CrimeNet: Neural Structured Learning using Vision Transformer for violence detection

Neural Networks 161 (2023) 318–329

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

CrimeNet: Neural Structured Learning using Vision Transformer for violence detection

Fernando J. Rendón-Segador [a,*], Juan A. Álvarez-García [a], Jose L. Salazar-González [a], Tatiana Tommasi [b]

[a] Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Spain
[b] Politecnico di Torino & Italian Institute of Technology, Italy

ARTICLE INFO

Article history:
Received 16 June 2022
Received in revised form 20 December 2022
Accepted 30 January 2023
Available online 2 February 2023

Keywords:
Deep learning
Neural Structured Learning
Vision Transformer
Violence detection
Adversarial Learning

ABSTRACT

The state of the art in violence detection in videos has improved in recent years thanks to deep learning models, but it is still below 90% of average precision in the most complex datasets, which may pose a problem of frequent false alarms in video surveillance environments and may cause security guards to disable the artificial intelligence system.

In this study, we propose a new neural network based on Vision Transformer (ViT) and Neural Structured Learning (NSL) with adversarial training. This network, called CrimeNet, outperforms previous works by a large margin and reduces practically to zero the false positives. Our tests on the four most challenging violence-related datasets (binary and multi-class) show the effectiveness of CrimeNet, improving the state of the art from 9.4 to 22.17 percentage points in ROC AUC depending on the dataset. In addition, we present a generalisation study on our model by training and testing it on different datasets. The obtained results show that CrimeNet improves over competing methods with a gain of between 12.39 and 25.22 percentage points, showing remarkable robustness.

1. Introduction

Violence detection is a very important functionality in public or private security. In smart cities, more and more surveillance cameras are being installed that enable several use cases such as traffic management, infraction or weapon detection (Salazar González, Zaccaro, Álvarez-García, Soria Morillo, & Sancho Caparrini, 2020). In addition, in schools, hospitals, shopping centres, and other buildings they are also generally used as a dissuasion and in the worst case to identify criminals once a crime has been committed. This growing number of cameras requires sufficient human resources to control the volume of video they generate, however the number of cameras needing attention is greater than the number of staff. Furthermore, after 20 min of monitoring a CCTV system, operators' attention spans are considerably reduced (Ainsworth, 2002; Velastin, Boghossian, & Vicencio-Silva, 2006). The current difficulties in tackling this problem become even more evident when considering that security is an issue of international concern which scales up from single cities to the World Wide Web. The amount of violent audiovisual content circulating on the network is excessive. To such an extent that the operators in charge of controlling and filtering videos of this type on social networks end up with mental health problems due to over-exposure to violent content.[1] Despite this, video surveillance systems are still being manned by humans because the number of false positives is not acceptable in production environments and a human in the loop is still necessary. When analysing the literature, we found that for the most challenging datasets the state of the art does not reach 90% accuracy (Lv et al., 2021). To overcome this issue, we aim at designing a robust and sufficiently accurate model to minimise the number of false positives, making it possible to use a video violence detector in real environments.

This work covers the video detection of all types of violent events, both visually intentional actions such as a fight between people, as well as unintentional acts such as an explosion. Moreover, we go beyond standard anomaly detection which differentiates violent from non-violent events: we target a model able to differentiate between different types of violence. This is a complex problem since many types of violent actions can be similar even if their classes are different (there are categories such as abuse, arrest or assault that can be confused with each other) or they can even occur in the same video. To tackle these

* Corresponding author.
E-mail addresses: frendon@us.es (F.J. Rendón-Segador), jaalvarez@us.es (J.A. Álvarez-García), jsalazar@us.es (J.L. Salazar-González), tatiana.tommasi@polito.it (T. Tommasi).

[1] https://www.bloomberg.com/news/articles/2021-12-24/tiktok-sued-by-content-moderator-traumatized-by-graphic-videos

https://doi.org/10.1016/j.neunet.2023.01.048

challenges we propose to combine the powerful self-attention learning paradigm with Neural Structured Learning (NSL). The former is obtained by exploiting the most recent Vision Transformer deep architecture (ViT, Dosovitskiy et al., 2021). The latter leverages the relation among neighbouring samples during training. Specifically, it provides a regularisation effect by biasing the network to learn similar hidden representations for close instances.

Our key contributions can be summarised as follows:

- We introduce CrimeNet: a deep model that combines adversarial NSL with ViT for violent activity recognition in videos. Up to our knowledge this is the first time that NSL is used with Transformers rather than with standard convolutional neural networks, and also the first time that it is applied for video recognition.
- CrimeNet improves the state of the art for violence detection on four datasets by reaching an accuracy over 99.98%, with an advantage ranging from 9.4 to 22.17 percentage points in ROC AUC over its competitors. A detailed ablation shows that NSL provides and improvement of 9.55% points in ROC AUC over the use of only ViT as the architecture of the model.
- We present an extensive cross-dataset analysis and show the generalisation abilities of CrimeNet. Despite the challenging setting of training and testing on different datasets, CrimeNet advances the state of the art from 12.39 to 25.22 percentage points with respect to existing methods that are affected by the domain shift.

This paper is organised as follows: Section 2 provides a brief survey of the state of the art of the problem. Section 3 describes in detail each of the datasets used in this work. Section 4 provides in-depth details on the proposed model architecture. Section 5 summarises the types of experiments proposed in this work. Then, in Sections 6 and 7 the results obtained from the experimentation are shown and described. Finally Section 8 summarises the relevance of the results obtained and the possible lines of progress for future development.

2. Related works

In this section, a review of the state of the art on video detection of violence is carried out. Furthermore, several Neural Graph Learning proposals are shown, where NSL is a specific case, applied to computer vision problems.

2.1. Deep learning for violence behaviour detection in videos

Crowd behaviours analysis (Li, Chen, Nie & Wang, 2017a, 2017b; Sadeghian, Alahi, & Savarese, 2017) and detection of violent actions in videos are well-known and well-studied research fields (Bermejo Nievas, Deniz Suarez, Bueno García, & Sukthankar, 2011; Deniz, Serrano, Bueno, & Kim, 2014), however, since 2014 when the first paper using a deep learning approach (Ding, Fan, Zhu, Feng, & Jia, 2014) appeared, violence detection has advanced by leaps and bounds. Specifically, for some datasets (Bermejo Nievas et al., 2011; Hassner, Itcher, & Kliper-Gross, 2012), based on short clips and labels for the whole clip, there exist approaches reaching up to 100% accuracy (Rendón-Segador, Álvarez-García, Enríquez, & Deniz, 2021).

Given that these datasets were not sufficiently challenging, researchers created new testbeds with hours of CCTV camera recordings: the videos are longer and often annotated at frame-level. In addition, some of them have moved from 2 classes (violence and non-violence) to multiple classes distinguishing the typology of violence in the video.

The first paper to introduce this new family of datasets was (Sultani, Chen, & Shah, 2018) in 2018: the UCF Crime collection incorporates the greatest variety of violence classes (14 types). This work also presented a model called DMIL Ranking, in which anomaly detection is approached as a regression problem considering video segments as instances in "multiple instances learning" (MIL) and using a ranking-based loss function to evaluate a fully connected neural network.

The NTU CCTV Fights dataset (Perez, Kot, & Rocha, 2019) was presented in 2019: it contains only long violent videos, labelled at frame-level with violent or non-violent classes. This dataset has been evaluated using different known feature extractors such as Two-stream Convolutional Neural Network (CNN) or 3D CNN and different classifiers such as End-to-End CNN, Long–short Time Memory (LSTM) and Support Vector Machine (SVM).

