







# Unsupervised Domain Adaptation for Video Violence Detection in the Wild

Luca Ciampi<sup>1</sup><sup>a</sup>, Carlos Santiago<sup>2</sup><sup>b</sup>, Joao Paulo Costeira<sup>2</sup><sup>c</sup>, Fabrizio Falchi<sup>1</sup><sup>d</sup>,  
Claudio Gennaro<sup>1</sup><sup>e</sup> and Giuseppe Amato<sup>1</sup><sup>f</sup>

<sup>1</sup>*Institute of Information Science and Technologies, National Research Council, Pisa, Italy*

<sup>2</sup>*Instituto Superior Técnico (LARSyS/IST), Lisbon, Portugal*


**Keywords:** Video Violence Detection, Video Violence Classification, Action Recognition, Unsupervised Domain Adaptation, Deep Learning, Deep Learning for Visual Understanding, Video Surveillance.


**Abstract:** Video violence detection is a subset of human action recognition aiming to detect violent behaviors in trimmed video clips. Current Computer Vision solutions based on Deep Learning approaches provide astonishing results. However, their success relies on large collections of labeled datasets for supervised learning to guarantee that they generalize well to diverse testing scenarios. Although plentiful annotated data may be available for some pre-specified domains, manual annotation is unfeasible for every ad-hoc target domain or task. As a result, in many real-world applications, there is a domain shift between the distributions of the train (source) and test (target) domains, causing a significant drop in performance at inference time. To tackle this problem, we propose an Unsupervised Domain Adaptation scheme for video violence detection based on single image classification that mitigates the domain gap between the two domains. We conduct experiments considering as the source labeled domain some datasets containing violent/non-violent clips in general contexts and, as the target domain, a collection of videos specific for detecting violent actions in public transport, showing that our proposed solution can improve the performance of the considered models.


## 1 INTRODUCTION


In recent years, in the Computer Vision field, there has been an increasing interest in developing applications and services that make life easier for citizens. Thanks to the significant growth of Deep Learning (DL) and the ubiquity of video surveillance cameras in modern cities, smart applications ranging from pedestrian detection (Amato et al., 2019b) (Ciampi et al., 2020a) (Cafarelli et al., 2022) to people tracking (Spremlola et al., 2016) (Staniszewski et al., 2020), crowd counting (Benedetto et al., 2022) (Avvenuti et al., 2022), parking lot management (Ciampi et al., 2022b) (Amato et al., 2019a) (Amato et al., 2018) (Ciampi et al., 2018) and even facial reconstruction (Peşzor


et al., 2016) have been proposed and are nowadays widely employed worldwide, helping to manage public spaces and preventing many criminal activities by exploiting AI systems that automatically analyze this deluge of visual data. However, the success of these supervised DL-based approaches hinges on two assumptions: (i) the existence of large collections of labeled data required for accurate model fitting during the training phase, and (ii) training (or *source*) and test (or *target*) datasets are independent and identically distributed (*i.i.d.*) (Huo et al., 2022). Although plentiful annotated data may be available for a few pre-specified domains, such as ImageNet (Deng et al., 2009) for image classification or COCO (Lin et al., 2014) for object detection, manual annotations are often prohibitive to obtain for every ad-hoc target domain or task. As a result, models trained by leveraging already existing labeled data are applied to target domains never seen during the training and consequently suffer from shifts in data distributions, i.e., *Domain Shifts* between source and target domains (Torralba and Efros, 2011).


<sup>a</sup> <https://orcid.org/0000-0002-6985-0439>

<sup>b</sup> <https://orcid.org/0000-0002-4737-0020>

<sup>c</sup> <https://orcid.org/0000-0001-6769-2935>

<sup>d</sup> <https://orcid.org/0000-0001-6258-5313>

<sup>e</sup> <https://orcid.org/0000-0002-3715-149X>

<sup>f</sup> <https://orcid.org/0000-0003-0171-4315>

One possible solution to tackle this issue is represented by *Unsupervised Domain Adaptation* - (UDA). Specifically, it aims at mitigating domain shifts between different domains, relying on labeled data in the source domain and *unlabelled* data in the target domain. In other words, UDA techniques exploit annotated data from the source domain as well as *non-annotated* data coming from the target domain that is easy to gather since it does not require human effort for labeling. The challenge here is to automatically infer some knowledge from this latter data flow to reduce the gap between the two domains and, specifically, to learn feature representations that should be (i) discriminative for the main learning task on the source domain and (ii) indiscriminative concerning the shift between the domains.

In this work, we focus on the specific task of violence detection in *trimmed* videos, i.e., capturing an exact action (either violent or non-violent). Therefore, this task is a subset of human action recognition. Specifically, the goal is to binary classify clips to predict if they contain (or not) any behaviors considered to be violent, differing from violent detection in *untrimmed* videos, a subset of action localization where the purpose is also to seek the action in the temporal dimension. Despite its importance in many practical, real-world scenarios, it is relatively unexplored compared to other action recognition tasks. Although some annotated datasets for video violence detection in general contexts already exist, they are limited in size and in the considered different scenarios. Therefore, existing Deep Learning-based solutions trained using these data systematically experience performance degradation when applied to new specific contexts, such as violence detection in public transport environments (Ciampi et al., 2022a).

