This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced UniformerV2 and Modified Efficient Additive Attention

Damith Chamalke Senadeera[1,2], Xiaoyun Yang[3], Dimitrios Kollias[1,2], Gregory Slabaugh[1,2]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]Queen Mary's Digital Environment Research Institute (DERI), London, UK
[3]Remark AI UK Limited, London, UK
d.c.senadeera@qmul.ac.uk, xiaoyun.yang@remarkai.co.uk, {d.kollias, g.slabaugh}@qmul.ac.uk

Abstract

In this paper we introduce CUE-Net, a novel architecture designed for automated violence detection in video surveillance. As surveillance systems become more prevalent due to technological advances and decreasing costs, the challenge of efficiently monitoring vast amounts of video data has intensified. CUE-Net addresses this challenge by combining spatial ***C****ropping with an enhanced version of the* ***U****niformerV2 architecture, integrating convolutional and self-attention mechanisms alongside a novel Modified* ***E****fficient Additive Attention mechanism (which reduces the quadratic time complexity of self-attention) to effectively and efficiently identify violent activities. This approach aims to overcome traditional challenges such as capturing distant or partially obscured subjects within video frames. By focusing on both local and global spatio-temporal features, CUE-Net achieves state-of-the-art performance on the RWF-2000 and RLVS datasets, surpassing existing methods. The source code is available at* [1].

1. Introduction

According to the World Bank, there has been an increase in the worldwide crime rate in the last five years [18]. Surveillance cameras are often deployed to help deter violence, provide real-time monitoring and collect evidence of criminal or violent activity. Thanks to advances in technology, surveillance systems are becoming increasingly affordable and easier to deploy. As the number of these deployed surveillance cameras grows, it rapidly becomes expensive and challenging for human operators to manually monitor camera feeds [22, 31]. Therefore there is substantial need for automated approaches to monitor surveillance cameras, simplifying the process of Violence Detection (VD) in a more accurate and an efficient manner [21, 31].

To respond to the challenge of efficient, automated violence detection from video, effective computer vision methods are required. Deep learning techniques such as Convolutional Neural Networks (CNNs) and more recently Transformer-based architectures have shown a great promise in solving computer vision related automated violence detection [21, 22, 31]. The success of violence detection is highly dependent on the objects and people present in the captured videos [22, 31]. Detection is difficult when the relevant features of the violent incidents are not captured properly, for example when the people involved in the violent incident are far away and occupy only a small part of the frame, as seen in one of the example videos from the RWF-2000 dataset [4] in Fig. 2 (a). Although different mechanisms have been explored for automated violence detection, the opportunity for improvement remains due to challenges such as tracking and extracting fast moving people or objects involved in violence, low resolution scenarios and occlusion-related issues [31].

Another research question relates to finding an effective and a robust processing architecture for violence detection in videos. An ideal architecture would be simultaneously capable of capturing the locally and globally important features across the temporal and the spatial dimensions. As discussed in [29, 30] CNN-based architectures have shown to better capture locally important features but not the globally important ones; whereas [14] argues that the self-attention mechanism in the transformer architecture seems to better capture globally important features temporally. However, transformer architectures may struggle with video data due to their quadratic computational complexity [24]. Therefore, a novel solution which combines the advantages of convolutions to capture local temporal features and transformers to capture global features using lightweight attention mechanisms is worthwhile exploring.

In this paper, we propose a novel architecture named CUE-Net which amalgamates spatial **C**ropping, with an