Another method combining the spatio-temporal feature extractor of ResNet 3D (Dubey, Boragule & Jeon, 2019) and the loss function of Sultani et al. (2018) was also published that year, improving the results for the UCF Crime dataset. Zhong et al. (2019) evaluated the same dataset using a convolutional graph to clean and refine the classifier based on Temporal Segment Networks (Wang et al., 2018).

New work focused on anomaly detection emerged in 2020. Degardin (2020) provided a new dataset, UBI Fights, and proposed an architecture based on the Gaussian mixture model (GMM) for the detection of abnormal events in videos applied under the weakly supervised learning paradigm.

Also noteworthy is the work of Kamoona, Gosta, Bab-Hadiashar, and Hoseinnezhad (2020) in the weakly supervised setting, using an encoder–decoder architecture (DMIL AutoEncoder). Furthermore in the same year, Wu et al. (2020) released the new dataset XD Violence, with 7 classes, the largest number of videos (4754), and hours (217), also incorporating sound. The authors presented a method that exploits visual and audio feature extractors whose output is mixed and provided to graph neural networks to detect short- and long-range temporal relationships.

In 2021 Tian et al. (2021) proposed an approach for anomaly detection in videos where a learning function recognises positive instances on the basis of the feature magnitude learning function and using self-attention mechanisms (Vaswani et al., 2017). In the same year, Lv et al. presented a study in which they propose a weakly supervised anomaly localization (WSLA) method that measures variations in both spatial and temporal contexts. Their results mark the current state of the art for the UCF Crime dataset.

In the work by Chang, Li, Shen, Feng, and Zhou (2021), frame anomalies are detected starting from a single binary annotation at the video level. Based on the extracted visual features, attention mechanisms are used to refine the classification of anomalous instances.

Dubey, Boragule, Gwak and Jeon (2021) proposed a model that addresses context-dependency by analysing motion and appearance features. They use a 3D ResNet network to extract spatio-temporal and motion feature sets which are then fused and provided as input to a network that learns context-dependencies in a weakly supervised manner using multiple classification measures (MRM).

Another noteworthy study is that of Feng, Hong, and Zheng (2021) in which they developed a multi-instance self-training framework (MIST) to efficiently refine task-specific discriminative representations with only video-level annotations. The model is composed of a multi-instance pseudo-label generator and a self-guided attention function encoder that aims to automatically focus on anomalous regions in frames while extracting task-specific representations.

Finally, Degardin and Proença (2021), presented an iterative learning framework composed of two expert systems working in

the weakly supervised and self-supervised paradigms. This work combines 3D convolutional neural networks for both paradigms with a Bayesian network that is responsible for performing data augmentation.

To provide some reference background, the next section reviews state of the art of Neural Graph Learning, and general cases for NSL and ViT.

2.2. Applied neural graph learning

Graph-based neural learning has been applied to several tasks related to human action recognition.

In 2019 Shi, Zhang, Cheng, and Lu (2019) proposed to represent skeleton data in a directed acyclic graph based on the kinematic dependence between joints and bones. In 2021, Xu and Takano (2021) presented a convolutional graph neural network to estimate 3-D human pose in which the input data were structured using an hourglass graph. In the same year, a model to classify sports videos was presented in Gao, Cai, and Liu (2021): it consists of a convolutional model designed with attention mechanisms for assigning weights to neighbouring nodes, combined with a third-order hourglass graph used to structure the features of the videos.

Graph-based learning has been also used in Yin, Shen, Gao, Crandall, and Yang (2021) to detect 3D objects in videos. In this case, the data were encoded through a grid message passing network (GMPNet). Considering each grid as a node, the data were structured using a k-NN network and the model was a spatio-temporal Transformer-GRU.

There are other fields of research that have experimented with NSL and adversarial learning with significant results, such as Ren, Wang, Zhang, and Chang (2020) for fake news detection through social networks. It has also been applied to protect against adversarial attacks using perturbed data (Jin et al., 2020), achieving significantly better performance compared to state of the art defense methods.

2.3. Applied vision transformer

ViT (Dosovitskiy et al., 2021) profits the Transformer (Vaswani et al., 2017) potential, avoiding inductive bias such as translation invariance and locally restricted receptive field in images. To do it, it splits an image in a sequence of patches, flattens them, produces linear embeddings, adds positional embedding to know where is located each patch in the original image, and feeds this sequence as an input to a standard transformer encoder (composed by a multi-head attention layer (Vaswani et al., 2017) that allows the model to jointly attend to information from different representation subspaces at different positions) as it can be seen in Fig. 4. The success of Transformers, ViT, and their variations (Khan et al., 2022) is beyond doubt, and have improved the state of the art in many areas such as frame synthesis (Liu et al., 2020), action recognition (Girdhar, Carreira, Doersch, & Zisserman, 2019), or object detection in videos (Chen, Cao, Hu, & Wang, 2020).

Our approach differs from previous proposals although being inspired by the use of graph neural networks already leveraged in Wu et al. (2020) and Zhong et al. (2019). As we will describe in the following, we propose a new approach based on supervised NSL (Bui, Ravi, & Ramavajjala, 2018; Gopalan et al., 2021) and ViT (Dosovitskiy et al., 2021). To our knowledge, the NSL paradigm is used here for the first time for violence recognition in videos.

Table 1
Information on datasets used.

Dataset	№ Items	№ Classes	№ Hours
NTU CCTV Fights (Perez et al., 2019)	1000	2	17.68
UBI Fights (Degardin, 2020)	1000	2	80
XD Violence (Wu et al., 2020)	4754	7	217
UCF Crime (Sultani et al., 2018)	1900	14	128

3. Datasets

For our work, we focus on four video datasets recording violent events. They all contain 1000 videos or more, each one ranging from hundreds to thousands of frames. A summary of the datasets' information is in Table 1, while the following list provides further details:

- NTU CCTV Fights (Perez et al., 2019) is a dataset containing 1000 videos depicting real-world fights. Of this dataset, 280 videos are recorded from CCTV and 720 from other sources such as mobile cameras or drones, containing different types of fights, ranging from 5 s to 12 min, with an average duration of 2 min.
- UBI Fights (Degardin, 2020) is a large-scale dataset of 80 h of video fully labelled at frame-level. It consists of 1000 videos, where 216 videos contain a fight event and 784 are normal everyday situations. All unnecessary video segments (e.g., video introductions, news, etc.) that could disrupt the learning process were removed. The title of the videos contains indicators related to the type of the respective video. The dataset is divided into binary classes: violence and normal.
- XD Violence (Wu et al., 2020) is a large-scale dataset with a total duration of 217 h, containing 4754 untrimmed videos with audio signals and video-level tagging. The dataset is divided into the following seven anomalous categories: abuse, car accident, explosion, fight, shooting, and riot.
- UCF Crime (Sultani et al., 2018) is a dataset consisting of long untrimmed surveillance videos covering 14 real-world violent classes, including abuse, arrest, arson, assault, traffic accident, burglary, explosion, fight, robbery, burglary, shooting, theft, shoplifting, and vandalism.

4. Model achitecture

This section shows the type of model and architecture used to address the problem. An in-depth definition of the model and the adaptations applied for our use case is provided.

4.1. Pre-processing

4.1.1. Optical flow

Since our model analyses the video frame by frame, including temporal information in each frame is critical. We use optical flow for this purpose (Farnebäck, 2003): it takes two adjacent frames and represents in an image the amount of pixel variation caused by the observed movements. Of course, the parts of a frame that move together will correspond to pixels with the same intensity as in the example of Fig. 1.

The state of the art demonstrates that the use of optical flow as input in violence detection typically improves the use of RGB input (Mahmoodi & Salajeghe, 2019; Rendón-Segador et al., 2021; Zhou, Ding, Luo, & Hou, 2018).

Fig. 1. Sequence of frames in RGB format and their corresponding optical flow. In this frame sequence, we go from a normal event to a violent event (explosion). Images from the XD-Violence dataset (Wu et al., 2020).

