SN Computer Science (2024) 5:834
https://doi.org/10.1007/s42979-024-03126-3

ORIGINAL RESEARCH

In the Wild Video Violence Detection: An Unsupervised Domain Adaptation Approach

Luca Ciampi[1] · Carlos Santiago[2] · Fabrizio Falchi[1] · Claudio Gennaro[1] · Giuseppe Amato[1]

Received: 11 October 2023 / Accepted: 9 July 2024

Abstract
This work addresses the challenge of video violence detection in data-scarce scenarios, focusing on bridging the domain gap that often hinders the performance of deep learning models when applied to unseen domains. We present a novel unsupervised domain adaptation (UDA) scheme designed to effectively mitigate this gap by combining supervised learning in the train (source) domain with unlabeled test (target) data. We employ single-image classification and multiple instance learning (MIL) to select frames with the highest classification scores, and, upon this, we exploit UDA techniques to adapt the model to unlabeled target domains. We perform an extensive experimental evaluation, using general-context data as the source domain and target domain datasets collected in specific environments, such as violent/non-violent actions in hockey matches and public transport. The results demonstrate that our UDA pipeline substantially enhances model performances, improving their generalization capabilities in novel scenarios without requiring additional labeled data.

Keywords Video violence classification · Deep learning for video understanding · Unsupervised learning · Video surveillance · Domain adaptation · Deep learning with scarce data

Introduction

In recent times, there has been an increasing enthusiasm to create computer vision applications and services that enhance the lives of citizens. The swift advancement of deep learning (DL) techniques, combined with the widespread presence of video surveillance cameras in modern cities, has given rise to intelligent applications designed for various purposes. These encompass face recognition [1, 2], crowd counting [3, 4], intelligent parking systems [5, 6], pedestrian detection and tracking [7, 8], among others. Such smart applications are now widely implemented worldwide, becoming pivotal in effectively managing public spaces and deterring criminal activities, and they are gradually replacing human supervision for monitoring.

However, state-of-the-art performances of DL algorithms are usually achieved through supervised learning, which relies on two key assumptions [9]. Firstly, it assumes the availability of extensive labeled datasets, which are crucial for accurately training the models. Secondly, it assumes that the training (or source) and test (or target) datasets are i.i.d., i.e., they are independent and have identical distributions. While abundant annotated data may be available for certain predefined domains, such as ImageNet [10] for image classification or COCO [11] for object detection, obtaining manual annotations for every specific target domain or task is often impractical and costly. Consequently, models are often applied to target domains not encountered in the existing labeled training data, and exhibit a drop in performance at inference time due to the domain shifts, i.e., the domain gap between source and target data distributions [12].

Unsupervised domain adaptation (UDA) provides one possible solution to tackle this issue. Its primary goal is to mitigate domain gaps by leveraging labeled data from the source domain and *unlabeled* data from the target domain. In essence, UDA techniques use annotated data from the source domain, along with non-annotated data from the target domain, which can be easily collected without requiring human effort for labeling. The key challenge here lies in automatically extracting knowledge from this latter data stream to narrow the gap between the two domains.

✉ Luca Ciampi
luca.ciampi@isti.cnr.it

[1] Institute of Information Science and Technologies of the National Research Council of Italy (ISTI-CNR), Pisa, Italy