[1] https://github.com/damith92/CUENet

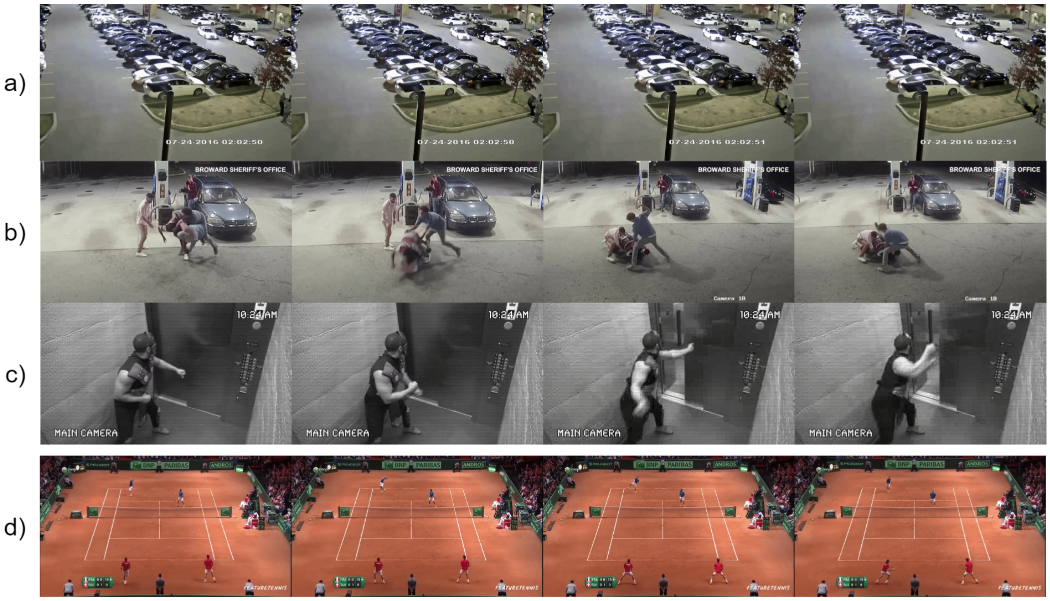

Figure 1. Sample violence detection videos. **(a)** is a set of frames from a challenging video from RWF-2000 where the people involved in the violent incident are far away from the camera, occupying only a small part of the frame. **(b)** shows a typical violent video from the RWF-2000 dataset correctly classified by CUE-Net. **(c)** is a video from the RWF-2000 dataset test split, where a man makes punching actions but is not really engaging in a fight. CUE-Net incorrectly classifies this as a violent video. **(d)** is a video from the RLVS dataset which CUE-Net correctly classifies as non-violent, but for which the ground-truth is mislabeled as violent.

enhanced version of the **U**niformerV2[16] architecture which incorporates the benefits of both the convolution and self-attention. In this architecture, we propose Modified **E**fficient Additive Attention (MEAA), a novel efficient attention mechanism which reduces the quadratic time complexity of self-attention to capture the important global spatio-temporal features, to mitigate the above mentioned bottlenecks. For the best of our knowledge, this is the first time that such a model which incorporates convolution and self-attention along with modified Efficient Additive Attention mechanism has been investigated in the context of violence detection in videos. Our contributions are as follows:

1. We propose CUE-Net, a novel architecture for violence detection video analytics which incorporates a novel enhanced version of the UniformerV2 architecture along with Modified Efficient Additive Attention (MEAA), a novel attention mechanism to capture the important global spatio-temporal features.
2. We incorporate a spatial cropping mechanism based on the detected number of people in our algorithm before the video is fed into the main learning algorithm, to focus the method on the area where violence is occurring without losing the important surrounding information.
3. Our results set a new state-of-the-art on the RWF-2000 and RLVS datasets, outperforming the most recently published methods.

2. Related Work

This section summarizes the current state-of-the-art methods for VD and categorizes different methods used in the context of violence detection as an action recognition task vs an anomaly detection task.

2.1. Deep Learning Architectures for Violence Detection using Anomaly Detection

In anomaly detection scenarios, violent events are considered as scarce abnormal events deviating from normal day-to-day events. Algorithms learn to characterise the features of normal events, and violence detection is based on detecting events that do not lie in the normal distribution. However, in practice, the boundary between normal and anomalous behaviors can be ambiguous. Under realistic situations, similar behaviors may be normal or anomalous given different conditions, for example the the action of punching will be normal for a friendly fist bump but anomalous for a violent punch [21, 31]. The work of [27] proposes to learn anomalies through a deep Multiple Instance Learning (MIL) framework that treats a video as a bag with short segments/clips of each video as instances in a bag. However, [28] argues that the recognition of the anomalous instances is largely biased by the dominant normal (non-violent) instances of the data, especially when the abnormal events are subtle anomalies that exhibit only small differences compared with normal events.

When trying to frame the VD problem in an anomaly detection context, violent (anomalous) events are identified by focusing mainly on learning how a normal situation looks like rather than focusing on the context of the violent behaviour. Often violence is dependent on context as well as the actions happening in the scene. Models trained to detect anomalies in this manner might not adequately understand the context in which certain violent actions are taking place and therefore may not be able to generalize well as their main task is not to learn the context-specific features for violent events [3].

2.2. Deep Learning Architectures for Violence Detection in an Action Recognition Context

Work by [30] introduced one of the earliest uses of 3D-CNNs along with a softmax classifier for violence detection. As a pre-processing step, first, frames where people are present are identified, with the premise that the violent actions will happen only when people are present. Then a 3D-CNN extracted spatio-temporal features out of the filtered frames and a soft-max layer classified the results. In another study, [26] introduced a novel approach for violence detection in the space of action recognition by learning contextual relationships between people using human skeleton points. Unlike the previous references, [26] formulated 3D skeleton point clouds from human skeleton sequences extracted from videos and then performed interaction learning on these 3D skeleton point clouds, considering them as non-Euclidean graphs using Graph CNNs. [26] is one of the first papers to evaluate performance on a real-world surveillance violence detection data set (RWF-2000) [4] where all most all the previous literature was evaluated on non-surveillance based datasets such as the Hockey Fight dataset [2]. [8] introduced a novel deep architecture comprising of two simultaneous pipelines, one to extract the skeletons of people using a pose estimation model and the other to estimate the dynamic temporal changes between frames where the outputs from the two pipelines were fused together using addition to transmit information even when one of the inputs provides a zero-valued signal.

The current state-of-the-art approach for violence detection on the RWF-2000 dataset relies on a Video Swin Transformer [14]. This work applies a method to extract keyframes from the videos based on frame colour, texture and motion features using colour histograms, gray level co-occurrence matrices and optical flow. Then, a Video Swin Transformer [17] starts with processing small patches of the videos and gradually merges them into deeper transformer layers in spatio-temporal context, creating a hierarchical representation. This approach has enabled the aggregation of features from a local to a global context.

