ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos”. In: *IEEE Access* (2023), pp. 1–1. DOI: 10.1109/ACCESS.2023.3293813.

[43] Altaf Hussain et al. “Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos”. In: *Computational Intelligence and Neuroscience* 2022 (Apr. 2022), p. 3454167. ISSN: 1687-5265. DOI: 10.1155/2022/3454167. URL: https://doi.org/10.1155/2022/3454167.

[44] Wenhai Wang et al. *InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions*. 2023. arXiv: 2211.05778 [cs.CV].

[45] Rujing Yue, Zhiqiang Tian, and Shaoyi Du. “Action recognition based on RGB and skeleton data sets: A survey”. In: *Neurocomputing* 512 (2022), pp. 287–306. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2022.09.071. URL: https://www.sciencedirect.com/science/article/pii/S0925231222011596.

[46] Hao-Shu Fang et al. “RMPE: Regional Multi-person Pose Estimation”. In: *ICCV*. 2017.

[47] Yufei Xu et al. *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*. 2022. arXiv: 2204.12484 [cs.CV].

[48] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[49] Zehua Sun et al. “Human Action Recognition From Various Data Modalities: A Review”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1–20. DOI: 10.1109/tpami.2022.3183112. URL: https://doi.org/10.1109%2Ftpami.2022.3183112.

[50] Wentao Zhu et al. “Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (Mar. 2016). DOI: 10.1609/aaai.v30i1.10451. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10451.

[51] Jun Liu et al. “Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks”. In: *IEEE Transactions on Image Processing* 27.4 (Apr. 2018), pp. 1586–1599. ISSN: 1941-0042. DOI: 10.1109/TIP.2017.2785279.

[52] Sijie Yan, Yuanjun Xiong, and Dahua Lin. *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition*. 2018. arXiv: 1801.07455 [cs.CV].

[53] Lei Shi et al. *Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition*. 2019. arXiv: 1805.07694 [cs.CV].

[54] Maosen Li et al. “Actional-structural graph convolutional networks for skeleton-based action recognition”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3595–3603.

[55] Vittorio Mazzia et al. "Action Transformer: A self-attention model for short-time pose-based human action recognition". In: *Pattern Recognition* 124 (Apr. 2022), p. 108487. DOI: 10.1016/j.patcog.2021.108487. URL: https://doi.org/10.1016%2Fj.patcog.2021.108487.

[56] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. "Spatial Temporal Transformer Network for Skeleton-Based Action Recognition". In: *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, pp. 694–701. DOI: 10.1007/978-3-030-68796-0_50. URL: https://doi.org/10.1007%2F978-3-030-68796-0_50.

[57] Qingtian Wang et al. *IIP-Transformer: Intra-Inter-Part Transformer for Skeleton-Based Action Recognition*. 2021. arXiv: 2110.13385 [cs.CV].

[58] Jun Kong, Yuhang Bian, and Min Jiang. "MTT: Multi-Scale Temporal Transformer for Skeleton-Based Action Recognition". In: *IEEE Signal Processing Letters* 29 (2022), pp. 528–532. DOI: 10.1109/LSP.2022.3142675.

[59] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. "Early vs Late Fusion in Multimodal Convolutional Neural Networks". In: *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. 2020, pp. 1–6. DOI: 10.23919/FUSION45008.2020.9190246.

[60] Jianan Li et al. "SGM-Net: Skeleton-guided multimodal network for action recognition". In: *Pattern Recognition* 104 (2020), p. 107356. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2020.107356. URL: https://www.sciencedirect.com/science/article/pii/S003132032030159X.

[61] Gueter Josmy Faure, Min-Hung Chen, and Shang-Hong Lai. "Holistic Interaction Transformer Network for Action Detection". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 3340–3350.

[62] Christoph Feichtenhofer et al. *SlowFast Networks for Video Recognition*. 2019. arXiv: 1812.03982 [cs.CV].

[63] Yanhao Jing and Feng Wang. "TP-VIT: A Two-Pathway Vision Transformer for Video Action Recognition". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 2185–2189. DOI: 10.1109/ICASSP43922.2022.9747276.