[2] Instituto Superior Técnico (LARSyS/IST), Lisbon, Portugal

Published online: 27 August 2024

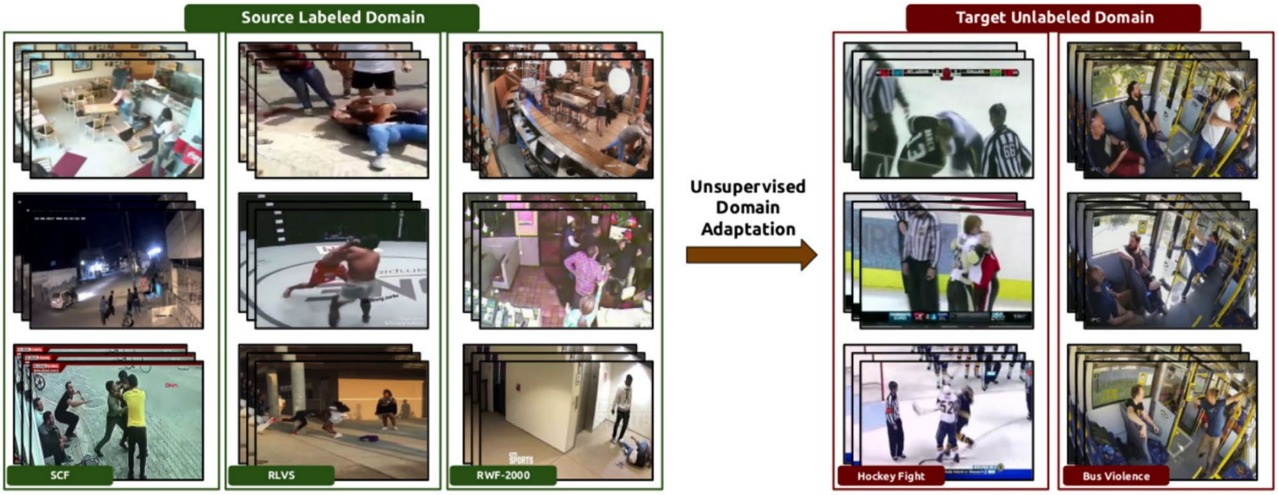


Fig. 1 **The considered scenario.** Our proposal introduces an unsupervised domain adaptation approach for video violence detection, aiming to bridge the domain gap that separates the source domain (depicted on the left) from the target domain (shown on the right). The source domain encompasses three datasets containing annotated videos portraying both violent and non-violent scenarios in broad contexts. In contrast, the target domain comprises two sets of *unlabeled* clips capturing instances of violent and non-violent actions within different and very specific scenarios, e.g., hockey matches and public transports

Specifically, the objective is to learn feature representations that are (i) discriminative for the primary learning task in the source domain and (ii) robust to the domain shift.

This paper focuses on detecting violent actions within trimmed videos, wherein the goal is to distinguish between violent and non-violent behaviors in clips that capture an exact action. Essentially, this task is a subcategory of human action recognition, with the objective of classifying video clips into binary categories encoding the presence or the absence of violent actions. Despite its significance through real-world scenarios involving the warning/prevention of criminal activities and public space management, this specific task has received relatively limited exploration compared to other action recognition tasks. Although some annotated datasets exist for general video violence detection, they suffer from shortcomings in terms of size and scenario diversity. Consequently, existing deep learning-based solutions trained on these datasets often exhibit decreased performance when applied to specific contexts, such as violence detection in public transports [13]. To tackle these limitations, we introduce an end-to-end deep learning-based UDA approach designed for improving performance in target scenarios where annotated data is scarce or absent. Our baseline takes inspiration from [14], a video violence detection technique that performs single-image classification by randomly selecting frames from the video. We improve this straightforward approach by exploiting a multiple instance learning (MIL) technique that instead considers the frame with the best classification score. Then, upon this baseline, we integrate a set of UDA techniques into the training process to automatically acquire knowledge from unlabeled data that pertains to the target domain. To the best of our knowledge, it is the first attempt at using a UDA schema for video violence detection. In our experiments, as the source domain, we utilize several annotated datasets of video violence detection in broader contexts. On the other hand, as the target domain, we exploit video clips in specific scenarios: the Hockey Fight dataset [15] containing violent/non-violent actions from hockey matches of the National Hockey League (NHL), and the recently introduced Bus Violence benchmark focusing on identifying violent behaviors occurring within a moving bus [13]. Figure 1 illustrates this experimental scenario. The outcomes indicate that our UDA pipeline substantially improves performance for the examined models. This suggests that these models exhibit enhanced generalization capabilities when adapting to this new scenario, all without the requirement of introducing new labels.