In summary, framing the violence detection problem using action recognition has advantages over anomaly detection. Recent literature has focused more on extracting rich and representative features of violent actions and derive a better contextual understanding in order to separate violent actions from normal activities [31].

3. Proposed Method

In this section we first motivate our work and then discuss our proposed CUE-Net method in detail.

Motivation: Our work takes inspiration from the action recognition literature, as it provides an effective supervised method for video action recognition. In the action recognition space, a novel deep architecture called Unified Transformer (UniFormer) [15] has been introduced which seamlessly integrates the merits of 3D convolution and spatio-temporal self-attention in a concise format by implementing modules of both convolution and self-attention together to achieve a balance between computational complexity and accuracy. Later, the Uniformer Version 2 (UniformerV2) architecture [16] modified these modules of the previous Uniformer architecture to implement them simultaneously and fuse at the end of the pipeline to capture the relevant spatio-temporal features. Also, UniformerV2 takes advantage of pre-trained ViT embeddings to initialize segments of the architecture to better make use of pretrained knowledge from large image datasets.

However, self-attention has a quadratic computational complexity with respect to the sequence length, making it challenging to process long sequences of tokens such as in videos [12, 24]. To alleviate this issue, [24] introduced a redesigned attention mechanism named Efficient Additive Attention as seen in Fig. 1 (a). This proposed mechanism replaces the expensive matrix multiplication operations with element-wise multiplications and linear transformations with the use of only the key-value interaction. However, such methods have not yet been investigated for the task of violence detection to the best of our knowledge. This poses an opportunity to modify and enhance the concepts discussed to create an improved, tailor-made solution for the problem of violence detection.

3.1. CUE-Net Architecture

We introduce our novel architecture, the spatial **C**ropping, enhanced **U**niformerV2 with Modified **E**fficient Additive Attention network (**CUE-Net**) for violence detection in videos as shown in Fig. 2. The architecture contains five main components, namely: (a) Spatial Cropping Module; (b) 3D Convolution Backbone; (c) Local UniBlock V2; (d) Global UniBlock V3; and (e) Fusion Block, inspired by the motivational factors discussed in the preceding paragraph.

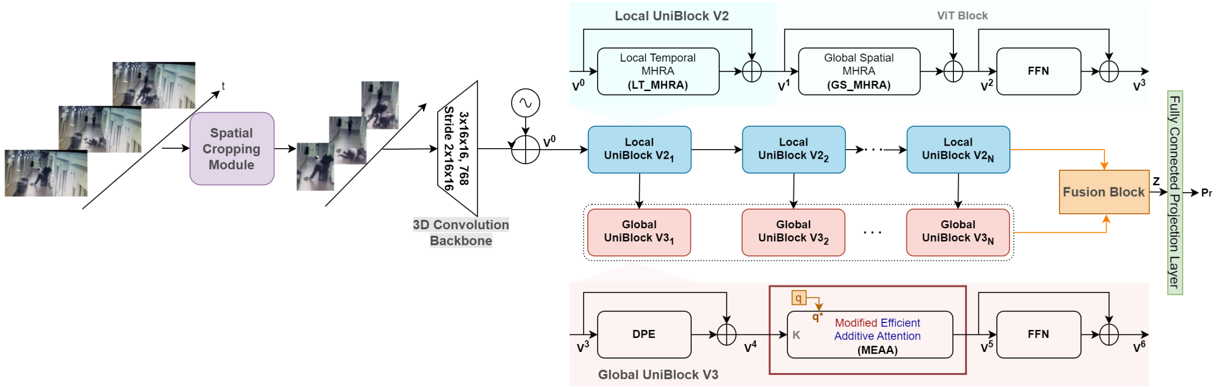

Figure 2. The overall CUE-Net architecture with its main components. **(a)** the Spatial Cropping Module uses the YOLO V8 algorithm to detect people and crop the video spatially; **(b)** the 3D Convolutional Block which is used to encode and downsample the frames spatio-temporally; **(c)** the Local UniBlock V2 which is mainly used to capture the important local dependencies with its main components LT_MHRA, GS_MHRA and a feed forward network (FFN); **(d)** the Global UniBlock V3 which is mainly used to capture the important global spatio-temporal dependencies, with its main components Dynamic Positional Embedding (DPE) unit, MEAA unit which implements a novel efficient self-attention mechanism and a feed forward network (FFN); **(e)** the Fusion Block which is used to fuse the outputs of the Local UniBlock V2 and Global UniBlock V3.

3.1.1 Spatial Cropping Module

The motivation for cropping the video spatially is based on the observation that violence is normally carried out between two or more people. We opted to extract the people and crop the video frames spatially with the maximum bounding box for the area where people are found, so as not to lose the information surrounding the people, but to maximize the important area to focus by removing the parts of the environment where the people are not present. We opted not to perform temporal cropping to avoid any information loss occurring from undetected people. When the video $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times c}$ (T, H, W and c represent temporal dimensions, height, width and colour channels of the video frames respectively) is input to this spatial cropping module, to detect people, we used the YOLO (You Only Look Once) V8 algorithm [11] which classifies objects in a single pass using a CNN-based architecture where a full image is taken as the input. Algorithm 1 elaborates the spatial cropping procedure for the maximum bounding box throughout the video. If more than one person is detected, it outputs $\mathbf{X}' \in \mathbb{R}^{T \times H \times W \times c}$ which is the spatially cropped video. If only a single person or no people are detected, $\mathbf{X}'$ will be the initial video as a whole to make sure the method does not miss out any information.