[64] Jing Shi et al. "A Novel Two-Stream Transformer-Based Framework for Multi-Modality Human Action Recognition". In: *Applied Sciences* 13.4 (2023). ISSN: 2076-3417. DOI: 10.3390/app13042058. URL: https://www.mdpi.com/2076-3417/13/4/2058.

[65] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[66] Kayleigh Boekhoudt et al. *HR-Crime: Human-Related Anomaly Detection in Surveillance Videos*. 2021. DOI: 10.34894/IRRDJE. URL: https://dataverse.nl/citation?persistentId=doi:10.34894/IRRDJE.

[67] Waqas Sultani, Chen Chen, and Mubarak Shah. *Real-world Anomaly Detection in Surveillance Videos*. 2019. arXiv: 1801.04264 [cs.CV].

[68] Yuliang Xiu et al. *Pose Flow: Efficient Online Pose Tracking*. 2018. arXiv: 1802.00977 [cs.CV].

[69] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

Appendix

A. Related Work

1. Vision Transformer: Figure 6 illustrates the working of vision transformer. Where, the image is first divided into non-overlapping patches. Consequently, the patches are projected to linearly projected to a higher dimension. The output of the projection is then passed to the temporal encoder to learn patch based descriptors.

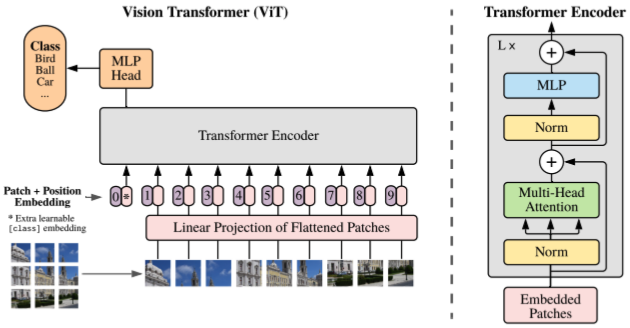


Fig. 6: Vision Transformer(ViT) architecture[69]

2. Tubelet Embedding: Figure 7 shows how 3D convolutions are employed to extract spatio-temporal tokens in videos.

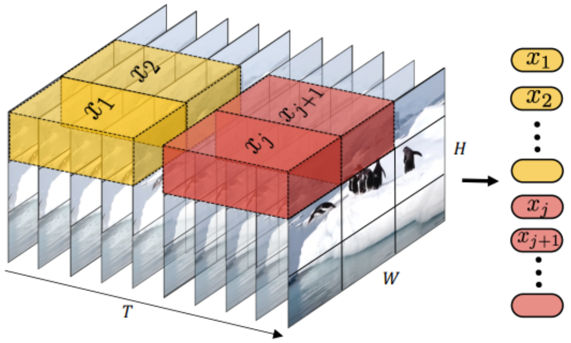


Fig. 7: Tubelete Embedding

B. Input Representation

1. Overall input representation

Given a skeleton trajectory and its corresponding RGB frames, we utilise only the middle frame as the context.

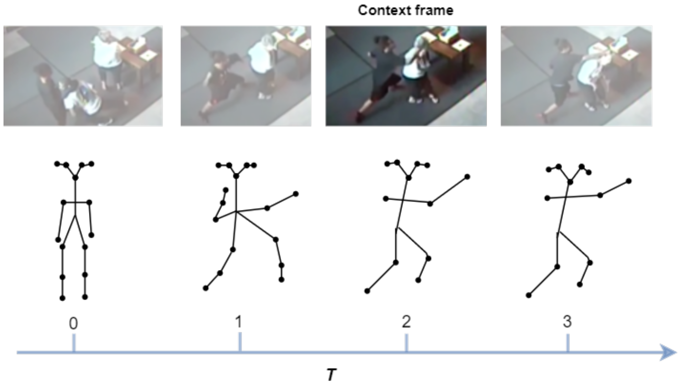


Fig. 8: Extracting the context frame given a sequence of skeleton trajectories. Where the visual context frame for a sequence of trajectory is the middle-frame of the sequence