4.1.2. Adversarial neighbours

As it will be seen when describing the model's details (Section 4.3), NSL is used, a new learning paradigm to train neural networks by using structured signals in a graph. This assumes that the model receives two inputs: the RGB frames, in our case transformed by an optical flow algorithm, and a similarity graph. This graph, or structured signal, is used to represent relationships between samples. The similarity graph regularises the training of a neural network, forcing the model to learn accurate predictions by minimising the supervised loss function while maintaining the input structural similarity by reducing the loss function of the neighbouring node.

The similarity graph is generated from the training examples using a graph builder. Each entry is assumed to have an ID and an embedded vector as features. On the one hand, the ID uniquely records and identifies each instance; on the other hand, the embedded vector is assumed to capture the essence of each example by representing it as a list of floating-point values. The graph generator compares the embedded vectors of all input pairs. The degree of similarity between any two samples is calculated as the cosine similarity of their embedded vectors. The brute-force approach to constructing a similarity graph from instance embeddings is $O(n^2)$, which does not scale well to large training sets.

To mitigate that problem, the graph builder uses a well-known randomisation technique called locality-sensitive hashing or LSH (Charikar, 2002). To describe this approach, it is assumed that we have a bidimensional embedding vector represented by n points plotted on a cartesian coordinate plane.

The first step of the LSH process is to choose some random hyperplanes through the origin as it can be seen in Fig. 2. These hyperplanes divide the space and thus the points into discrete sections that are called LSH buckets. Although there are several buckets, the number of points in each bucket is expected to be much smaller than the complete input set. The quadratic nature of comparing all pairs within each bin has a much smaller impact on performance; comparisons within each bin result in a certain number of graph edges within the bin. This generates a series of connected components each of which corresponds to a bucket, these components are separate and are not part of an overall graph. To complete the similarity graph, the bucketing process is repeated several times, with each round of bucketing choosing to select a different number of random hyperplanes until all connected components are connected.

We highlight that the similarity graph is needed only at training time, since during inference it would not be useful to regenerate this graph by adding only the samples to be inferred as the whole LSH process would be necessary again. The inference workflow will not change, simply providing predictions on new testing video frames.

Obtaining the exact similarity relationship between input data is difficult: it is not trivial to evaluate the likeness of one instance to another without a predefined feature embedding that differentiates the type of violence or separates crimes from normal actions. To overcome this issue we propose to generate synthetic neighbouring instances via Adversarial Learning (Zhang, Lemoine & Mitchell, 2018).

In the case of image classification networks, adversarial examples are nothing more than samples to which the pixels have been modified in order to alter the predicted class. Their appearance is similar to that of the original images, but they induce model confusion causing prediction errors. Generally, the adversarial examples are obtained by exploiting the inverse gradient direction of an optimised model. We adopt this logic on the optical flow images.

We start from the hypothesis that normal and violent events have different pixel clustering and variability in the amount of motion measured using optical flow, with the latter being greater in violent events. Likewise, between different types of violent events, there will be variations in the amount of motion and type of pixel clustering. Fig. 3 shows the scheme of a structured signal in which the optical flow images in Fig. 1 are grouped into two connected components α and β, assuming a binary classifier. The images and small circles are the new samples produced via adversarial learning.

With this procedure, we obtain adversarial instances, one per image with respect to the total training set. The samples in the same connected components are visually similar but were created on purpose to confuse the original model, thus this data makes it more robust and able to generalise better.

4.2. Vision transformer

As basic classification model that we expect to be initially fooled by the adversarial examples, we use a deep transformer network.

Specifically, we consider the Vision Transformer (ViT) model (Dosovitskiy et al., 2021) which has achieved remarkable results compared to CNN, in addition to reducing the computational cost required for training and exhibiting a generally weaker

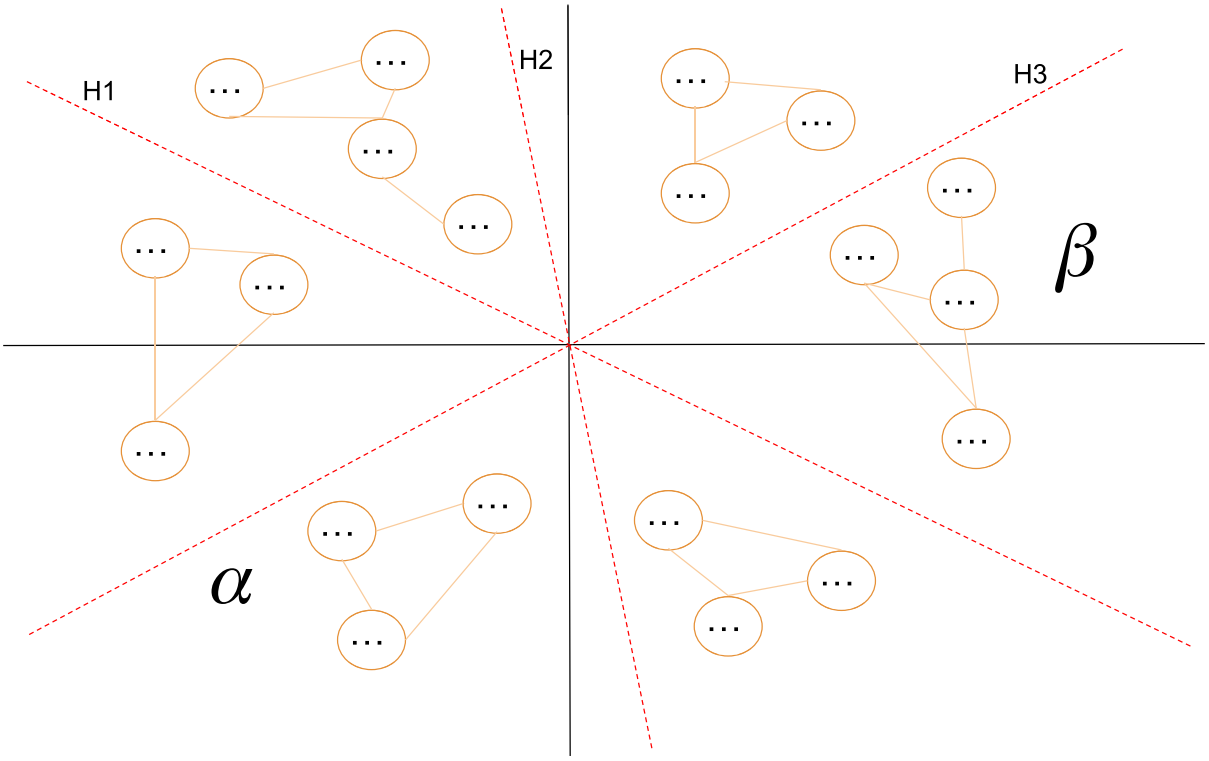


Fig. 2. A representation of the LSH process for the construction of the similarity graph. The dashed red lines corresponding to the labels H1, H2, and H3 are the set of hyperplanes that divide the different points and group them into buckets. α and β correspond to two connected components of the similarity graph. The circles represent instances of the training dataset, in this paper, frames from each training video.