To mitigate this problem, in this paper, we propose an end-to-end DL-based UDA solution to detect violent situations in videos in specific target scenarios where annotated data is scarce or lacking. Our proposal relies on *single* image classification randomly sampled from the frames making up the video, a simple technique already addressed by (Akti et al., 2022). Starting from this, some UDA techniques for image classification are employed during the training pipeline, automatically gathering some knowledge from the unlabeled data belonging to the target domain. To the best of our knowledge, it is the first attempt at using a UDA schema for video violence detection. We conducted experiments by exploiting, as the source domain, several annotated datasets present in the literature dealing with video violence detection in general contexts and, as the target domain, the recently introduced *Bus Violence* benchmark (Ciampi

et al., 2022a), a collection of clips specific for detection of violent behaviors inside a moving bus. Experimental results show that by using our UDA pipeline, we can improve the performance of the considered models by a significant margin, thus suggesting that they generalize better over this new scenario without the need to use new labels.

Summarizing, the contribution of this work can be listed as follows:

- we introduce a UDA scheme for video violence detection based on single-image classification, which can mitigate the domain gap between a labeled source dataset and an unlabeled target one: to the best of our knowledge, this is the first time that UDA has been applied to video violence detection;
- we conduct an experimental evaluation taking into account as the source domain some annotated dataset containing violent/non-violent clips in general contexts and, as the target domain, a recently introduced collection of videos specific for detection of violent behaviors in public transport;
- preliminary results show that our proposed UDA scheme can improve the performance of the considered models, which can better generalize against new scenarios for which labels are absent.

The rest of the paper is structured as follows. Section 2 reviews some works related to ours. Section 3 describes the proposed methodology. Section 4 shows the performed experimental evaluation. Finally, Section 5 concludes the paper, suggesting some insights on future directions.

## 2 RELATED WORKS

In the literature, there are several methods and datasets specific to video violence detection. Most deal with *trimmed* clips, i.e., capturing an exact action (either violent or non-violent). Therefore, this task lies with action recognition aiming at binary classifying videos to predict if they contain (or not) violent human behaviors. On the other hand, a few works also deal with *untrimmed* videos. In this case, the task is no longer a subset of action recognition but is treated as action *localization*, i.e., it is also needed to seek the actions' starting and ending time points. This distinction is also reflected in the datasets required for the learning phase: in the former case, they are annotated at a video level, while in the second case, frame-annotated data is necessary. In this paper, we consider video violence detection in *trimmed* videos. Hereafter we describe some of the more popular techniques and

collections of trimmed clips in the literature, concluding the section by reviewing some existing UDA approaches.

## 2.1 Video Violence Detection Methods

In (Sudhakaran and Lanz, 2017), the authors introduced a Deep Learning-based model consisting of a series of convolutional layers for spatial features extraction, followed by Convolutional Long Short Memory (ConvLSTM) (Shi et al., 2015) for encoding the frame level changes. On the other hand, a variant of this architecture is presented in (Hanson et al., 2019), where a spatio-temporal encoder built on a standard convolutional backbone for features extraction is combined with the Bidirectional Convolutional LSTM (BiConvLSTM) architecture for extracting the long-term movement information present in the clips. Differently, the authors in (Akti et al., 2022) proposed classifying videos using single frames randomly sampled from the clips. Alternatively, it is also possible to exploit methods suitable for human action recognition: in this case, fine-tuning is needed to recognize only two classes – violence and non-violence. For instance, the ResNet 3D network (Tran et al., 2018) considers actions as spatiotemporal objects and handles both spatial and temporal dimensions using 3DConv layers (Tran et al., 2015); on the other hand, the ResNet 2+1D architecture (Tran et al., 2018), decomposes the convolutions into separate 2D spatial and 1D temporal filters (Feichtenhofer et al., 2016). Another popular model is represented by SlowFast (Feichtenhofer et al., 2019). In this two-pathway architecture, the first one is designed to capture the semantic information that can be given by images or a few sparse frames operating at low frame rates. In contrast, the other one is responsible for capturing rapid changing motion by working at a fast refreshing speed. Finally, recently, architecture relying on Transformer attention modules have been introduced, such as Video Swim Transformer (Liu et al., 2022), which extends the sliding-window Transformers proposed for image processing (Liu et al., 2021) to the temporal axis, obtaining an excellent efficiency-effective trade-off.

## 2.2 Video Violence Detection Datasets

In the last years, some benchmarks of trimmed clips suitable for video violence detection have been introduced. In (Padamwar, 2020), the authors presented two video benchmarks for violence detection — the *Hockey Fight* and the *Movies Fight* datasets. The former consists of 200 clips extracted from short movies.

On the other hand, the second one has 1,000 fight and non-fight clips from the ice hockey game. More recently, another dataset, named *Surveillance Camera Fight*, has been presented in (Akti et al., 2019). It consists of 300 videos in total, 150 of which describe fight sequences and 150 depict non-fight scenes, recorded from several surveillance cameras located in public spaces. Moreover, the *RWF-2000* (Cheng et al., 2021) and the *Real-Life Violence Situations* (Soliman et al., 2019) datasets were gathered from public surveillance cameras. In both collections, the authors collected 2000 video clips: half of them include violent behaviors, while the others belong to non-violent activities. Finally, in the *Bus Violence* benchmark (Ciampi et al., 2022a), the authors gathered and made publicly available 1,400 videos of violent/non-violent actions simulated by several actors in a moving bus.