2. Bounding box input representation

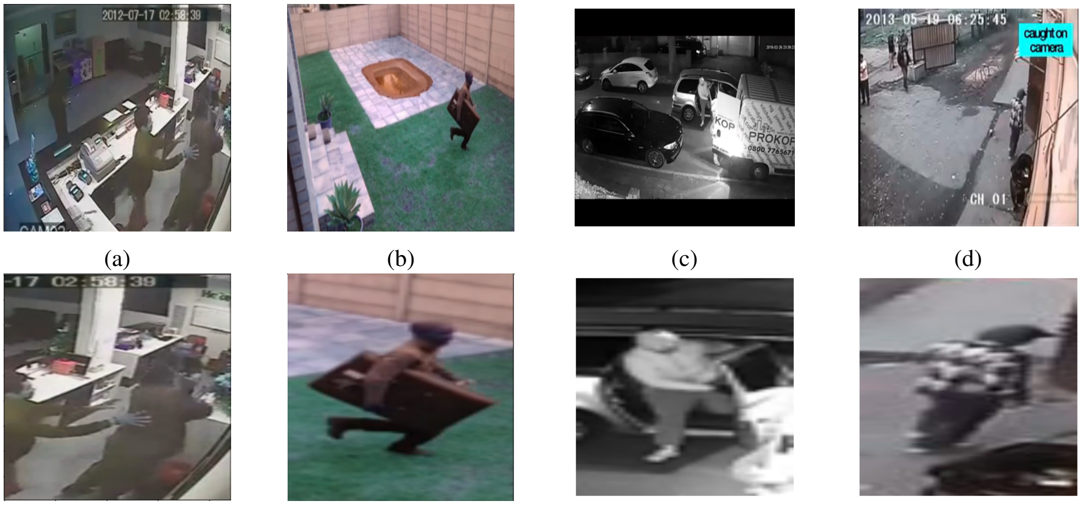


Fig. 9: On the first row(a,b,c,d) we have full frame context of different videos. The second row(e,f,g,h) corresponds to the extracted ROI with scaling factor $\alpha = 0.5$

C. Model

1. Temporal Encoder

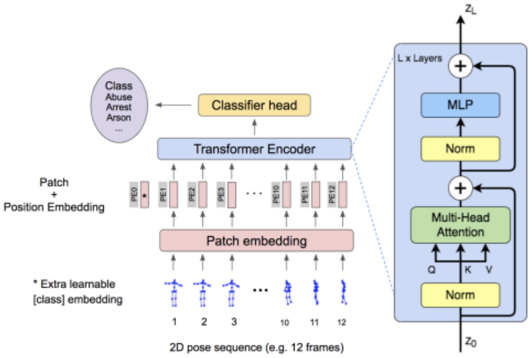


Fig. 10: Temporal transformer architecture[20]