3.1.2 3D Convolution Backbone

The spatially cropped video frames from the previous module $\mathbf{X}'$ are then passed as input to the 3D Convolution Backbone, where a 3D convolution (i.e., 3×16×16) is used to encode and project the input video as spatio-temporal tokens $\mathbf{V}^0 \in \mathbb{R}^{T \times H \times W \times d}$, ($T$, H, W and d represent temporal dimensions, height and width of the frames and hidden dimensions respectively). Afterwards, according to the original ViT design [7], spatial downsampling by 16× is performed and then a temporal downsampling by 2× is performed to reduce spatio-temporal resolution. The encoded hidden dimension d was maintained the same throughout the architecture modules to facilitate residual connections. At the end of this stage the processed input is sent to the Local UniBlock V2.

3.1.3 Local UniBlock V2

The Local UniBlock V2, has been introduced specifically to model the local dependencies in our CUE-Net architecture. This was extracted from the UniformerV2 architecture without modifications as a result of the ablation study we performed. Here, two types of Multi-Head Relation Aggregator (MHRA) units are used namely, Local Temporal MHRA (LT_MHRA) and Global Spatial (GS_MHRA) along with a Feed Forward Network (FFN) module. The input to this block is $\mathbf{V}^0 \in \mathbb{R}^{T \times H \times W \times d}$ which is the output of the previous 3D Convolution Backbone and this block outputs $\mathbf{V}^3 \in \mathbb{R}^{T \times H \times W \times d}$ at the end of the FFN. The processing inside a Local UniBlock V2 can be represented as:

$$\mathbf{V}^1 = \mathbf{V}^0 + \text{LT\_MHRA}\left(\text{LN}\left(\mathbf{V}^0\right)\right), \tag{1}$$

$$\mathbf{V}^2 = \mathbf{V}^1 + \text{GS\_MHRA}\left(\text{LN}\left(\mathbf{V}^1\right)\right), \tag{2}$$

$$\mathbf{V}^3 = \mathbf{V}^2 + \text{FFN}\left(\text{LN}\left(\mathbf{V}^2\right)\right), \tag{3}$$

where $\text{LN}(\cdot)$ represents layer normalization. A Multi-Head Relation Aggregator (MHRA) unit concatenates multiple

Algorithm 1 Spatial Cropping Mechanism with YOLO V8

```
Input: $x$                                        ▷ Input Video
Output: $x'$                                      ▷ Cropped Video
    ▷ A crop box for a video is defined by the coordinates
    $(x_{min}, y_{min})$ and $(x_{max}, y_{max})$
    $(x_{min}, y_{min}) \leftarrow (inf, inf)$
    $(x_{max}, y_{max}) \leftarrow (0, 0)$
    $max\_people \leftarrow 0$                     ▷ Max no of people

    $F \leftarrow YOLO\_V8(x)$        ▷ $F$ is the list of frames of
    the video where the $i$th entry $f_i$ is another list of people
    bounding boxes $P$ found in each frame, the $j$th bounding
    box $p_j$ denoted as $(x^j_{min}, y^j_{min}), (x^j_{max}, y^j_{max})$.

    for each $f_i$ in $F$ do
        $n_i \leftarrow 0$                   ▷ no. of people in each frame
        for each $p_j$ in $f_i$ do
            $x_{min} \leftarrow \min(x_{min}, x^j_{min})$
            $y_{min} \leftarrow \min(y_{min}, y^j_{min})$
            $x_{max} \leftarrow \min(x_{max}, x^j_{max})$
            $y_{max} \leftarrow \min(y_{max}, y^j_{max})$
            $people \leftarrow people + 1$
        end for
        if $n_i > 0$ then
            $max\_people \leftarrow max(max\_people, n_i)$
        end if
    end for
    if $max\_people > 1$ then
        $x' \leftarrow crop\_video(x_{min}, y_{min}, x_{max}, y_{max})$
    else
        $x' \leftarrow x$
    end if
```

heads and can be described as:

$$\mathrm{S}_n(\mathbf{V^i}) = \mathbf{B}_n \cdot \mathrm{L}_n(\mathbf{V^i}), \tag{4}$$

$$\mathrm{MHRA}(\mathbf{V^i}) = [\mathrm{S}_1(\mathbf{V^i}); \mathrm{S}_2(\mathbf{V^i}); \cdots; \mathrm{S}_N(\mathbf{V^i})] \cdot \mathbf{M}, \tag{5}$$

where the relational aggregator of the n-th head is represented by $\mathrm{S}_n(\cdot)$ where $\mathbf{B}_n$ represents an affinity matrix that characterizes the relationships between tokens and $\mathbf{B}_n$ is changed accordingly in LT_MHRA and in GB_MHRA to achieve their respective goals. A linear projection is represented by $\mathrm{L}_n(\cdot)$. A fusion matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ which is learnable is used to integrate N heads during concatenation of the heads represented by [...] at the end of a general MHRA unit.