This research extends our previous work [16] by modifying the baseline architecture with a MIL approach, considering additional state-of-the-art models for performance comparison, and experimenting with an additional target dataset. The main contributions are summarized below:

- we present a novel UDA scheme for video violence detection to effectively reduce the domain gap between a labeled source dataset and an unlabeled target dataset;
- we perform an empirical assessment, using general violent/non-violent videos as the source domain and other clips designed for detecting violent behaviors in specific scenarios (such as hockey matches and public transports) as the target domain;

- the outcomes reveal that our UDA approach enhances the performance of the examined models, enabling them to achieve improved generalization in situations where labels are unavailable, accommodating novel scenarios.

The paper's structure is as follows: In Sect. "Related Works", we review related works. Section "Methodology" outlines our proposed methodology. Section "Experimental Analysis" presents the results from our experimental evaluation. Finally, Sect. "Conclusion" provides the paper's conclusion.

Related Works

Numerous methods and datasets are explicitly designed for video violence detection within the existing literature. Many of these approaches are tailored to analyze trimmed clips [14, 17–24], which capture precise actions, whether violent or non-violent. Consequently, this task falls under the umbrella of action recognition, where the goal is to classify videos, predicting the presence or absence of violent human behaviors. However, a few studies also delve into the realm of untrimmed videos [25–27]. In this case, the objective expands beyond action recognition to include action localization, which entails identifying the temporal boundaries of the actions. This distinction is also reflected in the datasets used for model training: trimmed video datasets typically provide annotations at the video level, while untrimmed video datasets require frame-level annotations. Our effort focuses explicitly on video violence detection in trimmed videos, where we introduce a UDA scheme to tackle a scenario characterized by scarce annotated data. In the subsequent sections, we will explore some collections of trimmed clips and prominent techniques in the literature. Finally, we will conclude this section by reviewing some existing UDA approaches.

Video Violence Detection Datasets

Over the past few years, several benchmarks comprising trimmed video clips suitable for video violence detection have been introduced. There exist numerous challenges concerning these collections of videos that consequently impact the performance of the video violence detection models, such as small amounts of data, video quality, and size. Some notable examples in the literature are (i) the Surveillance Camera Fight (SCF) [24] dataset, a collection of 300 videos from surveillance camera footage, 150 of which describe fight sequences and 150 depict non-fight scenes, (ii) the Real-Life Violence Situations (RLVS) [17] dataset, a set of 2000 video clips of violent/non-violent actions in general real-world scenarios, and (iii) the RWF-2000 [28] dataset which includes 2000 trimmed video clips from YouTube capturing violent/non-violent scenes from surveillance cameras. Furthermore, some other datasets focus on specific environments, such as the Hockey Fight [15] dataset containing 1000 labeled "violent" or "non-violent" trimmed clips of actions from hockey matches of the National Hockey League (NHL), and the Bus Violence [13] benchmark which instead includes 1400 videos of violent/non-violent scenes from several cameras located inside a moving bus, representing the first public dataset for human violence detection in public transport. We report all these accounted benchmarks in Table 1, where details and sample frames are also given. In this work, we point out the difficulties concerning the generalization capabilities of the DL-based techniques for video violence detection when trained with general-context data [17, 24, 28] and applied to specific scenarios [13, 15], providing a solution to mitigate this issue without using further annotations.