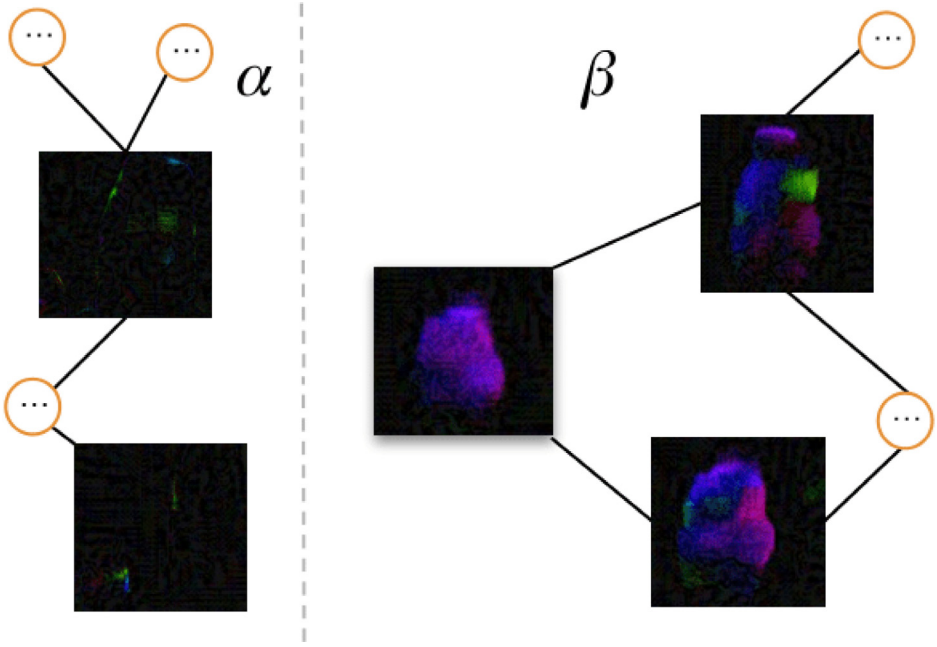


Fig. 3. Similarity graph of optical flow frames generated using adversarial perturbation of the original input (Fig. 1). Component α are images of non-violent events and component β are images of violent events. This similarity graph is used as the second input to the model. The images inside small circles correspond to other adversarial samples.

inductive bias. ViT is a model based on a Transformer architecture (Vaswani et al., 2017) initially used for image classification but later adapted to other visual tasks such as next-frame prediction (Jahanbakht, Xiang, & Azghadi, 2022) or video classification (Arnab et al., 2021).

It divides an image into a sequence of fixed-size fragments called patches, correctly embeds each of them, and includes their positional encoding as input to the Transformer encoder. The fragments are used similarly to the series of embedded words

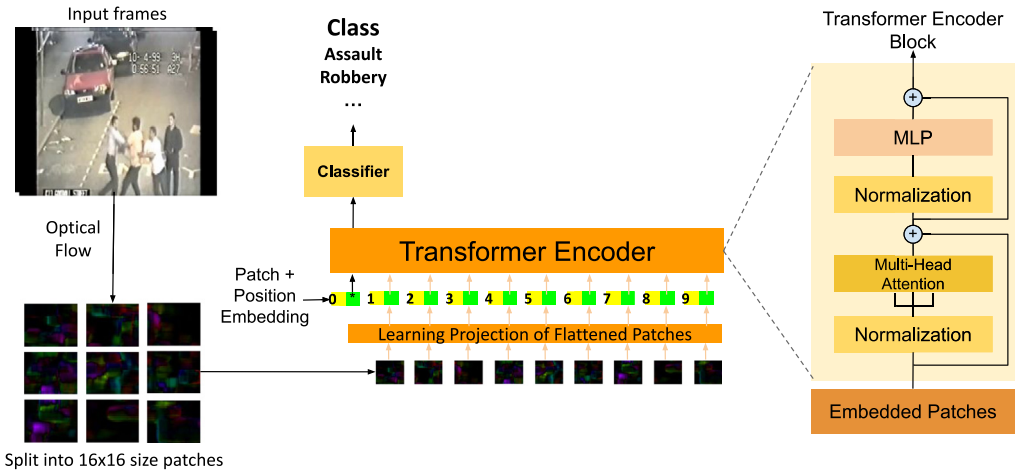


Fig. 4. ViT Architecture for frame classification, based on Dosovitskiy et al. (2021) and Vaswani et al. (2017) works.

for text Transformers, and the output is a prediction label for the image.

Although training a ViT model that has good performance requires from 10 million to more than 100 million images and a huge amount of time and resources (Dosovitskiy et al., 2021), the authors of Steiner et al. (2022) released more than 50000 ViT models trained under diverse settings (including patch size of 8, 16 and 32) on various datasets.[2] The one selected in this work is called ViT-S or DeiT-S (Touvron et al., 2021) and it is trained using a different strategy than the original ViT (Dosovitskiy et al., 2021), achieving good metrics (83.73% in ImageNet (Russakovsky et al., 2015) top-1) with reduced size (115 MB). It is pre-trained with ImageNet 21K and fine-tuned with ImageNet 1K and the patches used are of dimension 16 × 16 pixels. Given that the input frames are resized to 224 × 224, that means the ViT model ingests 14^2 patches. It is worth considering that smaller patch sizes are computationally more expensive. For our work we keep this standard decomposition cardinality of the video frames in patches since we are mainly interested in how ViT combines with NSL: we are aware that the patch size influences the performance of the ViT models (Dosovitskiy et al., 2021; Steiner et al., 2022; Touvron et al., 2021), but this is orthogonal to our analysis and we expect that any improved choice of the patch size would inevitably further improve the observed results.

ViT architecture (Dosovitskiy et al., 2021; Paul & Chen, 2022), is shown in Fig. 4.

We split the image into fixed-size patches and linearly embed each of them. We add positional embeddings by generating a vector sequence with which a standard Transformer encoder is fed. For classification, an extra classification element is added to the vector sequence. Classification is performed by a multilayer perceptron.

4.3. Neural structured learning

In order to take advantage of the power of the similarity graph generated by means of adversarial instances, we follow the scheme proposed in NSL (Bui et al., 2018; Gopalan et al., 2021) (See Fig. 5). In it we can see that we have a pair of inputs, the frames of the training set transformed by optical flow, and a similarity graph or structured signal with instances generated by adversarial modified versions of the sample frame to which it is applied with small perturbations. The generated adversarial neighbours form a similarity graph. In the next step, the original instances are combined with their neighbours and serve as input to the ViT. An encoding of examples and their neighbours is obtained as output. The final regularised graph is the sum of the discriminate loss and the regularisation loss of the graph (Bui et al., 2018).

Using this connection, neural networks learn to maintain the similarities between the sample and the adversarial neighbours while avoiding the confusion resulting from misclassifications, thus improving the quality and accuracy of the overall neural network.

Being $T = t_1 \dots t_n$ the training dataset an $Y = y_1 \dots y_m$ the set of labels associated with the instances of the training dataset, through the adversarial learning process a set of neighbour nodes n is generated for each instance of the training dataset, see Eq. (1).

$$\forall t_i \in T \rightarrow t_i : \{n_1 \dots n_k\} \tag{1}$$

For each subset of neighbours associated with an instance t_i a subgraph $H_{t_i} = (V', E')$ is generated, where each of the neighbours will be connected to the instance, see Eqs. (2) and (3).

$$V' = \{t_i, n_1 \dots n_k\} \tag{2}$$

$$E' = \{(t_i, n_1) \dots (t_i, n_k)\} \tag{3}$$

The sum of all subgraphs forms the graph $G(V, E)$ which we call the structured signal (Eq. (4)).

$$G(V, E) = \sum_{t_i \in T} H_{t_i} \tag{4}$$

From the structured signal $S = G(V, E)$ one can distinguish vertices containing the instances (vertices) of the pure training set $V_t = \{t_1 \dots t_n\}$ and those of the neighbouring vertices generated by adversarial learning $V_n = \{n_1 \dots n_i\}$ both of which form the input to the ViT model that finish in a multi-layer perceptron (MLP) generating a feature vector. The model returns an embedding $\Phi(V_t)$ from V_t and an embedding $\Phi(V_n)$ from V_n, the latter only during training. The feature vector $\Phi(F_v)$ is the result of the graph regularisation between the vectors $\Phi(V_t)$ and $\Phi(V_n)$, see Eq. (5).