## 2.3 Unsupervised Domain Adaptation

Traditional UDA approaches have been developed to address the problem of image classification, and they try to align features across the two domains. Some notable examples are (Ganin et al., 2016) (Jin et al., 2020) (Tzeng et al., 2017). However, their usage in other applications is not straightforward, as pointed out by (Zhang et al., 2017), and in the literature, there are a limited number of UDA approaches suitable for different tasks. More recent advances involve semantic segmentation (Hong et al., 2018) (Chen et al., 2019) and visual counting (Ciampi et al., 2020b) (Ciampi et al., 2021). In this work, we propose a UDA scheme for video violence detection in videos. To the best of our knowledge, it is the first attempt to exploit UDA in this task.

# 3 METHOD

## 3.1 Background

Following the notation introduced in (Pan and Yang, 2010) (Csurka, 2017), we define a *domain*  $\mathcal{D}$  consisting of two components: a  $d$ -dimensional feature space  $\mathcal{X} \subset \mathbb{R}^d$  and a marginal probability distribution  $P(X)$ , where  $X = \{x_1, \dots, x_n\} \subset \mathcal{X}$ . Given a specific domain,  $\mathcal{D} = \{\mathcal{X}, P(X)\}$ , we formulate a *task*  $\mathcal{T}$  defined by a label space  $\mathcal{Y}$  and the conditional probability distribution  $P(Y|X)$ , where  $Y = \{y_1, \dots, y_n\} \subset \mathcal{Y}$  is the set of the corresponding labels for  $X$ . In general,  $P(Y|X)$  can be learned in a supervised manner from these feature-label pairs  $\langle x_i, y_i \rangle$ .

When considering Unsupervised Domain Adaptation (UDA), there is (i) a *source* domain  $\mathcal{D}_S =$

$\{X_S, P(X_S)\}$  with  $\mathcal{T}_S = \{\mathcal{Y}_S, P(Y_S|X_S)\}$  and (ii) a *target* domain  $\mathcal{D}_T = \{X_T, P(X_T)\}$  with  $\mathcal{T}_T = \{\mathcal{Y}_T, P(Y_T|X_T)\}$ , where  $\mathcal{Y}_T$  is unknown, i.e., we do not have any labels. Due to the difference between the two domains, the distributions are assumed to be different, i.e.,  $P(X_S) \neq P(X_T)$  and  $P(Y_S|X_S) \neq P(Y_T|X_T)$ . UDA aims to learn a model with lower generalization error in the target domain by mitigating the domain discrepancy.

### 3.2 UDA for Video Violence Detection

In this work, the source domain  $\mathcal{D}_S$  consists of a labeled set of videos with  $\mathcal{Y}_S = \{0, 1\}$ , where 0 and 1 indicate the non-presence/presence of violent actions occurring in the clips, respectively. Specifically, we considered some general violence detection datasets present in the literature collecting very heterogeneous and everyday life violent and non-violent actions. On the other hand, the target domain  $\mathcal{D}_T$  consists of a different set of videos for which we do not have annotations. In this case, clips include violent/non-violent actions performed in a more specific and different scenario compared to the ones characterizing the source domain. The goal is to infer some knowledge from the unlabeled target domain during the training phase, mitigating the domain discrepancy present with the source domain so that the model can be able to better generalize to the new specific scenario for which the annotations are absent.

Our method relies on Deep Learning-based models trained end-to-end together with some UDA techniques attached to them. The peculiarity of our UDA scheme is that it is based on *image* classification. Specifically, we cast the task of video classification to image classification since the scenes including violent actions can be discriminated from non-violent scenes just by classifying an image randomly sampled from the entire video clip (Akti et al., 2022). Starting from this baseline, we put into the training pipeline two different UDA techniques native for image classification that we fed with images sampled from the target domain, which are responsible for the intra-domain transfer knowledge.

More in detail, we considered some Convolutional Neural Networks (CNNs) as backbones for feature extraction, cutting off the final classification layers. We replaced the last classification head with a binary classification layer, outputting the probability that the given video contains (or does not contain) violent actions, and we added an additional linear layer followed by a ReLU to map the feature maps coming from the feature extractor to a fixed dimension. This latter fixed dimensional feature map is then fed to a

UDA module.

We considered two different UDA strategies. The first one is the *Domain-Adversarial Neural Network* - (*DANN*) (Ganin et al., 2016) where a domain regressor competes against the classifier in an adversarial way. Here, UDA is achieved by connecting the domain classifier to the feature extractor via a gradient reversal layer that produces an adversarial loss by multiplying the gradient by a certain negative constant during the backpropagation-based training. Otherwise, the training proceeds in a standard way by minimizing the label prediction loss (for source examples) and the domain classification loss (for all samples). The adversarial loss ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in domain-invariant features. We refer the reader to (Ganin et al., 2016) for further details. The second one is the *Minimum Class Confusion* - (*MCC*) (Jin et al., 2020), a loss function that can be considered as a UDA approach that does not explicitly perform domain alignment. It is based on class confusion, i.e., the tendency of a classifier to confuse the predictions between the correct and the ambiguous classes. Specifically, given the feature extractor, MCC is defined on the class prediction given by the classifier on the target data. Provided that less class confusion implies more transferability, during the training pipeline, MCC is optimized using standard backpropagation to obtain more generalized features. We refer the reader to (Jin et al., 2020) for further details.