LT_MHRA: The Local Temporal MHRA (LT_MHRA) takes the input $\mathbf{V}^0$ from the 3D Convolution Backbone, implements depth-wise convolution (DWConv) with the help of the affinity matrix $\mathbf{B}_n$ described in the preceding paragraph, as the goal of this unit is to reduce the local temporal redundancy and to learn local representations form the local spatio-temporal context. This unit outputs $\mathbf{V}^1 \in \mathbb{R}^{T \times H \times W \times d}$.

GT_MHRA: The Global Temporal MHRA (GT_MHRA) receives the output of the LT_MHRA unit $\mathbf{V}^1$ and implements multi-headed self-attention (MHSA) from the ViT architecture [7] with the help of the affinity matrix $\mathbf{B}_n$ described earlier as the goal of this unit is to make use of the rich image pretraining of ViTs learned from large image databases. To achieve this target, the GT_MHRA units are initialized with image-pretrained ViT embeddings inflated along the temporal dimension and the output of this unit is $\mathbf{V}^2 \in \mathbb{R}^{T \times H \times W \times d}$.

FFN: The Feed Forward Network (FFN) module accepts the output $\mathbf{V}^2$ of GT_MHRA, and consists of two linear projections separated by a GeLU [10] activation function. FFN is implemented at the end of the Local UniBlock V2 to output $\mathbf{V}^3 \in \mathbb{R}^{T \times H \times W \times d}$.

3.1.4 Global UniBlock V3

The Global UniBlock V3 has been introduced specifically to perform global long-range dependency modeling on the spatio-temporal scale in our CUE-Net. This Global UniBlock V3 consists of three basic units namely, Dynamic Positional Embedding (DPE) unit, Modified Efficient Additive Attention (MEAA) unit, and finally a Feed Forward Network (FFN) module. The input to this block is $\mathbf{V}^3 \in \mathbb{R}^{T \times H \times W \times d}$ which is the output of the previous Local UniBlock V2 and the Global UniBlock V3 outputs $\mathbf{V}^6 \in \mathbb{R}^{1 \times d}$ at the end of the FFN unit. The processing inside this block where $\mathrm{LN}(\cdot)$ represents layer normalization can be represented as:

$$\mathbf{V}^4 = \mathbf{V}^3 + \mathrm{DPE}\left(\mathbf{V}^3\right), \tag{6}$$

$$\mathbf{V}^5 = \mathrm{MEAA}\left(\mathrm{LN}\left(\mathbf{q}\right), \mathrm{LN}\left(\mathbf{V}^4\right)\right), \tag{7}$$

$$\mathbf{V}^6 = \mathbf{V}^5 + \mathrm{FFN}\left(\mathrm{Norm}\left(\mathbf{V}^5\right)\right). \tag{8}$$

DPE: The Dynamic Positional Embedding (DPE) unit receives the input $\mathbf{V}^3$ from the previous Local UniBlock V2, and uses simple 3D depth-wise spatio-temporal convolution with zero padding (DWConv) to encode spatio-temporal positional information for token representations, as the videos vary both spatially and temporally. The output of the DPE block is $\mathbf{V}^4 \in \mathbb{R}^{T \times H \times W \times d}$.

Modified Efficient Additive Attention (MEAA): In the Modified Efficient Additive Attention (MEAA) unit, a learnable query $\mathbf{q} \in \mathbb{R}^{1 \times d}$ is converted into a video representation, through modeling a relationship between this query $\mathbf{q}$ and all the spatio-temporal tokens $\mathbf{V^4}$ received from the DPE unit, with the help of this modified version of Efficient Additive Attention. As depicted in Fig. 3 (b), the learnable query vector $\mathbf{q}$ is projected into query $(\mathbf{q}^*)$ and $\mathbf{V}^4$ is projected into the key $(\mathbf{K})$ using two linear layers where

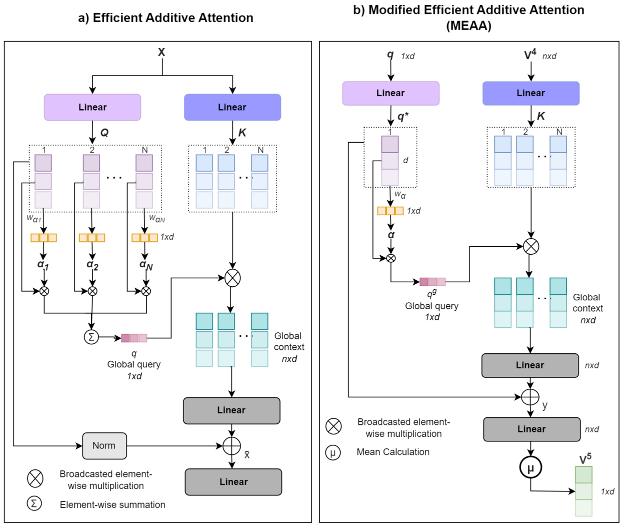

Figure 3. **(a)** illustrates the Efficient Additive Attention where the expensive matrix multiplication operations have been replaced with element-wise multiplications and linear transformations via a query-key pair interaction. **(b)** represents the Modified Efficient Additive Attention (MEAA) which only uses a query vector instead of a whole query matrix when computing Efficient Additive Attention, reducing the computational complexity along with memory usage.