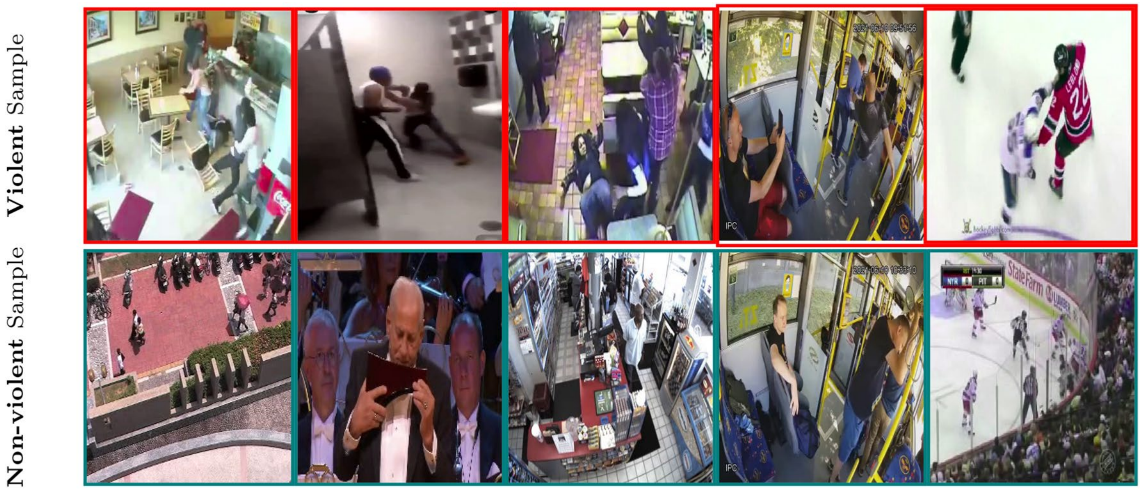
Video Violence Detection Approaches

Many of the existing video violence detection methods follow an architecture comprising a series of convolutional layers to extract spatial features, one or more long short memory (LSTM) layers [29] (or some variants of it) to encode the long-term frame level changes from a temporal perspective, and, finally, a sequence of fully connected layers for the final video classification. Some notable works are [17, 18] where the authors proposed a pre-trained VGG-16 on ImageNet as spatial feature extractor followed by LSTM as temporal feature extractor, or [19, 22, 23] that instead exploited ImageNet pre-trained AlexNet, VGG16, and ResNet50 as backbones, respectively, and convolutional LSTM (ConvLSTM) [30] for temporal feature encoding. Similarly, in [20], a spatio-temporal encoder built on a pre-trained VGG13 combined with bidirectional convolutional LSTM (BiConvLSTM) has been introduced, while the authors in [24] proposed a combination of Xception and bidirectional LSTM (BiLSTM) layers. It is also worth noting that these latter three works [19, 20, 23] do not use the raw RGB video stream as input, but they instead employ the frame-difference video stream, i.e., the difference of adjacent frames; frame-difference represents a computationally efficient alternative for optical flow, and it has been successfully exploited to capture short-term frame level changes. On the other hand, in [31], the authors used a different architecture relying on 3D convolutional layers [32] to handle both spatial and temporal dimensions, while in [14], the videos have been classified using single frames randomly sampled within the clips.

Alternatively, methods suitable for human action recognition can also be exploited. In this case, fine-tuning is required to classify videos into two classes: violence and non-violence. For instance, the ResNet 2+1D network [21] treats actions as spatio-temporal objects using a sort of 3D

Table 1 **Summary of datasets.** We report statistics and sample frames of some collections of trimmed clips in the literature suitable for video violence detection. SCF, RLVS, and RWF-2000 comprise general-context data, while Bus Violence and Hockey Fight focus on specific environments

	SCF [24]	*RLVS* [17]	*RWF-2000* [28]	*Bus Violence* [13]	*Hockey Fight* [15]
# clips	300	2000	2000	1400	1000
FPS	25	Variable	30	30	25
Length/Clip	2	3–7	5	–	1.6−1.96
Resolution	480x360	Variable	Variable	Variable	360x288
Year	2020	2019	2020	2022	2011
Balanced	✓	✓	✓	✓	✓



convolutional layer obtained by decomposing the convolutions into separate 2D spatial and 1D temporal filters [33]. Another widely-used model is SlowFast [34]. This architecture incorporates two branches. The first branch aims to capture semantic information through images or a few sparse frames, operating at a low frame rate. In contrast, the second branch captures fast-changing motion by operating at a higher refresh rate. Lastly, recent advancements have introduced architectures based on Transformer attention modules. An example is the Video Swin Transformer [35], which extends the sliding-window Transformers proposed for image processing [36] to the temporal axis. This extension achieves an excellent balance between efficiency and effectiveness.