$$\Phi(F_v) = \Phi(V_t) + \Phi(V_n) \tag{5}$$

It should be noted that the ViT model takes each of the vertices of the structured signal S and fragments the image it contains into a sequence of patches $P(v) = \{pv_1 \dots pv_n\}$ that embeds a

[2] https://github.com/google-research/vision_transformer

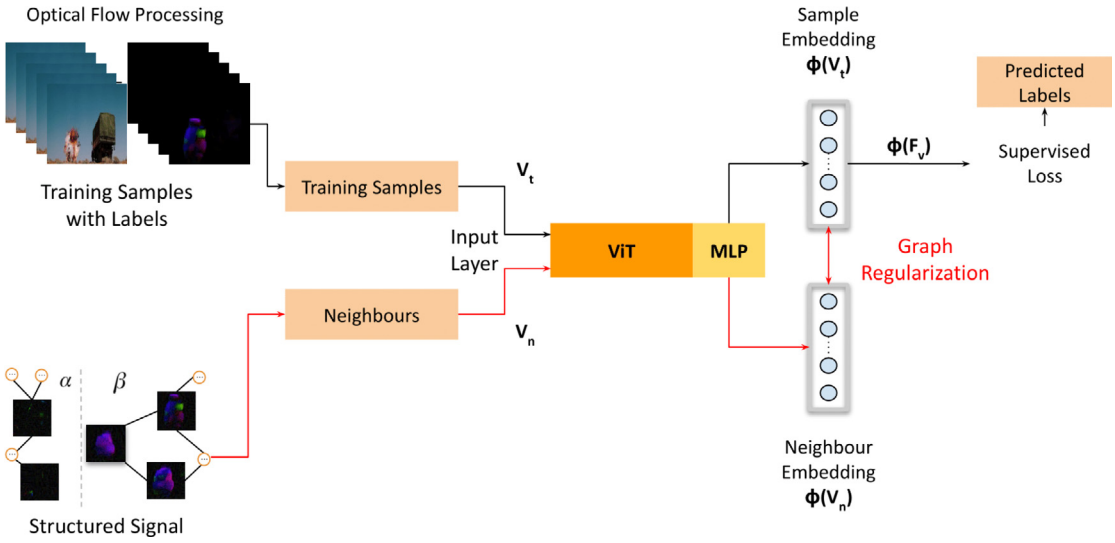


Fig. 5. Neural Structured Learning using as processing model ViT (Fig. 4). The neural network minimises two loss functions, the supervised and the adversarial loss function, which is shown in Eq. (9). Figure inspired in Juan et al. (2020) and Gopalan et al. (2021) works.

vector Γ and then by the process of structured learning is re-embedded under the vector $\Phi(F_v)$, the actual embedding process is represented mathematically by the Eqs. (6), (7) and (8).

$$V_t = \bigcup_{\forall v_t \in V} P(v_t) \tag{6}$$

$$V_n = \bigcup_{\forall v_n \in V} P(v_n) \tag{7}$$

$$\Phi(F_v) = \Phi(\Gamma(V_t)) + \Phi(\Gamma(V_n)) \tag{8}$$

Once the vectors are embedded, the final step consists of regularising the vector $\Phi(F_v)$ and applying the adversarial loss (sparse categorical cross-entropy) function (Goodfellow, Shlens, & Szegedy, 2015). Being $y_i \in Y$ the actual label value of a vertex i and $g_\theta(n_j)$ the prediction of the neighbour node j, the adversarial loss function is shown in Eq. (9).

$$\sum_{n_j \in V(t_i)} \epsilon(y_i, g_\theta(n_j)) \tag{9}$$

5. Experimental setting

This section shows the guidelines followed during the experimentation. What kind of experimentation has been performed, how it has been performed, and the motivations related to such experimentation.

All experiments have been conducted using a computer with an Intel(R) Core(TM) i7-9700F CPU 3.00 GHz, 16 GB of RAM and an NVIDIA RTX 2080 Super GPU. Details of experimentation are shown through GitHub.[3]

5.1. Datasets partitions

In the first batch of experiments, the model is trained and tested with each of the datasets following the predefined partitions for each of them. Each of the datasets opts for a different type of labelling. These types of labelling can be grouped into three. Firstly, frame-level labelling, where each frame has a label associated with it depending on whether it is a frame that harbours violent behaviour or not, this is the case of UBI Fights. Secondly, interval labelling, in which for each instance of the dataset the interval in frames or seconds in which the violent action occurs is provided, this is the case of NTU CCTV Fights (intervals per second) and XD Violence (intervals per frame). And the third and last case is video-level labelling, where the entire instance of the dataset has a single label associated with it that classifies the entire video, this is the case of UCF Crime training set (the testing set is labelled using intervals per frame). We consider the finest labelling to be at frame level so we will use it for our model. For interval–labelled datasets, we take each interval and label each frame belonging to that interval based on interval class. In the case of UCF Crime training set, all frames belonging to an instance will have the same label as the instance.

The datasets provide pre-defined guidelines for their division into training, validation and testing datasets. Rather than creating our own divisions, we strictly follow the guidelines provided by each of the datasets: NTU CCTV Fights uses three randomly selected partitions: 50% training, 25% validation and 25% testing; UBI Fights three fixed subsets (80%, 5% and 15%); XD Violence use 3954 videos for training and 800 for testing; UCF Crime uses 800 normal and 810 anomalous videos for training and 150 normal and 140 anomalous for testing in 4-fold cross-validation. The divisions are made at the instance level (videos) so first, the partition is made into training, validation and test subsets and then the frames are extracted from each instance and labelled.

5.2. Preparation of ablation study data

The second batch of experiments aims to see the contributions of NSL to a deep learning model, in this case, a ViT. For this purpose, an ablation study is performed where NSL is eliminated, including the similarity graph using adversarial learning, and replaced by supervised learning, only standard optical flow frames are used as input of the ViT. In this ablation study, the same type of experimentation is performed as with the first batch (following the predefined partitions), substituting one type of learning (ViT + NSL) for another (only ViT). The previous and this experiments are called In-domain experiments.

5.3. Cross-datasets data preparation

In the last batch of experiments, the objective is to measure the generality of the concept of violent action by means of cross-datasets experiments (do not confuse it with cross-validation).

[3] https://github.com/FernandoJRS/CrimeNet-ViT-NSL

Fig. 6. Comparison between two frame sequences of the classes Car Accident XD Violence (Top) and Road Accident UCF Crime (Bottom).

Table 2
Matching classes between the UCF Crime and XD Violence datasets.

UCF crime classes	XD violence classes	Match classes
Normal	Normal	✓
Abuse	Abuse	✓
Arrest	–	×
Arson	–	×
Assault	–	×
Burglary	–	×
Explosion	Explosion	✓
Fighting	Fighting	✓
–	Riot	×
Road Accident	Car Accident	✓
Robbery	–	×
Shooting	Shooting	✓
Shoplifting	–	×
Stealing	–	×
Vandalism	–	×

For this purpose, the whole of the source dataset is used as the training set and the whole of the target dataset with which the model is to be evaluated as the test dataset.

On the one hand, the single-class datasets, NTU CCTV Fights, and UBI Fights are crossed with each other by training the model with one set and validating with the other. On the other hand in the multi-class datasets, XD Violence and UCF Crime, cross-datasets experiment with all the classes cannot be performed since they do not have the same amount of classes and not all of them are coincident, for that reason, the model must be trained only with those classes coincident between both datasets. The matched classes between XD Violence and UCF Crime are shown in the Table 2. Given the similarity of the instances of the classes 'Road Accident' in UCF Crime and 'Car Accident' in XD Violence, we consider them matched classes.