2. Spatial Encoder

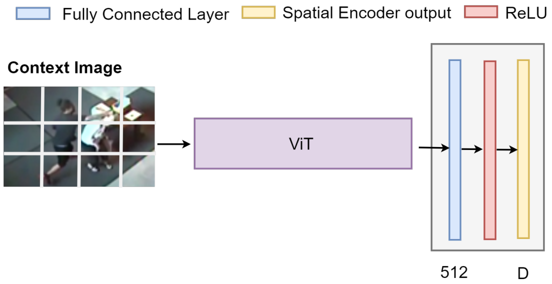


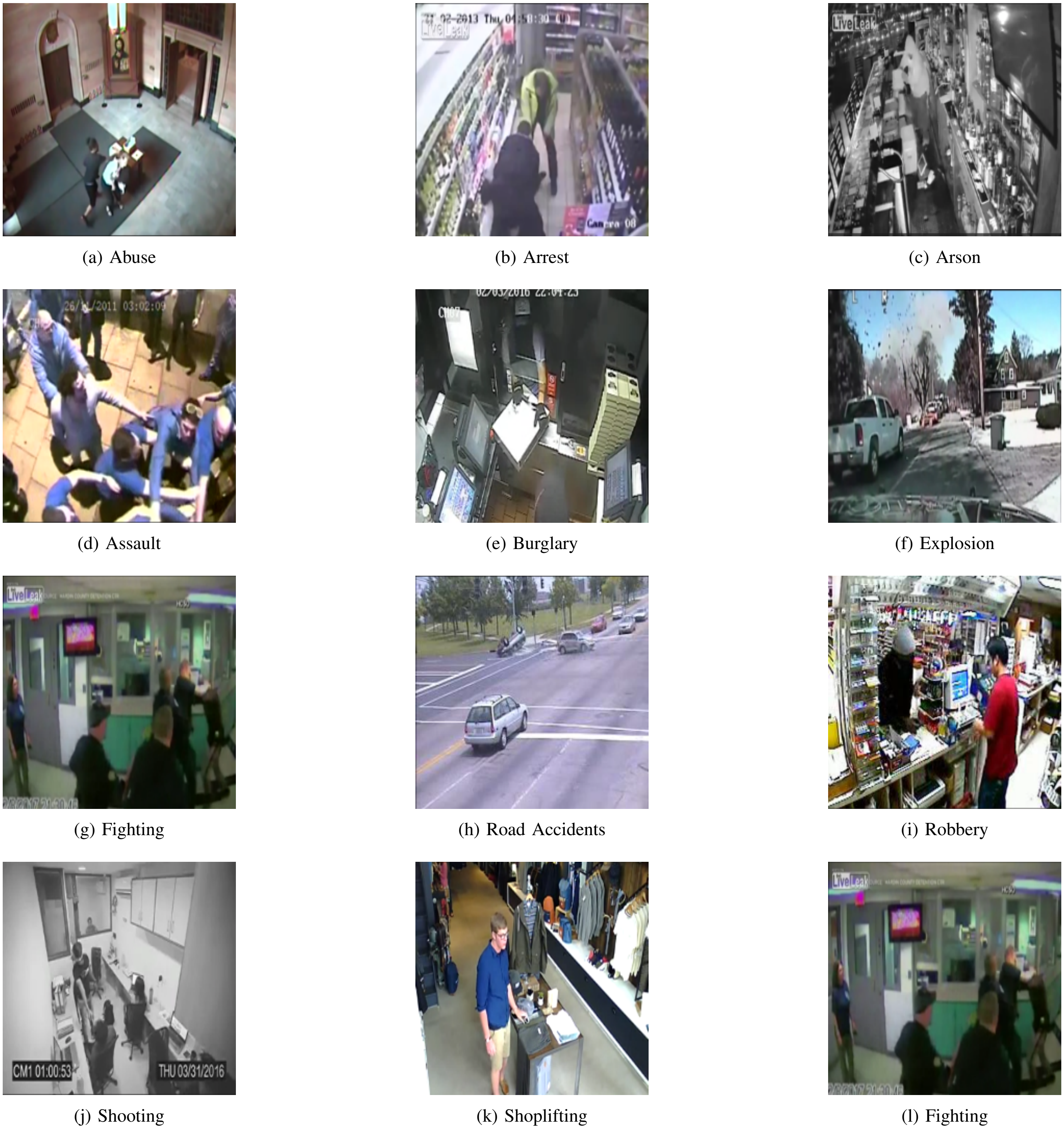
Fig. 11: Spatial Encoder, namely, ViT. Which takes in non-overlapping image patches as input and passes through ViT to extract patch based features. Which is further refined by a series of fully connected layers

D. HRC Class Dataset

Number of Videos & Trajectories			Video distribution	
Category	Total Videos	Number of Trajectories	**Train**	**Test** Video
Abuse	50	718	40	10
Arrest	50	1465	40	10
Arson	50	373	40	10
Assault	100	1210	80	20
Burglary	100	856	80	20
Explosion	50	513	40	10
Fighting	50	1640	40	10
Robbery	150	2011	120	30
Road Accidents	150	982	120	30
Shoplifting	50	1666	40	10
Stealing	100	1418	80	20
Shooting	50	830	40	10
Vandalism	50	787	40	10