Unsupervised Domain Adaptation Approaches

Traditional unsupervised domain adaptation (UDA) methods have predominantly focused on solving image classification problems by aligning features between two domains. Prominent examples of these methods include [37, 38]. However, extending these approaches to different applications is not straightforward, as underscored by [39]. Consequently, the existing literature provides a limited number of UDA techniques that are suitable for diverse tasks. Recent advancements have expanded the scope of UDA techniques to encompass areas like semantic segmentation [40, 41] and visual counting [42, 43]. This research introduces a UDA framework tailored specifically for video violence detection. To the best of our knowledge, this represents the first attempt to leverage UDA for this task.

Methodology

Background

In line with the notation introduced in [44, 45], we define a domain $\mathcal{D}$ consisting of two main components: a d-dimensional feature space $\mathcal{X} \subset \mathbb{R}^d$, and a marginal probability distribution $P(X)$, where $X = \{x_1, \dots, x_n\} \subset \mathcal{X}$ represents the set of feature samples. For a specific domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, we formulate a task $\mathcal{T}$, which is defined by a label space $\mathcal{Y}$ and the conditional probability distribution $P(Y|X)$, where $Y = \{y_1, \dots, y_n\} \subset \mathcal{Y}$ corresponds to the set of labels associated with X. In a supervised setting, $P(Y|X)$ can be learned from the provided feature-label pairs $\langle x_i, y_i \rangle$.

In the context of unsupervised domain adaptation, we are presented with two distinctive domains:

(i) a source domain $\mathcal{D}_S = \{\mathcal{X}_S, P(X_S)\}$, with $\mathcal{T}_S = \{\mathcal{Y}_S, P(Y_S|X_S)\}$,

(ii) a target domain $\mathcal{D}_T = \{\mathcal{X}_T, P(X_T)\}$, with $\mathcal{T}_T = \{\mathcal{Y}_T, P(Y_T|X_T)\}$,

where $\mathcal{Y}_T$ is unknown, meaning that we do not have any labels available for the target samples. Due to the inherent differences between the two domains, the distributions are assumed to be distinct, i.e., $P(X_S) \neq P(X_T)$ and $P(Y_S|X_S) \neq P(Y_T|X_T)$. The main goal of UDA is to train a model that exhibits decreased generalization error in the target domain, achieved through the effective reduction of domain discrepancy.

UDA for Video Violence Detection

In this work, the source domain $\mathcal{D}_S$ comprises a labeled collection of highly diverse videos depicting both violent and non-violent everyday-life situations. Here, $\mathcal{Y}_S = 0, 1$ indicates whether violent actions are absent or present in these clips, respectively. In contrast, the target domain $\mathcal{D}_T$ comprises an entirely separate set of videos lacking any annotations. These videos capture instances of violent or non-violent actions occurring in a distinct and unique context compared to the scenarios observed in the source domain. The main objective is to leverage knowledge from the unlabeled target domain in the training process, aiming to minimize the dissimilarity between the source and target domains. This adaptation enhances the model's capacity to generalize effectively to scenarios where annotations are not available.

Our approach relies on a deep learning-based model, trained end-to-end with an attached UDA module. A distinctive feature of our UDA scheme is that it is based on single-image classification. We transform the task of video classification into image classification, as scenes with violent actions can be distinguished from non-violent scenes by classifying a sampled image from the entire video clip [14, 16]. Specifically, we improve the idea introduced in [14, 16] where the authors picked up a frame from a clip at random, and we propose a multiple instance learning (MIL) technique that instead considers the frame with the best classification score. MIL is a type of weakly supervised learning in which training instances are organized into groups, referred to as "bags", and a single label is assigned to the entire bag [46]; in our context, bags are represented by the trimmed videos while instances are the frames composing the clips themselves. A straightforward MIL approach involves applying a max pooling operator against the classification scores associated with the instances, therefore obtaining a single score associated with the bag. Building on this baseline, we incorporate into the training pipeline two UDA techniques initially designed for image classification, feeding them with images sampled from the target domain to facilitate inter-domain knowledge transfer.