Two sequences of frames from the classes Car Accident from the XD Violence dataset and Road Accident from the UCF Crime dataset are shown in Fig. 6. The frames from the UCF Crime dataset are of lower quality than those from XD Violence, in this case from a movie. However, both classes capture the same concept of traffic accidents and can be compared and considered the same class despite the difference in quality and camera focus.

5.4. Metrics

The metrics used in the experiments are as follows:

- Confusion Matrix (CM): Allows the display of the number of predictions made by a model that matches the labels.
- Receiver Operating Characteristic Area Under the Curve (ROC AUC): Calculates sensitivity versus specificity for a classifier system as the discrimination threshold is varied.
- Average Precision (AP): Summarises the area under the precision–recall curve (PR AUC) as the weighted average of the accuracy achieved at each threshold.

6. In-domain: Results from each dataset

The results in Table 3 show the effectiveness of CrimeNet with respect to its competitors. More precisely, Fig. 7 presents the nearly perfect confusion matrices of CrimeNet for the multi-class XD Violence and UCF Crime datasets. We highlight that the UCF Crime dataset comes with four different train/test dataset splits: for all of them the CrimeNet results are stable with almost zero standard deviation, confirming the robustness of the model. The inference time for these results is about 40 ms.

By reporting the results of ViT we provide an ablation on the role of NSL: indeed CrimeNet builds over ViT and further exploits the sample neighbour graph via NSL. The results indicate that CrimeNet has an advantage over ViT of around 10% points confirming that NSL generates a more robust model by establishing stronger relationships between similar image features. These surprisingly good results are in agreement with those obtained in other works using NSL such as Uddin and Soylu (2021).

We note that ViT is already surpassing all the state of the art models except in the case of UBI Fights where results are very close in ROC AUC to Sultani et al. (2018) (89.76% vs 90.60%). The approximate inference time of ViT remains at 40 ms, showing the inclusion of NSL in the model does not cause a prediction delay.

Finally, CrimeNet outperforms the current state of the art by far. For the less studied datasets such as NTU CCTV Fights and UBI Fights, perfect results are obtained, which is 20.5% points higher in AP for NTU CCTV Fights and 9.4% points higher ROC AUC in UBI Fights. For the case of the multi-class datasets, the results obtained are again significantly higher than the state of the art. For the XD Violence dataset, the improvement is 22.17% in ROC AUC while for the UCF Crime dataset is 14.6%.

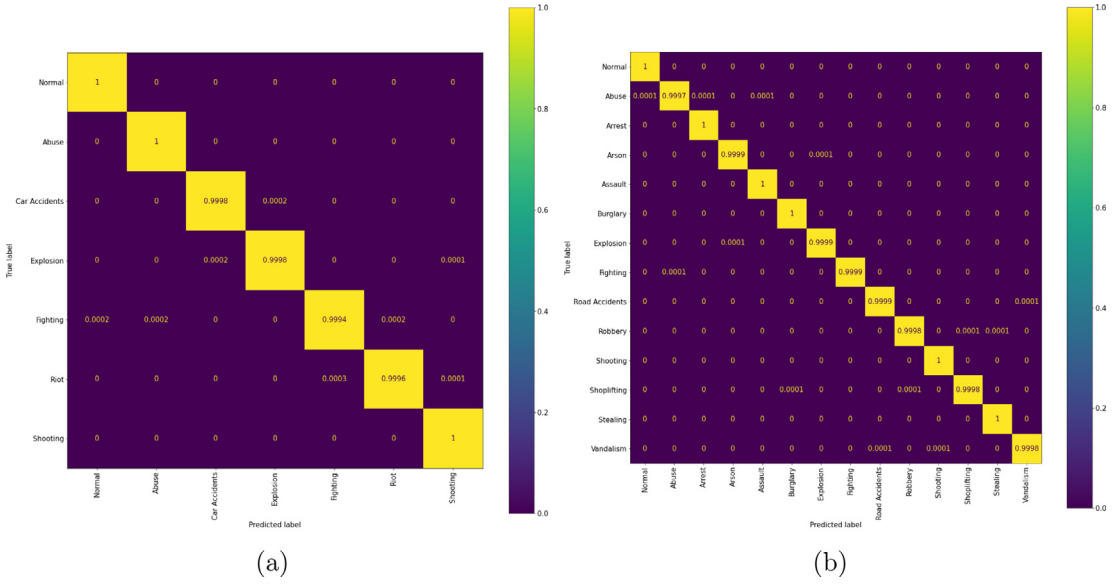
7. Cross-domain: Results across datasets

Given the results of our model CrimeNet using ViT and NSL so close to 100% of the metrics, this experiment is essential to check if overfitting is occurring. The results of this experiment show a good performance of the model in generalising the concept of violent action. Figs. 8 and 9 show the confusion matrices for the cross-datasets experiment between NTU CCTV Fights - UBI Fights and XD Violence-UCF Crime datasets (retrained using the common classes shown in Table 2).

Table 4 shows the accuracy results for each of the cross-datasets experiments. These results show that the single-class datasets generalise the concept of violence better, achieving results of around 80% accuracy. The multi-class datasets show significantly lower performance, but higher than 70% in both cases.

Table 3
Comparison of CrimeNet results with the state of the art for the case study datasets. Our ablation study using only ViT is also included.

Dataset	Method	ROC AUC	AP
NTU CCTV Fights (Perez et al., 2019)	3D CNN (Perez et al., 2019)	–	79.50%
	ViT	90.45%	90.40%
	CrimeNet	**100%**	**100%**
UBI Fights (Degardin, 2020)	BayesianNet (Degardin & Proença, 2021)	84.60%	–
	ViT	89.76%	89.73%
	GMM (Degardin, 2020)	90.60%	–
	CrimeNet	**100%**	**100%**
XD Violence (Wu et al., 2020)	HL-Net (Wu et al., 2020)	–	78.64%
	Contrastive Attention (Chang et al., 2021)	–	76.90%
	RTFM (Tian et al., 2021)	77.81%	–
	ViT	87.23%	87.20%
	CrimeNet	**99.98%**	**99.95%**
UCF Crime (Sultani et al., 2018)	DMIL Ranking (Sultani et al., 2018)	75.41%	–
	GMM (Degardin, 2020)	75.90%	–
	3D ResNet (Dubey, Boragule, Jeon, 2019)	76.67%	–
	DMIL AutoEncoder (Kamoona et al., 2020)	80.10%	–
	DMIL-MRM (Dubey, Boragule, Gwak, et al., 2021)	81.91%	–
	GCN (Zhong et al., 2019)	82.12%	–
	MIST (Feng et al., 2021)	82.30%	–
	RTFM (Tian et al., 2021)	84.30%	–
	Contrastive Attention (Chang et al., 2021)	84.62%	–
	WSAL (Lv et al., 2021)	85.38%	–
	ViT	87.50%	87.50%
	CrimeNet	**99.98%**	**99.97%**



(a) (b)

Fig. 7. (a) XD Violence Confusion Matrix, (b) UCF Crime Confusion Matrix.

It should be noted that none of the state of the art works studied used this experiment to check the robustness of their models, so we carried out it with the state of the art models that have obtained the best results for each dataset in Table 3, that is, GMM (Degardin, 2020) for UBI Fights, RTFM (Tian et al., 2021) for XD Violence and WSAL (Lv et al., 2021) for UCF Crime and whose code is available to reproduce. ViT (without NSL) is also included to check the performance of CrimeNet over the state of the art and ViT. As it can be seen, CrimeNet far outperforms the cross-datasets results of the other models, improving in 25.22% points ROC AUC for UBI Fights, 18.73% for XD Violence and 12.39% for UCF Crime respectively. For the NTU CCTV Fights dataset, there is no reproducible code is available with which to obtain results for comparison.