## 4 PERFORMANCE ANALYSIS

### 4.1 Evaluation Metrics

Following previous works regarding video violence detection, we used *Accuracy* to measure the performance of the considered methods, defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

where TP, TN, FP, and FN are the True Positives, True Negatives, False Positives, and False Negatives, respectively. To have a more in-depth comparison between the obtained results, we also considered as metrics the *F1-score*, the *False Alarm*, and the *Missing Alarm*, defined as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (2)$$

$$FalseAlarm = \frac{FP}{TN + FP}, \quad (3)$$

$$MissingAlarm = \frac{FN}{TP + FN}, \quad (4)$$

where Precision and Recall are defined as  $\frac{TP}{TP+FP}$  and  $\frac{TP}{TP+FN}$ , respectively. Finally, to account also for the probabilities of the detections, we employed the *Area Under the Receiver Operating Characteristics (ROC AUC)*, computed as the area under the curve plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold settings, where  $TPR = \frac{TP}{TP+FN}$  and  $FPR = \frac{FP}{TN+FP}$ .

## 4.2 Experimental Setting

We exploited three datasets present in the literature as the source domain — *Surveillance Camera Fight* (Akti et al., 2019), *Real-Life Violence Situations* (Soliman et al., 2019), and *RWF-2000* (Cheng et al., 2021), already mentioned in Section 2. These videos have been gathered from fixed security cameras and include trimmed heterogeneous violent and non-violent scenes, thus containing very general violent situations. On the other hand, we considered the recently introduced *Bus Violence* dataset (Ciampi et al., 2022a) as the target domain. In this case, trimmed clips are recorded inside a moving bus where some actors simulated violent/non-violent actions. This latter scenario is, therefore, more specific as it involves violent situations in public transport, and it represents the perfect testing ground for evaluating the generalization capabilities of Deep Learning models trained with more generic labeled data. We depict the considered scenario in Figure 1.

As the backbone for feature extraction, we considered two popular CNNs — ResNet50 (He et al., 2016) and VGG16 (Simonyan and Zisserman, 2015). As already mentioned, we replaced their final classification head with a binary classification layer and exploited them as baselines, i.e., without any UDA modules, as well as the feature extractors and classifier for our proposed UDA schemes. Furthermore, to compare the obtained results with the literature, we also considered other existing approaches tailored for video violence detection and video action recognition. Specifically, we exploited the architectures introduced in (Sudhakaran and Lanz, 2017) and (Hanson et al., 2019) that employ ConvLSTM and BiConvLSTM as spatio-temporal encoders, together with some popular video action classifiers — the (2+1)D network (Tran et al., 2018), the SlowFast (Feichtenhofer et al., 2019) architecture and the Video Swim Transformer (Liu et al., 2022). We refer the reader

to Section 2 and the related papers for further details about the employed models. For a fair comparison, we accounted for the ImageNet pre-trained versions of all these models as the starting point without using any additional extra data. Furthermore, we always applied the same data augmentation strategy during the learning phase: horizontal flipping with a probability of 0.5 and image resizing to  $256 \times 256$  pixels.

## 4.3 Results and Discussion

We employed the following evaluation protocol to have reliable statistics on the final metrics. For each of the three considered source (training) domains, i.e., *Surveillance Camera Fight*, *Real-Life Violence Situations*, and *RWF-2000*, we randomly varied the training and validation subsets three times, picking up the best model in terms of accuracy and testing it over the target (test) domain, i.e., the *Bus Violence* benchmark. Finally, we reported the mean and the standard deviation of these three runs.

Results are shown Table 1. Overall, all the considered models exhibit moderate performance, indicating the difficulties in generalizing their abilities in detecting violent actions in videos coming from the target domain. However, the model which turns out to be the most performing in terms of the golden metrics, i.e., the *Accuracy*, is the ResNet50 architecture with the MCC UDA module. Specifically, we gain 7.4%, 0.37%, and 12.9% of accuracy compared with the ResNet50 network without UDA concerning the *Surveillance Camera Fight*, the *RWF-2000*, and the *Real-life Violence Situations* source domains, respectively, overcoming all the other considered methods present in the literature.

Considering *False Alarms* and *Missing Alarms*, it can be noted that, in general, all the methods obtained very good results regarding the first metric, while they struggled with the latter. Considering that missing alarms are crucial for video violence detection since they indicate violent actions that happened but were not detected, this represents the main limitation for all the violence detectors. However, it is worth noting that the proposed approach made of the ResNet50 architecture and the MCC module can mitigate this issue, achieving better performance compared with the single ResNet50 model and often overtaking all the other techniques. This behavior is linked with a lower number of detected False Negatives and consequently affects and improves the *Recall* and *F1-score*. In Figure 2, we report some samples of True Positive, True Negative, False Positive, and False Negative coming out from the best model, i.e., the ResNet50 architecture with attached the MCC UDA module.