n is the token length and d is the number of hidden dimensions. Afterwards, another vector of learnable parameters $\mathbf{w}_a \in \mathbb{R}^d$ is multiplied with the query $\mathbf{q}^*$ with the intention of learning the attention weights of the query. This results in outputting $\alpha \in \mathbb{R}^1$ which can be considered the global attention query vector:

$$\alpha = \left(\frac{\mathbf{q}^* \cdot \mathbf{w}_a}{\sqrt{d}} \right) \tag{9}$$

The global query vector $\mathbf{q}^g \in \mathbb{R}^{1\times d}$ is afterwards derived using the attention weight which was learned as:

$$\mathbf{q}^g = \alpha \odot \mathbf{q}^*, \tag{10}$$

where $\odot$ represents element-wise multiplication.

Finally, element-wise multiplication is performed between the global query vector $\mathbf{q}^g \in \mathbb{R}^{1\times d}$ and the key matrix $\mathbf{K} \in \mathbb{R}^{n\times d}$ in order to fuse these two entities, where the end result has dimensions $\mathbb{R}^{n\times d}$. The above process is inexpensive, with linear complexity in relation to token length, compared to obtaining self-attention which has a quadratic complexity. A linear layer is then applied to this element-wise multiplication with a residual connection from $\mathbf{q}^*$ along with a final linear layer to produce the output:

$$\mathbf{V}^5 = \mathbf{Mean}\big(\mathbf{W}_2 \cdot ((\mathbf{W}_1 \cdot (\mathbf{q}^g \odot \mathbf{K}) + \mathbf{b}_1) + \mathbf{q}^*) + \mathbf{b}_2\big). \tag{11}$$

To obtain $\mathbf{V}^5 \in \mathbb{R}^{1\times d}$ as the output, the mean is calculated along the n dimension to get an overall representation.

FFN: Similar to the FFN module in the previous Local UniBlock v2, this Feed Forward Network (FFN) accepts the output $\mathbf{V}^5$ of GT_MHRA module and consists of two linear projections separated by a GeLU [10] activation function, at the end of the Global UniBlock V3 to output $\mathbf{V}^6 \in \mathbb{R}^{1\times d}$.

3.1.5 Fusion Block

At the very end of the CUE-Net architecture, a fusion block integrates the final token from the Global UniBlock $\mathbf{V}^6 \in \mathbb{R}^{1\times d}$ with the final video class token $\mathbf{V}^{3'} \in \mathbb{R}^{1\times d}$ extracted from the final output $\mathbf{V}^3 \in \mathbb{R}^{T\times H\times W\times d}$ of the Local UniBlock. These tokens $\mathbf{V}^6$ and $\mathbf{V}^{3'}$ are dynamically fused to obtain $\mathbf{Z}$ as:

$$\beta' = Sigmoid(\beta), \tag{12}$$

$$\mathbf{Z} = (1-\beta') \odot \mathbf{V}^6 + \beta' \odot \mathbf{V}^{3'}. \tag{13}$$

using another learnable parameter $\beta \in \mathbb{R}^{1\times d}$ passed through the Sigmoid function. Finally, the target class Pr is obtained by passing $\mathbf{Z}$ through a fully connected projection layer.

4. Experiments and Results

4.1. Datasets

The most challenging datasets so far in the VD domain are the Real-World Fighting (RWF-2000) dataset [4] and the Real Life Violence Situations (RLVS) dataset [25], that contain video footage of fighting in real life scenarios. But of these two datasets, only the RWF-2000 dataset contains exclusive surveillance footage.

4.1.1 Real World Fighting (RWF-2000) Dataset

The Real World Fighting (RWF-2000) dataset [4] was introduced in 2020 and is the most comprehensive dataset, containing real world fighting scenarios sourced purely through surveillance footage. A typical violent example can be seen at Fig. 1 (b). RWF-2000 contains 2,000 trimmed video clips captured by surveillance cameras from real-world scenes collected from YouTube. Each video is trimmed to 5 seconds where the fighting occurs. The dataset is balanced with 1000 violent videos and 1000 non-violent videos, with a 80%-20% predefined train-test split which has been thoroughly checked for data leakage between the splits.

4.1.2 Real Life Violence Situations (RLVS) Dataset

The Real Life Violence Situations (RLVS) dataset [25] consists of 2000 video clips with 1000 violent and another 1000

Method	Model Type	Accuracy (%)
ConvLSTM[4]	CNN+LSTM	77.00
X3D[14]	3DCNN	84.75
I3D[9]	3DCNN	83.40
Flow Gated Network[4]	Two Stream Graph CNN	87.25
SPIL[26]	Graph CNN	89.30
Structured Keypoint Pooling[9]	CNN	93.40
Video Swin Transformer[17]	ViViT	91.25
ACTION-VST[14]	CNN + ViViT	93.59
CUE-Net (Ours)	**Enhanced UniformerV2 + MEAA**	**94.00**

Table 1. Results comparison for the RWF-2000 Dataset.

non-violent videos collected from YouTube. These contain many real street fight situations in several environments and conditions with an average length of 5s from different sources such as surveillance cameras, movies, video recordings, etc. Similar to RWF-2000, a 80%-20% train-test split has been created for this dataset.