When XD Violence is used as the training dataset, better generalisation can be observed than when UCF Crime is used. We consider that the size of the first dataset, with more than twice as many videos as the second one, allows better training and ROC AUC than the second one (73.5% vs. 70.20%).

The ablation study on the role of NSL again shows that CrimeNet has a significant advantage over ViT, around 10% (+9.47%, +11.06%, +10.89%, and +9.47%) in the cross-validation experiments. This difference is in agreement with the results obtained in the In-domain experiments and shows that ViT applied to video overpass state of the art but NSL includes much more robustness.

7.1. Recommendations

The practically perfect in-domain results motivated us to carry out the cross-dataset experiment, showing that CrimeNet maintains its advantage over the competitors even in this challenging setting.

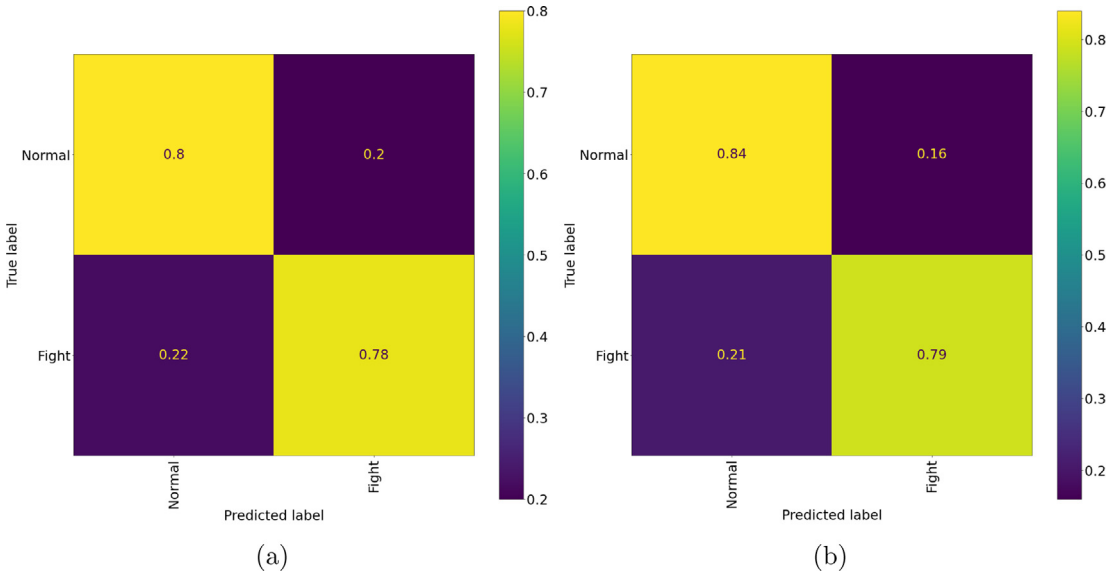


Fig. 8. (a) Crossvalidation NTU CCTV Fights - UBI Fights, (b) Crossvalidation UBI Fight - NTU CCTV Fight.

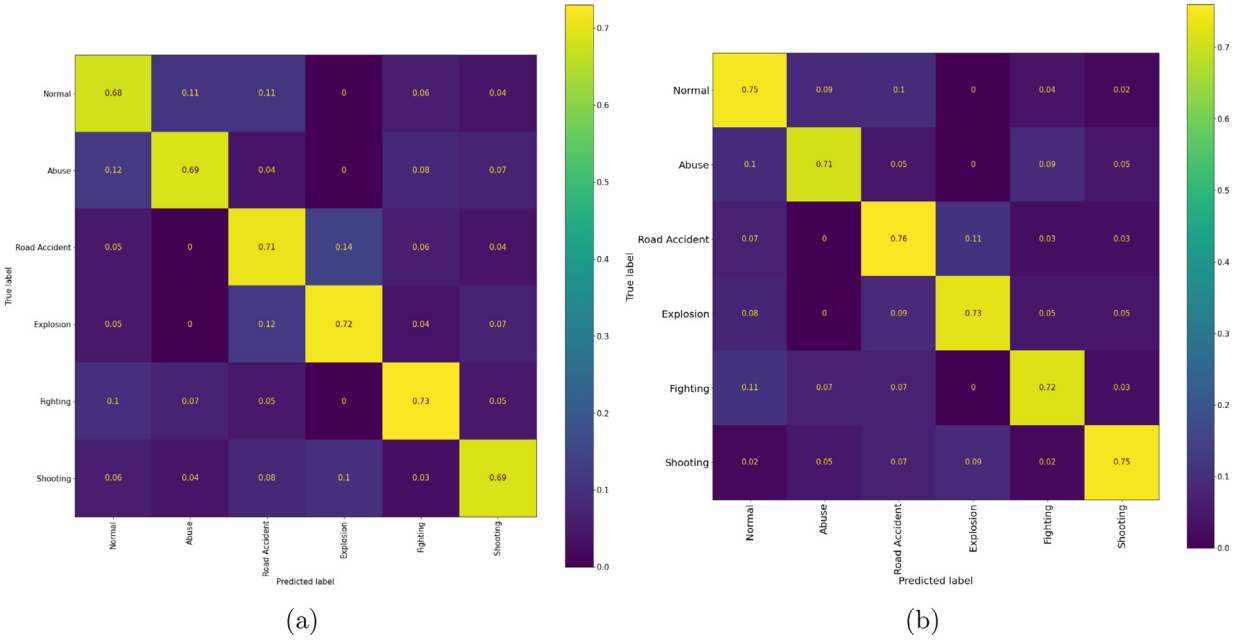


Fig. 9. (a) Crossvalidation UCF Crime - XD Violence, (b) Crossvalidation XD Violence - UCF Crime.

Table 4
Cross-datasets experiments between the four datasets. Binary: NTU CCTV Fights and UBI Fights; Multiclass: UCF Crime and XD Violence. Our ablation study using only ViT is also included.

Dataset train	Dataset test	Method	ROC AUC	AP
NTU CCTV Fights[a] (Perez et al., 2019)	UBI Fights (Degardin, 2020)	ViT	69.43%	69.50%
		CrimeNet	**78.90%**	**78.87%**
UBI Fights (Degardin, 2020)	NTU CCTV Fights (Perez et al., 2019)	GMM (Degardin, 2020)	56.13%	–
		ViT	70.29%	70.26%
		CrimeNet	**81.35%**	**81.35%**
XD Violence (Wu et al., 2020)	UCF Crime (Sultani et al., 2018)	RTFM (Tian et al., 2021)	54.77%	–
		ViT	62.59%	62.55%
		CrimeNet	**73.50%**	**73.44%**
UCF Crime (Sultani et al., 2018)	XD Violence (Wu et al., 2020)	WSAL (Lv et al., 2021)	57.81%	–
		ViT	60.73%	60.73%
		CrimeNet	**70.20%**	**70.10%**

[a]Code not available to reproduce.

Interestingly we noticed that our result in the cross-dataset experiment training in UBI Fights and testing in NTU CCTV Fights, achieves a better performance than the current state of the art (81.35% vs 79.50% Perez et al., 2019).

We also consider that the cross-dataset experiment is a good practice that must be used in these benchmarks, which normally is not present in the related works.

On the basis of the results we showed, it is clear that this challenging setting is still in need of further research.

Finally, we showed how NSL approach using Adversarial learning for generating the similarity graph is a very good approach to obtaining better results and making the model more robust.

8. Conclusions and future work

In this paper, we present a case study on the detection of violent events in videos and the generalisation of this concept. To address the case study we proposed an innovative deep learning model that combines NSL with ViT architectures called CrimeNet.

This model has far surpassed the current state of the art in violence detection after analysing four datasets and outperforming each of the best performing works by 9.4 to 22.17% points in ROC AUC.