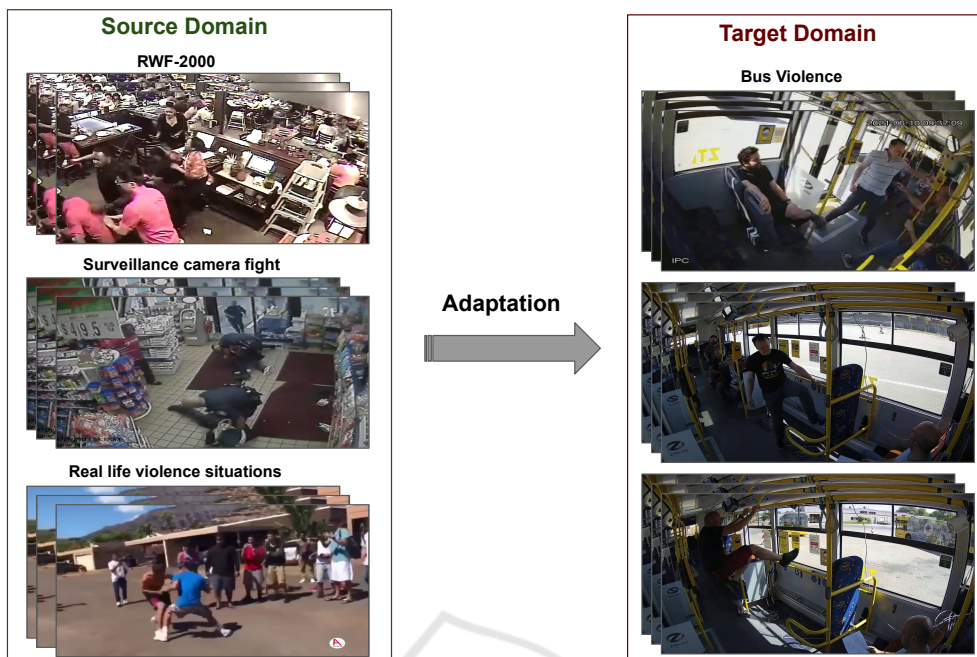


Figure 1: **The considered scenario.** We propose an Unsupervised Domain Adaptation scheme for video violence detection to mitigate the domain gap existing between a source domain (on the left) and a target domain (on the right). The source domain consists of three collections of annotated videos depicting violent/non-violent scenes in general contexts. On the other hand, the target domain is represented by a set of *unlabeled* clips of violent/non-violent actions in public transport.

## 5 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we tackled the problem of video violence detection in the context of data scarcity. Indeed, current Deep Learning solutions hinge on vast quantities of labeled data needed for supervised learning, and they suffer when applied to new scenarios never seen during the training phase. Thus, a model trained on one domain, named *source*, usually experiences a drastic drop in performance when applied on another domain, named *target*. To tackle this issue, we proposed an Unsupervised Domain Adaptation scheme for detecting violent/non-violent actions present in trimmed videos, which relies on supervised learning in the source domain and, at the same time, exploits an *unlabeled* target dataset to reduce the domain shift between the two sets. Our proposed solution is based on *single* image classification, randomly sampled from the frames making up the clips. The feature representations generated by the target images have been hooked and fed to a UDA module responsible for making them indiscriminative concerning the shift between the domains. To the best of our knowledge, it is the first attempt at using a UDA schema for video violence detection. We conducted experiments considering as source domain three datasets

composed of videos of violent/non-violent scenes in general contexts and, as the target domain, a collection of clips of violent/non-violent actions in public transport. Preliminary results showed that our UDA scheme can help to improve the generalization capabilities of the considered models mitigating the domain gap.

In the future, we plan to extend our experimentation by considering and designing other UDA strategies to be attached to the classifier. Indeed, although we obtained a significant performance boost, the considered models still exhibit moderate generalization capabilities, suggesting that a more effective domain gap reduction is needed. Furthermore, we plan to put into the pipeline also the spatio-temporal information provided by consecutive frames making up the clips.

## ACKNOWLEDGEMENTS

This work was partially funded by: AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911); PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by European Union - NextGenerationEU.

Table 1: **Performance Evaluation.** We considered three datasets for video violence detection in general contexts as source domains and a collection of clips with violent situations in public transport as the target domain. We randomly varied the training and validation subsets of the source domains three times, picking up the best model in terms of accuracy. Mean  $\pm$  st.dev is reported.

**Source Domain: Surveillance Camera Fight (Akti et al., 2019) - Target Domain: Bus Violence (Ciampi et al., 2022a)**

Model	Accuracy $\uparrow$	F1 $\uparrow$	False Alarm $\downarrow$	Miss Alarm $\downarrow$	ROC AUC $\uparrow$
(Hanson et al., 2019)	0.54 $\pm$ 0.02	0.19 $\pm$ 0.11	0.04 $\pm$ 0.03	0.89 $\pm$ 0.07	0.68 $\pm$ 0.02
(Sudhakaran and Lanz, 2017)	0.52 $\pm$ 0.01	0.27 $\pm$ 0.18	0.16 $\pm$ 0.17	0.79 $\pm$ 0.18	0.55 $\pm$ 0.02
ResNet (2+1)D (Tran et al., 2018)	0.52 $\pm$ 0.02	0.44 $\pm$ 0.34	0.52 $\pm$ 0.44	0.44 $\pm$ 0.46	0.54 $\pm$ 0.05
SlowFast (Feichtenhofer et al., 2019)	0.55 $\pm$ 0.01	0.40 $\pm$ 0.21	0.27 $\pm$ 0.32	0.62 $\pm$ 0.35	0.62 $\pm$ 0.02
VideoSwimTransformer (Liu et al., 2022)	0.52 $\pm$ 0.01	0.65 $\pm$ 0.01	0.86 $\pm$ 0.01	0.10 $\pm$ 0.01	0.50 $\pm$ 0.01
ResNet50 (He et al., 2016)	0.54 $\pm$ 0.02	0.52 $\pm$ 0.06	0.44 $\pm$ 0.12	0.48 $\pm$ 0.12	0.55 $\pm$ 0.03
VGG16 (Simonyan and Zisserman, 2015)	0.51 $\pm$ 0.01	0.45 $\pm$ 0.07	0.39 $\pm$ 0.21	0.59 $\pm$ 0.19	0.51 $\pm$ 0.01
ResNet50 + DANN (Ganin et al., 2016)	0.55 $\pm$ 0.01	0.51 $\pm$ 0.04	0.39 $\pm$ 0.03	0.51 $\pm$ 0.06	0.56 $\pm$ 0.03
ResNet50 + MCC (Jin et al., 2020)	<b>0.58 <math>\pm</math> 0.01</b>	0.52 $\pm$ 0.03	0.45 $\pm$ 0.05	0.47 $\pm$ 0.04	0.63 $\pm$ 0.01
VGG16 + DANN (Ganin et al., 2016)	0.53 $\pm$ 0.01	0.51 $\pm$ 0.04	0.49 $\pm$ 0.12	0.46 $\pm$ 0.10	0.51 $\pm$ 0.01
VGG16 + MCC (Jin et al., 2020)	0.53 $\pm$ 0.01	0.43 $\pm$ 0.01	0.28 $\pm$ 0.03	0.64 $\pm$ 0.01	0.52 $\pm$ 0.01