More in detail, we utilize several convolutional neural networks (CNNs) as backbones for extracting features, excluding the final classification layers. We substitute the last classification head with a binary classification layer, which provides the probability of the presence (or absence) of violent actions in the given video. Additionally, we introduce an extra linear layer followed by a ReLU activation function to transform the feature maps originating from the feature extractor into a fixed-dimensional representation. This fixed-dimensional feature map is subsequently input into a UDA module.

We have explored two distinct UDA approaches. The first approach is known as the Domain-Adversarial Neural Network - (DANN) [37], which involves a domain regressor engaged in an adversarial competition with the classifier. This method achieves UDA by connecting the domain classifier to the feature extractor through a gradient reversal layer. During training, this layer introduces an adversarial loss by reversing the gradient with a specific negative constant. Otherwise, the training process is standard for source examples, minimizing the label prediction loss, and includes domain classification loss for all samples. The adversarial loss ensures that the feature distributions between the two domains become as similar as possible, resulting in domain-invariant features. The second approach is referred to as the Minimum Class Confusion - (MCC) [38], a method that can be exploited as UDA without explicitly aligning domains. MCC is grounded in the concept of class confusion, where the classifier tends to confuse predictions between correct and ambiguous classes. Specifically, MCC operates on the class predictions made by the classifier for the target data, given the feature extractor. During training, MCC is optimized using standard backpropagation to reduce class confusion and enhance feature generalization.

Experimental Analysis

Experimental Setting

Experimental scenario. We exploited three datasets from existing literature as the source domain $\mathcal{D}_S$: Surveillance Camera Fight (SCF) [24], Real-Life Violence Situations (RLVS) [17], and RWF-2000 [28], which were previously mentioned in Sect. "Related Works". These datasets comprise annotated videos captured by stationary security cameras, featuring a diverse array of trimmed violent and non-violent scenes that span various real-life situations. In contrast, we adopted the Hockey Fight [15] dataset and the Bus Violence dataset [13] as the target domains $\mathcal{D}_T$. The former consists of trimmed clips from National Hockey League (NHL) matches, while the latter includes trimmed clips recorded within a moving bus, featuring actors simulating both violent and non-violent actions. These scenarios are notably more specific, involving instances of violence within the context of hockey matches or public transportation. More in detail, we divided the Hockey Fight and the Bus Violence datasets into two distinct splits: one has been used as the unlabeled set from which infer some domain-specific knowledge, while the second one served as testing grounds for evaluating the generalization capabilities of all the considered deep learning models.

We employed two popular CNNs, ResNet50 [47] and VGG16 [48], as the primary feature extraction backbones. We replaced their final classification head with a binary classification layer to adapt them for our video violence detection task. These networks served as our baseline models, as already did in our previous work [16] and in [14], and we used both without any UDA modules and as feature extractors and classifiers in our proposed UDA approaches. We also exploited these two feature extractors in our modified setting with the added MIL strategy, again, both with and without our UDA modules. Furthermore, to compare with existing literature, we considered other established methods tailored for video violence detection and video action recognition. Specifically, we leveraged architectures introduced in [17, 19, 20, 24], which utilize LSTM, BiLSTM, ConvLSTM, and BiConvLSTM as spatio-temporal encoders, and the network proposed in [31] that exploits 3D convolutional layers. We also considered popular video action classifiers, including the ResNet 2+1D network [21], the SlowFast [34] architecture, and the Video Swin Transformer [35]. More information about these models can be found in Sect. "Related Works" and their respective papers. We initiated these models with pre-trained versions from ImageNet [10] or Kinetics-400 [49] datasets, with no additional external data.