4.2. Implementation Details

Our algorithm was implemented in PyTorch using the AdamW optimizer [20] with a cosine learning rate schedule [19] starting with a learning rate of 1e-5 and Cross-Entropy Loss, taking insights from training recipes of the original UniformerV2 architecture [16]. To initialize the Global MHRA units of the Local UniBlocks, pretrained embeddings from CLIP-ViT [23] model are used as [16] states this yields the best results in their architecture due to the well learned representations by vision-language contrastive learning. All models were trained for 50 epochs where the best validation model was saved after each epoch. We utilized NVidia A100 GPUs with 40GB/80GB memory. For data augmentation, RandAugment by [5] was used. Our best performing CUE-Net architecture consisted of 354M parameters where the number of frames selected (T) to be inputted was 64 with a resized frame height (H) and width (W) of 336 × 336 in RGB channels ($c = 3$).

Method	Model Type	Accuracy (%)
CNN-LSTM[25]	VGG16+LSTM	88.20
Temporal Fusion CNN +LSTM[6]	CNN+LSTM	91.02
DeVTr[1]	ViViT	96.25
ACTION-VST[14]	CNN + ViViT	98.69
CUE-Net (Ours)	**Enhanced UniformerV2 + MEAA**	**99.50**

Table 2. Results comparison for the RLVS Dataset.

4.3. Results

In this section we perform an in-depth analysis comparing our CUE-Net architecture with other leading architectures using the two different datasets, RWF-2000 and RLVS. Following the practice of other researchers [9, 14], we also use classification accuracy as the metric to evaluate the performance as both of the trained and tested upon datasets are balanced. Tab. 1 and Tab. 2 present the results comparison of our CUE-Net architecture with other state-of-the-art methods on RWF-2000 and RLVS datasets respectively. Our CUE-Net architecture outperforms all others in classification accuracy. On the RWF-2000 dataset, our CUE-Net architecture reaches an accuracy of 94.00%, and on the RLVS dataset, it records an accuracy of 99.50%, setting a new state-of-the-art on both datasets.

4.3.1 Visual Analysis of Results

RWF-2000 Dataset: For the RWF-2000 test set, we performed a visual evaluation of the misclassified instances. As the accuracy was 94.00%, there were only 24 misclassified instances where 15 non-violent videos were misclassified as violent and 8 violent videos were misclassified as non-violent. This gives the idea that our model is better able to learn the specifics of the violent action markers. Supporting this proposition, we were able to identify a video shown in Fig. 1 (c) where a man makes punching actions but is not really engaging in a fight. Our method misclassifies this non-violent video as a violent video.

RLVS Dataset: We also performed a visual evaluation of the misclassified instances in RLVS test set. Since our accuracy was 99.5%, there were only 2 misclassified videos where 1 non-violent video was misclassified and vice versa. When analysing the 2 misclassified videos, we noted the

Spatial Cropping	Local UniBlock	Global UniBlock	Accuracy (%)	FLOPs (Giga)
×	Self-Attention	Self-Attention	92.00	6108
✓	Self-Attention	Self-Attention	92.50	6108
✓	MEAA	Self-Attention	50.00	5929
✓	MEAA	MEAA	50.00	**5749**
✓	**Self-Attention**	**MEAA**	**94.00**	5826

Table 3. Ablation showing the effect of Spatial Cropping, Self-Attention and MEAA in the Local UniBlock and Global UniBlock.

video shown in Fig. 1 (d) was labelled violent and was misclassified, but was a *mislabeled instance of a non-violent video* where two players were playing tennis without any violence, thus increasing the true accuracy of our model with this correction to 99.75%. This strongly shows CUE-Net has learned the dynamics of violent actions.

4.4. Ablation Study

We performed a series of ablation studies to asses the efficacy of the components of CUE-Net.

4.4.1 Ablation on Spatial Cropping, Self-Attention and MEAA in Local UniBlock and Global UniBlock

Four ablation experiments were conducted to explore the use of spatial cropping and the MEAA module as shown in in Tab. 3. First, we remove the spatial cropping module, and use Self-Attention both in the Local UniBlock and in the Global UniBlock. In the second row of the table, we add spatial cropping, which enhances the performance of the model. In the last row, we replace the Self-Attention with Modified Efficient Additive Attention (MEAA) in the Global UniBlock, forming our full CUE-Net model. This provides a considerable boost of 1.5% in accuracy. We speculate traditional Self-Attention may have an information overload especially while trying to capture representative features temporally. In contrast, with the simpler MEAA, it may be easier for the Global UniBlock to learn the discriminative features temporally when it comes to identifying violent actions. The remaining rows in the table explore the use of MEAA in the Local UniBlock. Here the algorithm performance becomes random as shown by the results in Tab. 3. In this setting, the local UniBlocks are not initialized with pretrained ViT embeddings and underperform. Also, it is evident from Tab. 3 that the FLOPs count reduces when MEAA is used in place of Self Attention depicting a reduction in computational complexity. Therefore our proposed approach of using Self-Attention in the Local UniBlock and MEAA in the Global UniBlock has the best performance along with a reduced FLOPs count.