**Source Domain: RWF-2000 (Cheng et al., 2021) - Target Domain: Bus Violence (Ciampi et al., 2022a)**

Model	Accuracy $\uparrow$	F1 $\uparrow$	False Alarm $\downarrow$	Miss Alarm $\downarrow$	ROC AUC $\uparrow$
(Hanson et al., 2019)	0.51 $\pm$ 0.01	0.07 $\pm$ 0.03	0.01 $\pm$ 0.01	0.96 $\pm$ 0.02	0.67 $\pm$ 0.05
(Sudhakaran and Lanz, 2017)	0.51 $\pm$ 0.01	0.08 $\pm$ 0.08	0.03 $\pm$ 0.03	0.95 $\pm$ 0.05	0.52 $\pm$ 0.02
ResNet (2+1)D (Tran et al., 2018)	0.53 $\pm$ 0.03	0.43 $\pm$ 0.05	0.29 $\pm$ 0.01	0.64 $\pm$ 0.05	0.54 $\pm$ 0.03
SlowFast (Feichtenhofer et al., 2019)	0.53 $\pm$ 0.03	0.40 $\pm$ 0.10	0.26 $\pm$ 0.08	0.67 $\pm$ 0.12	0.55 $\pm$ 0.03
VideoSwimTransformer (Liu et al., 2022)	0.53 $\pm$ 0.01	0.52 $\pm$ 0.04	0.45 $\pm$ 0.12	0.49 $\pm$ 0.09	0.57 $\pm$ 0.01
ResNet50 (He et al., 2016)	0.54 $\pm$ 0.01	0.49 $\pm$ 0.04	0.34 $\pm$ 0.05	0.56 $\pm$ 0.06	0.58 $\pm$ 0.01
VGG16 (Simonyan and Zisserman, 2015)	0.54 $\pm$ 0.01	0.41 $\pm$ 0.03	0.25 $\pm$ 0.06	0.67 $\pm$ 0.04	0.54 $\pm$ 0.01
ResNet50 + DANN (Ganin et al., 2016)	0.55 $\pm$ 0.01	0.52 $\pm$ 0.01	0.40 $\pm$ 0.01	0.50 $\pm$ 0.01	0.57 $\pm$ 0.01
ResNet50 + MCC (Jin et al., 2020)	<b>0.56 <math>\pm</math> 0.01</b>	0.59 $\pm$ 0.02	0.49 $\pm$ 0.05	0.37 $\pm$ 0.05	0.60 $\pm$ 0.02
VGG16 + DANN (Ganin et al., 2016)	0.55 $\pm$ 0.02	0.52 $\pm$ 0.03	0.39 $\pm$ 0.04	0.51 $\pm$ 0.03	0.54 $\pm$ 0.02
VGG16 + MCC (Jin et al., 2020)	0.55 $\pm$ 0.01	0.41 $\pm$ 0.02	0.20 $\pm$ 0.05	0.69 $\pm$ 0.06	0.55 $\pm$ 0.01

**Source Domain: Real-life Violence Situations (Soliman et al., 2019) - Target Domain: Bus Violence (Ciampi et al., 2022a)**

Model	Accuracy $\uparrow$	F1 $\uparrow$	False Alarm $\downarrow$	Miss Alarm $\downarrow$	ROC AUC $\uparrow$
(Hanson et al., 2019)	0.58 $\pm$ 0.02	0.49 $\pm$ 0.09	0.26 $\pm$ 0.12	0.57 $\pm$ 0.14	0.61 $\pm$ 0.01
(Sudhakaran and Lanz, 2017)	0.52 $\pm$ 0.01	0.45 $\pm$ 0.02	0.35 $\pm$ 0.04	0.61 $\pm$ 0.04	0.55 $\pm$ 0.02
ResNet (2+1)D (Tran et al., 2018)	0.51 $\pm$ 0.01	0.01 $\pm$ 0.01	0.01 $\pm$ 0.01	0.99 $\pm$ 0.01	0.57 $\pm$ 0.08
SlowFast (Feichtenhofer et al., 2019)	0.51 $\pm$ 0.01	0.02 $\pm$ 0.02	0.01 $\pm$ 0.01	0.99 $\pm$ 0.01	0.54 $\pm$ 0.04
VideoSwimTransformer (Liu et al., 2022)	0.51 $\pm$ 0.02	0.30 $\pm$ 0.20	0.22 $\pm$ 0.17	0.76 $\pm$ 0.20	0.53 $\pm$ 0.02
ResNet50 (He et al., 2016)	0.54 $\pm$ 0.01	0.49 $\pm$ 0.03	0.38 $\pm$ 0.08	0.54 $\pm$ 0.06	0.56 $\pm$ 0.01
VGG16 (Simonyan and Zisserman, 2015)	0.53 $\pm$ 0.01	0.54 $\pm$ 0.02	0.33 $\pm$ 0.09	0.51 $\pm$ 0.08	0.58 $\pm$ 0.01
ResNet50 + DANN (Ganin et al., 2016)	0.57 $\pm$ 0.01	0.49 $\pm$ 0.03	0.25 $\pm$ 0.04	0.59 $\pm$ 0.03	0.57 $\pm$ 0.02
ResNet50 + MCC (Jin et al., 2020)	<b>0.61 <math>\pm</math> 0.01</b>	0.54 $\pm$ 0.09	0.32 $\pm$ 0.15	0.51 $\pm$ 0.13	0.61 $\pm$ 0.01
VGG16 + DANN (Ganin et al., 2016)	0.54 $\pm$ 0.01	0.52 $\pm$ 0.03	0.40 $\pm$ 0.05	0.49 $\pm$ 0.03	0.54 $\pm$ 0.02
VGG16 + MCC (Jin et al., 2020)	0.57 $\pm$ 0.01	0.54 $\pm$ 0.04	0.36 $\pm$ 0.08	0.50 $\pm$ 0.08	0.59 $\pm$ 0.01