Performance metrics. Consistent with previous research on video violence detection, we employed accuracy as the key metric to assess the performance of the methods being examined. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP represents true positives, TN stands for true negatives, FP denotes false positives, and FN represents false negatives. For a more comprehensive comparison of the results obtained, we also incorporated additional metrics, including the F1-score, false alarm (FA), and missing alarm, which are defined as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{2}$$

$$FalseAlarm = \frac{FP}{TN + FP}, \tag{3}$$

$$MissingAlarm = \frac{FN}{TP + FN}, \tag{4}$$

where precision and recall are defined as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$, respectively. Finally, we incorporated the area under the receiver operating characteristics (ROC AUC) metric to account for the probabilities associated with the detections. It is computed by measuring the area under the curve obtained by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Implementation details and evaluation protocol. We implemented our models using PyTorch. Training and inference of all models are performed on an NVIDIA GeForce RTX4090. We used SGD for training, setting the starting learning rate to 0.005, the momentum at 0.9, and the weight decay to 0.001. The number of epochs is set at 60, 75, and 100 concerning the SCF, RLVS, and RWF-2000 datasets, respectively. We consistently employed a uniform data augmentation strategy throughout the training phase that included horizontal flipping with a probability of 0.5 and resizing to dimensions of 256×256 pixels.

To ensure the robustness of our results, we implemented the following evaluation protocol. Within each of the three selected source (training) domains, i.e., SCF, RLVS, and RWF-2000, we randomly partitioned the training and validation subsets three times. We then selected the best-performing model based on accuracy and tested it on the target (test) domain, i.e., the splits of Hockey Fight and Bus Violence benchmarks selected as the performance testing ground. Our reported results represent the mean and standard deviation of these three independent runs. We repeated the experiments five times instead of three times for some of the results obtained in our previous work [16] that showed a high standard deviation.

Results and Discussion

The results concerning the Hockey Fight target dataset are shown in Table 2. In general, all the models we examined demonstrate only moderate performance, highlighting the challenges in adapting their capabilities to detect violent actions in videos from the target domain effectively. Specifically, they particularly struggle when the considered source domain is the SCF dataset, while they exhibit better achievements with the RLVS data collection. However, our modified VGG16 architecture with our MIL-based approach together with the MCC UDA module stands out as the top performer in terms of the key metric, i.e., accuracy, in all the considered scenarios. More in detail, when compared to the same architecture without UDA, our proposed technique attains a gain of about 6%, 4%, and 3% in accuracy concerning the SCF, the RWF-2000, and the RLVS source domains, respectively. More importantly, it

Table 2 Performance evaluation over the Hockey Fight dataset [15]. We report the obtained results considering the Hockey Fight benchmark as the target domain and three sets of clips for video violence detection in general contexts