TABLE VI: Comprehensive overview of the HRC dataset

2. Visualization of HRC dataset

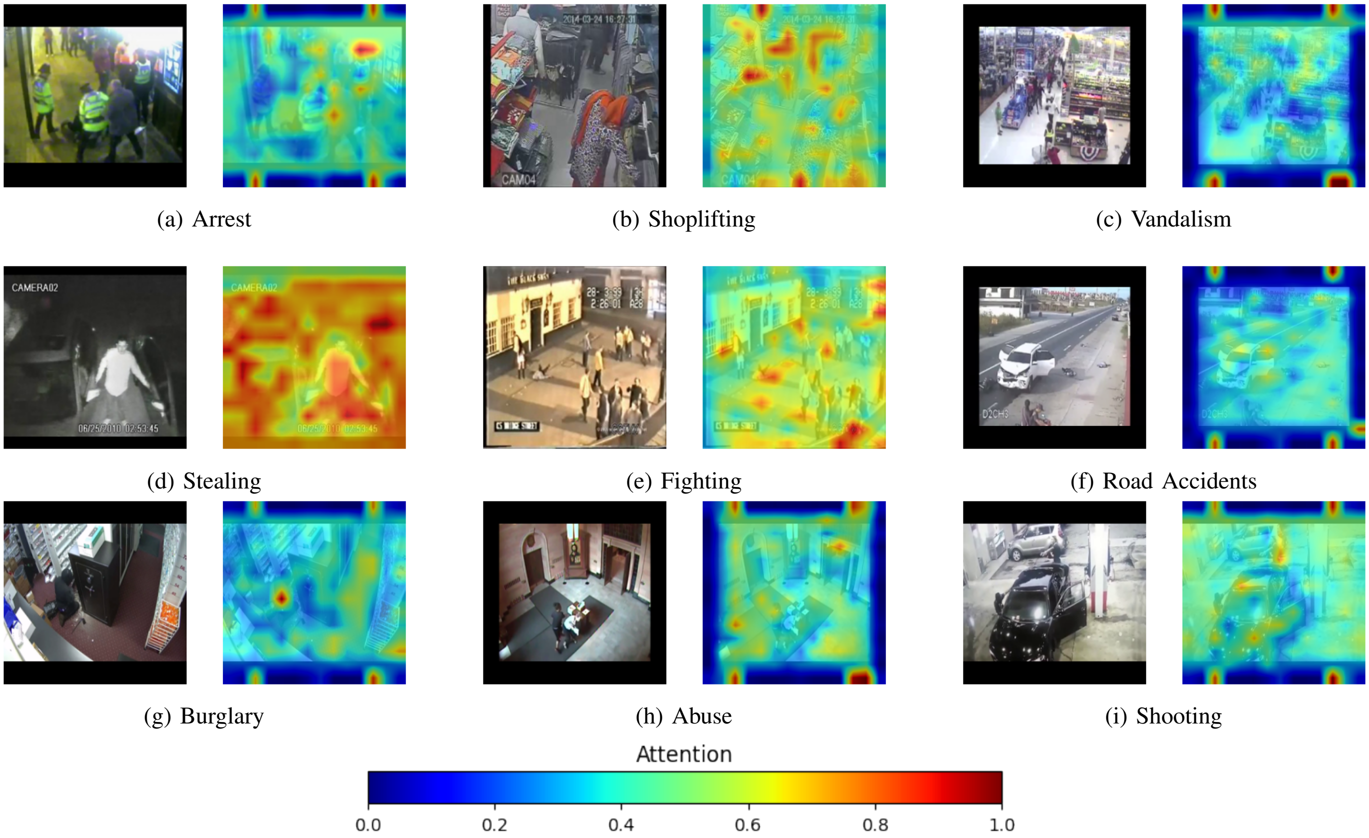


(a) Abuse (b) Arrest (c) Arson (d) Assault (e) Burglary (f) Explosion (g) Fighting (h) Road Accidents (i) Robbery (j) Shooting (k) Shoplifting (l) Fighting

Fig. 12: Class visualization HRC dataset

E. Results

1. Spatial encoder attention heatmap


(a) Arrest

(b) Shoplifting

(c) Vandalism

(d) Stealing

(e) Fighting

(f) Road Accidents

(g) Burglary

(h) Abuse

(i) Shooting

Fig. 13: Visualizing attention head of the spatial stream