In view of the results, we can state that the combination of ViT with NSL and Adversarial learning for the similarity graph input generates a unique model with high efficiency and robustness in the task of video detection of violence. After the ablation study, we checked the importance of using NSL instead of supervised learning directly. It is shown that applying NSL improves the results by about 10% points compared to supervised learning (from 9.55 to 12.75% points in ROC AUC).

The model also shows a state of the art performance when carrying out cross-dataset experiments to check the generalisation of the concept of violent action. The results reach values above 70% ROC AUC for multiclass (73.5% and 70.2%) and close to 80% for binary-class datasets (81.35% and 78.9%), meaning an improvement with respect to previous models from 12.39 to 25.22 points.

These results again manifest the robustness of the model and the possibility to use this approach in a number of lines of future work such as video classification, detection, classification of human actions, tracking of elements in a video, etc. Another study that we want to carry out is the comparison of different backbones, substituting for instance ViT by 3D ResNet, and the generation of different types of data augmentation such as mixup (Zhang, Cisse, Dauphin & Lopez-Paz, 2018).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research is partially supported by the DISARM project - Grant n. PDC2021-121197, and the HORUS project - Grant n. PID2021-126359OB-I00 funded by MCIN/AEI/310.13039/5011000 11033 and by the "European Union NextGenerationEU/PRTR".

References

Ainsworth, T. (2002). Buyer beware. *Security Oz*, *19*, 18–26.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6836–6846).

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *International conference on computer analysis of images and patterns* (pp. 332–339). Springer.

Bui, T. D., Ravi, S., & Ramavajjala, V. (2018). Neural graph learning: Training neural networks using graphs. In *Proceedings of the eleventh ACM International conference on web search and data mining* (pp. 64–71).

Chang, S., Li, Y., Shen, J. S., Feng, J., & Zhou, Z. (2021). Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia*.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM Symposium on theory of computing* (pp. 380–388).

Chen, Y., Cao, Y., Hu, H., & Wang, L. (2020). Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10337–10346).

Degardin, B. M. (2020). *Weakly and partially supervised learning frameworks for anomaly detection* (Ph.D. thesis), Portugal: Universidade da Beira Interior.

Degardin, B., & Proença, H. (2021). Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognition Letters*, *145*, 50–57.

Deniz, O., Serrano, I., Bueno, G., & Kim, T. -K. (2014). Fast violence detection in video. In *2014 International conference on computer vision theory and applications: Vol. 2* (pp. 478–485). IEEE.

Ding, C., Fan, S., Zhu, M., Feng, W., & Jia, B. (2014). Violence detection in video by using 3D convolutional neural networks. In *International symposium on visual computing* (pp. 551–558). Springer.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Dubey, S., Boragule, A., Gwak, J., & Jeon, M. (2021). Anomalous event recognition in videos based on joint learning of motion and appearance with multiple ranking measures. *Applied Sciences*, *11*(3), 1344.

Dubey, S., Boragule, A., & Jeon, M. (2019). 3D ResNet with ranking loss function for abnormal activity detection in videos. In *2019 International conference on control, automation and information sciences* (pp. 1–6). IEEE.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on image analysis* (pp. 363–370). Springer.

Feng, J. -C., Hong, F. -T., & Zheng, W. -S. (2021). MIST: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14009–14018).

Gao, M., Cai, W., & Liu, R. (2021). AGTH-Net: Attention-based graph convolution-guided third-order hourglass network for sports video classification. *Journal of Healthcare Engineering*, *2021*.

Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 244–253).

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations*. URL http://arxiv.org/abs/1412.6572.

Gopalan, A., Juan, D. -C., Magalhaes, C. I., Ferng, C. -S., Heydon, A., Lu, C. -T., et al. (2021). Neural structured learning: Training neural networks with structured signals. In *Proceedings of the 14th ACM international conference on web search and data mining* (pp. 1150–1153).

Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 1–6). IEEE.

Jahanbakht, M., Xiang, W., & Azghadi, M. R. (2022). Sediment prediction in the great barrier reef using vision transformer with finite element analysis. *Neural Networks*, *152*, 311–321.

Jin, W., Ma, Y., Liu, X., Tang, X., Wang, S., & Tang, J. (2020). Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 66–74).

Juan, D. -C., Lu, C. -T., Li, Z., Peng, F., Timofeev, A., Chen, Y. -T., et al. (2020). Ultra fine-grained image semantic embedding. In *Proceedings of the 13th international conference on web search and data mining* (pp. 277–285).

Kamoona, A. M., Gosta, A. K., Bab-Hadiashar, A., & Hoseinnezhad, R. (2020). Multiple instance-based video anomaly detection using deep temporal encoding-decoding. arXiv Preprint arXiv:2007.01548.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, *54*(10s), http://dx.doi.org/10.1145/3505244.

Li, X., Chen, M., Nie, F., & Wang, Q. (2017a). A multiview-based parameter free framework for group detection. In *Thirty-first AAAI conference on artificial intelligence*.

Li, X., Chen, M., Nie, F., & Wang, Q. (2017b). Locality adaptive discriminant analysis. In *IJCAI: Vol. 2201*, (2207).

Liu, Z., Luo, S., Li, W., Lu, J., Wu, Y., Sun, S., et al. (2020). Convtransformer: A convolutional transformer network for video frame synthesis. arXiv preprint arXiv:2011.10185.

Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., & Yang, J. (2021). Localizing anomalies from weakly-labeled videos. *IEEE Transactions on Image Processing*, *30*, 4505–4515.

Mahmoodi, J., & Salajeghe, A. (2019). A classification method based on optical flow for violence detection. *Expert Systems with Applications*, *127*, 121–127.

Paul, S., & Chen, P. -Y. (2022). Vision transformers are robust learners. In *AAAI conference on artificial intelligence*.

Perez, M., Kot, A. C., & Rocha, A. (2019). Detection of real-world fights in surveillance videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 2662–2666). IEEE.

Ren, Y., Wang, B., Zhang, J., & Chang, Y. (2020). Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE international conference on data mining* (pp. 452–461). IEEE.

Rendón-Segador, F. J., Álvarez-García, J. A., Enríquez, F., & Deniz, O. (2021). Violencenet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence. *Electronics*, *10*(13), 1601.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sadeghian, A., Alahi, A., & Savarese, S. (2017). Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE international conference on computer vision* (pp. 300–311).

Salazar González, J. L., Zaccaro, C., Álvarez-García, J. A., Soria Morillo, L. M., & Sancho Caparrini, F. (2020). Real-time gun detection in CCTV: An open problem. *Neural Networks*, *132*, 297–308.

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7912–7921).

Steiner, A. P., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2022). How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, URL https://openreview.net/forum?id=4nPswr1KcP.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., & Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4975–4986).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347–10357). PMLR.

Uddin, M. Z., & Soylu, A. (2021). Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. *Scientific Reports*, *11*(1), 1–15.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Velastin, S. A., Boghossian, B. A., & Vicencio-Silva, M. A. (2006). A motion-based image processing system for detecting potentially dangerous situations in underground railway stations. *Transportation Research Part C (Emerging Technologies)*, *14*(2), 96–113.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., et al. (2018). Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(11), 2740–2755.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., et al. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision* (pp. 322–339). Springer.

Xu, T., & Takano, W. (2021). Graph stacked hourglass networks for 3D human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16105–16114).

Yin, J., Shen, J., Gao, X., Crandall, D., & Yang, R. (2021). Graph neural network and spatiotemporal transformer attention for 3D video object detection from point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International conference on learning representations*.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 335–340).

Zhong, J. -X., Li, N., Kong, W., Liu, S., Li, T. H., & Li, G. (2019). Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1237–1246).

Zhou, P., Ding, Q., Luo, H., & Hou, X. (2018). Violence detection in surveillance video using low-level features. *PLoS One*, *13*(10), Article e0203668.