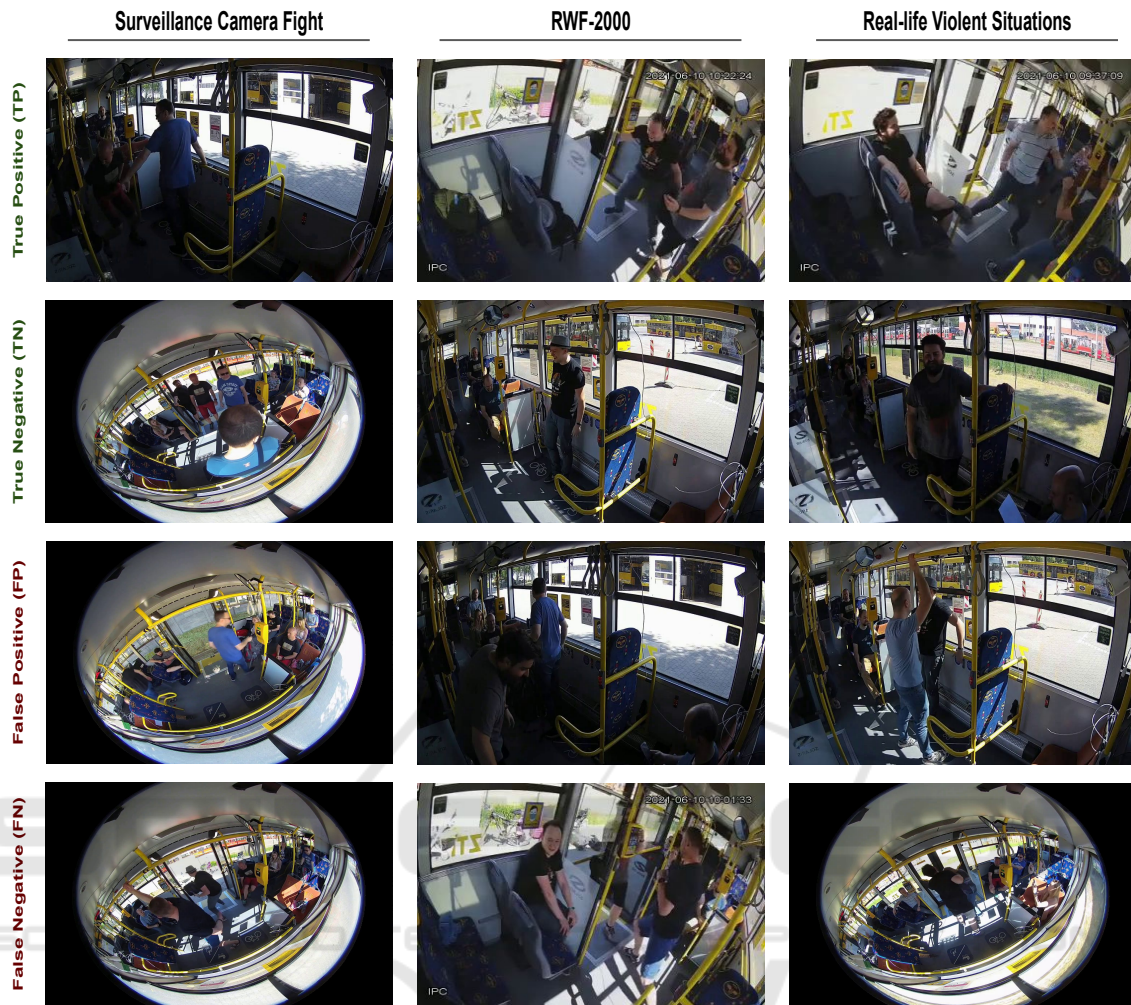


Figure 2: Some samples of predictions over the target domain. In the four rows, we report some samples of True Positives, True Negatives, False Positives, and False Negatives concerning the best model, i.e., ResNet50 + MCC, for each of the considered source domains (one for each column).