Efficient Additive Attention Variant	Accuracy (%)	GPU Memory Usage
Original	93.00	47.33 GB
MEAA	**94.00**	**35.04 GB**

Table 4. Ablation showing the Original Efficient Additive Attention vs Modified Efficient Additive Attention (MEAA).

4.4.2 Ablation on Original Efficient Additive Attention vs Modified Efficient Additive Attention (MEAA)

We also experimented with the original Efficient Additive Attention with a n-dimensional query matrix instead of a 1-dimensional query vector in the Global UniBlock in our CUE-Net architecture, but it under-performed, with 1% less accuracy compared to MEAA as seen in Tab. 4. We also note the GPU memory consumption was considerably higher (47 GB compared to 35 GB) when the original Efficient Additive Attention was used. Therefore, we can state that our MEAA gives a competitive edge over original Efficient Additive Attention when it comes to memory usage.

5. Conclusion

This paper introduces CUE-Net, a novel framework for violence detection in videos which implements cropping with an enhanced version of UniformerV2 architecture. CUE-Net uses convolution-based mechanisms to capture the local features and attention mechanisms to capture the global spatio-temporal features fused with a novel attention mechanism named Modified Efficient Additive Attention. We incorporated video cropping spatially, based on the detected number of people before the video is fed into the main processing algorithm to focus the method on the areas where violence takes place. We also proposed Modified Efficient Additive Attention instead of Self-Attention in the Global UniBlock V3 of the CUE-Net architecture, to capture the important global spatio-temporal features, as it has shown to be effective and efficient. Our proposed CUE-Net algorithm has achieved new state-of-the-art performance on the RWF-2000 and RLVS datasets, surpassing the results of most recently published methods.

Acknowledgment. This work has been funded through an EPSRC DTP studentship at Queen Mary University of London in collaboration with Remark AI UK Ltd. This paper utilised Queen Mary's Andrena HPC facility, supported by QMUL Research-IT Services [13].

References

[1] Almamon Rasool Abdali. Data efficient video transformer for violence detection. In *2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pages 195–199. IEEE, 2021. 7

[2] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*, pages 332–339. Springer, 2011. 3

[3] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 3

[4] Ming Cheng, Kunjing Cai, and Ming Li. Rwf-2000: an open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190. IEEE, 2021. 1, 3, 6, 7

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 7

[6] Jean Phelipe de Oliveira Lima and Carlos Maurício Seródio Figueiredo. A temporal fusion approach for video classification with convolutional and lstm neural networks applied to violence detection. *Inteligencia Artificial*, 24(67):40–50, 2021. 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4, 5

[8] Guillermo Garcia-Cobo and Juan C SanMiguel. Human skeletons and change detection for efficient violence detection in surveillance videos. *Computer Vision and Image Understanding*, 233:103739, 2023. 3

[9] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified keypoint-based action recognition framework via structured keypoint pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22962–22971, 2023. 7

[10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5, 6

[11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO V8, 2023. https://github.com/ultralytics/ultralytics [Accessed: 2024-01-16]. 4

[12] Abbas Khan, Muhammad Asad, Martin Benning, Caroline Roney, and Gregory Slabaugh. Crop and couple: cardiac image segmentation using interlinked specialist networks. In *International Symposium on Biomedical Imaging*, 2024. 3

[13] Thomas King, Simon Butcher, and Lukasz Zalewski. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*, 2017. 8

[14] Chenghao Li, Xinyan Yang, and Gang Liang. Keyframe-guided video swin transformer with multi-path excitation for violence detection. *The Computer Journal*, page bxad103, 2023. 1, 3, 7

[15] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2022. 3

[16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1632–1643, 2023. 2, 3, 7

[17] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3, 7

[18] Macrotrends LLC. World crime rate statistics, 2023. https://www.macrotrends.net/countries/WLD/world/crime-rate-statistics [Accessed: 2023-12-16]. 1

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 7

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7

[21] Nadia Mumtaz, Naveed Ejaz, Shabana Habib, Syed Muhammad Mohsin, Prayag Tiwari, Shahab S Band, and Neeraj Kumar. An overview of violence detection techniques: current challenges and future directions. *Artificial Intelligence Review*, 56(5):4641–4666, 2023. 1, 2

[22] Batyrkhan Omarov, Sergazi Narynov, Zhandos Zhumanov, Aidana Gumar, and Mariyam Khassanova. State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. *PeerJ Computer Science*, 8: e920, 2022. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[24] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17425–17436, 2023. 1, 3

[25] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85. IEEE, 2019. 6, 7

[26] Yukun Su, Guosheng Lin, Jinhui Zhu, and Qingyao Wu. Human interaction learning on 3d skeleton point clouds for video violence recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 74–90. Springer, 2020. 3, 7

[27] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 2

[28] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021. 2

[29] Abdarahmane Traoré and Moulay A Akhloufi. Violence detection in videos using deep recurrent and convolutional neural networks. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 154–159. IEEE, 2020. 1

[30] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11):2472, 2019. 1, 3

[31] Fath U Min Ullah, Mohammad S Obaidat, Amin Ullah, Khan Muhammad, Mohammad Hijji, and Sung Wook Baik. A comprehensive review on vision-based violence detection in surveillance videos. *ACM Computing Surveys*, 55(10):1–44, 2023. 1, 2, 3