## REFERENCES

- Akti, S., Ofli, F., Imran, M., and Ekenel, H. K. (2022). Fight detection from still images in the wild. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE.
- Akti, S., Tataroglu, G. A., and Ekenel, H. K. (2019). Vision-based fight detection from surveillance cameras. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE.
- Amato, G., Bolettieri, P., Moroni, D., Carrara, F., Ciampi, L., Pieri, G., Gennaro, C., Leone, G. R., and Vairo, C. (2018). A wireless smart camera network for parking monitoring. In *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE.
- Amato, G., Ciampi, L., Falchi, F., and Gennaro, C. (2019a). Counting vehicles with deep learning in onboard UAV imagery. In *2019 IEEE Symposium on Computers and Communications (ISCC)*. IEEE.
- Amato, G., Ciampi, L., Falchi, F., Gennaro, C., and Messina, N. (2019b). Learning pedestrian detection from virtual worlds. In *Image Analysis and Processing - ICIAP 2019 - 20th International Conference, Trento, Italy, September 9-13, 2019, Proceedings, Part I*, volume 11751 of *Lecture Notes in Computer Science*, pages 302–312. Springer.
- Avvenuti, M., Bongiovanni, M., Ciampi, L., Falchi, F., Gennaro, C., and Messina, N. (2022). A spatio-temporal attentive network for video-based crowd counting. *CoRR*, abs/2208.11339.
- Benedetto, M. D., Carrara, F., Ciampi, L., Falchi, F., Gennaro, C., and Amato, G. (2022). An embedded toolset for human activity monitoring in critical environments. *Expert Systems with Applications*, 199:117125.



- Cafarelli, D., Ciampi, L., Vadicano, L., Gennaro, C., Berton, A., Paterni, M., Benvenuti, C., Passera, M., and Falchi, F. (2022). MOBDrone: A drone video dataset for man OverBoard rescue. In *Image Analysis and Processing – ICIAP 2022*, pages 633–644. Springer International Publishing.
- Chen, Y., Li, W., Chen, X., and Gool, L. V. (2019). Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Cheng, M., Cai, K., and Li, M. (2021). RWF-2000: An open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE.
- Ciampi, L., Amato, G., Falchi, F., Gennaro, C., and Rabitti, F. (2018). Counting vehicles with cameras. In *Proceedings of the 26th Italian Symposium on Advanced Database Systems, Castellana Marina (Taranto), Italy, June 24-27, 2018*, volume 2161 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Ciampi, L., Foszner, P., Messina, N., Staniszewski, M., Gennaro, C., Falchi, F., Serao, G., Cogiel, M., Golba, D., Szcześna, A., and Amato, G. (2022a). Bus violence: An open benchmark for video violence detection on public transport. *Sensors*, 22(21):8345.
- Ciampi, L., Gennaro, C., Carrara, F., Falchi, F., Vairo, C., and Amato, G. (2022b). Multi-camera vehicle counting using edge-AI. *Expert Systems with Applications*, 207:117929.
- Ciampi, L., Messina, N., Falchi, F., Gennaro, C., and Amato, G. (2020a). Virtual to real adaptation of pedestrian detectors. *Sensors*, 20(18):5250.
- Ciampi, L., Santiago, C., Costeira, J., Gennaro, C., and Amato, G. (2021). Domain adaptation for traffic density estimation. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications.
- Ciampi, L., Santiago, C., Costeira, J. P., Gennaro, C., and Amato, G. (2020b). Unsupervised vehicle counting via multiple camera domain adaptation. In *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020*, volume 2659 of *CEUR Workshop Proceedings*, pages 82–85. CEUR-WS.org.
- Csurka, G. (2017). Domain adaptation for visual applications: A comprehensive survey. *CoRR*, abs/1702.05374.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.
- Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). SlowFast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.
- Feichtenhofer, C., Pinz, A., and Wildes, R. P. (2016). Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3468–3476.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. S. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.
- Hanson, A., PNVR, K., Krishnagopal, S., and Davis, L. (2019). Bidirectional convolutional LSTM for the detection of violence in videos. In *Lecture Notes in Computer Science*, pages 280–295. Springer International Publishing.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Hong, W., Wang, Z., Yang, M., and Yuan, J. (2018). Conditional generative adversarial network for structured domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.
- Huo, X., Xie, L., Hu, H., Zhou, W., Li, H., and Tian, Q. (2022). Domain-agnostic prior for transfer semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Jin, Y., Wang, X., Long, M., and Wang, J. (2020). Minimum class confusion for versatile domain adaptation. In *Computer Vision – ECCV 2020*, pages 464–480. Springer International Publishing.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.
- Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211.
- Padamwar, B. (2020). Violence detection in surveillance video using computer vision techniques. *International Journal for Research in Applied Science and Engineering Technology*, 8(8):533–536.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.
- Pęszor, D., Staniszewski, M., and Wojciechowska, M. (2016). Facial reconstruction on the basis of video surveillance system for the purpose of suspect identification. In *Intelligent Information and Database Systems*, pages 467–476. Springer Berlin Heidelberg.
- Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., and Woo, W. (2015). Convolutional LSTM network: A

- machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Soliman, M. M., Kamal, M. H., Nashed, M. A. E.-M., Mostafa, Y. M., Chawky, B. S., and Khattab, D. (2019). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE.
- Spremolla, I. R., Antunes, M., Aouada, D., and Ottersten, B. (2016). RGB-d and thermal sensor fusion. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications.
- Staniszewski, M., Foszner, P., Kostorz, K., Michalczuk, A., Wereszczyński, K., Cogiel, M., Golba, D., Wojciechowski, K., and Polański, A. (2020). Application of crowd simulations in the evaluation of tracking algorithms. *Sensors*, 20(17):4960.
- Sudhakaran, S. and Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE.
- Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*. IEEE.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE.
- Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Zhang, Y., David, P., and Gong, B. (2017). Curriculum domain adaptation for semantic segmentation of urban scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.