Model	*Accuracy*↑	*F1*↑	*FA*↓	*MA*↓	*ROC AUC*↑
(a) **Source Domain:** *Surveillance Camera Fight* [24]					
Hanson et al. [20] † $	0.54 ± 0.05	0.67 ± 0.01	0.85 ± 0.08	0.05 ± 0.07	0.62 ± 0.07
Sudhakaran et al. [19] †$	0.52 ± 0.02	0.67 ± 0.04	0.93 ± 0.05	0.02 ± 0.01	0.79 ± 0.05
Akti et al. [24] †	0.59 ± 0.01	0.61 ± 0.04	0.50 ± 0.08	0.29 ± 0.08	0.69 ± 0.08
Li et al. [31] *$	0.52 ± 0.01	0.67 ± 0.01	0.93 ± 0.03	0.01 ± 0.01	0.65 ± 0.07
Soliman et al. [17] †	0.67 ± 0.01	0.63 ± 0.07	0.21 ± 0.05	0.43 ± 0.06	0.78 ± 0.05
ResNet (2+1)D [21] *	0.62 ± 0.03	0.70 ± 0.01	0.64 ± 0.09	0.11 ± 0.03	0.77 ± 0.01
SlowFast [34] *	0.68 ± 0.07	0.63 ± 0.04	0.21 ± 0.05	0.41 ± 0.08	0.79 ± 0.02
VideoSwinTransformer [35] *	0.52 ± 0.03	0.67 ± 0.06	0.94 ± 0.08	0.01 ± 0.01	0.89 ± 0.04
ResNet50 [47] [14] † ‡	0.56 ± 0.01	0.62 ± 0.06	0.60 ± 0.06	0.26 ± 0.08	0.60 ± 0.04
VGG16 [48] [14] † ‡	0.54 ± 0.02	0.68 ± 0.01	0.89 ± 0.05	0.02 ± 0.01	0.69 ± 0.06
ResNet50-MIL †	0.70 ± 0.07	0.61 ± 0.06	0.12 ± 0.03	0.46 ± 0.07	0.84 ± 0.05
VGG16-MIL †	0.69 ± 0.06	0.75 ± 0.03	0.52 ± 0.07	0.07 ± 0.04	0.80 ± 0.05
ResNet50 † ‡ + DANN [37]	0.59 ± 0.02	0.64 ± 0.02	0.57 ± 0.04	0.25 ± 0.05	0.63 ± 0.03
ResNet50-MIL † + DANN [37]	0.72 ± 0.04	0.63 ± 0.03	0.13 ± 0.03	0.42 ± 0.02	0.85 ± 0.03
ResNet50 † ‡ + MCC [38]	0.61 ± 0.02	0.66 ± 0.01	0.56 ± 0.02	0.23 ± 0.02	0.66 ± 0.02
ResNet50-MIL † + MCC [38]	0.72 ± 0.02	0.62 ± 0.03	0.12 ± 0.02	0.41 ± 0.02	0.85 ± 0.02
VGG16 † ‡ + DANN [37]	0.56 ± 0.02	0.69 ± 0.01	0.86 ± 0.03	0.03 ± 0.01	0.70 ± 0.04
VGG16-MIL † + DANN [37]	0.71 ± 0.03	0.77 ± 0.02	0.50 ± 0.03	0.06 ± 0.02	0.82 ± 0.03
VGG16 † ‡ + MCC [38]	0.59 ± 0.02	0.71 ± 0.02	0.83 ± 0.02	0.03 ± 0.01	0.72 ± 0.03
VGG16-MIL † + MCC [38]	**0.73 ± 0.02**	0.78 ± 0.02	0.48 ± 0.02	0.05 ± 0.01	0.84 ± 0.02
(b) **Source Domain:** *RWF-2000* [28]					
Hanson et al. [20] †$	0.65 ± 0.03	0.64 ± 0.07	0.31 ± 0.04	0.38 ± 0.03	0.69 ± 0.05
Sudhakaran et al. [19] †$	0.63 ± 0.02	0.63 ± 0.06	0.40 ± 0.05	0.32 ± 0.03	0.72 ± 0.04
Akti et al. [24] †	0.70 ± 0.03	0.71 ± 0.03	0.30 ± 0.02	0.20 ± 0.02	0.79 ± 0.03
Li et al. [31] *$	0.69 ± 0.03	0.71 ± 0.03	0.21 ± 0.03	0.40 ± 0.03	0.78 ± 0.04
Soliman et al. [17] †	0.62 ± 0.04	0.46 ± 0.03	0.08 ± 0.05	0.66 ± 0.03	0.69 ± 0.03
ResNet (2+1)D [21] *	0.73 ± 0.01	0.69 ± 0.02	0.14 ± 0.01	0.39 ± 0.02	0.79 ± 0.01
SlowFast [34] *	0.63 ± 0.02	0.44 ± 0.03	0.01 ± 0.03	0.71 ± 0.02	0.69 ± 0.04
VideoSwinTransformer [35] *	0.63 ± 0.03	0.68 ± 0.02	0.51 ± 0.01	0.22 ± 0.03	0.69 ± 0.03
ResNet50 [47] [14] † ‡	0.65 ± 0.01	0.71 ± 0.02	0.54 ± 0.06	0.13 ± 0.08	0.75 ± 0.03
VGG16 [48] [14] † ‡	0.75 ± 0.01	0.79 ± 0.01	0.43 ± 0.02	0.05 ± 0.01	0.87 ± 0.01
ResNet50-MIL †	0.68 ± 0.03	0.73 ± 0.02	0.53 ± 0.04	0.11 ± 0.03	0.79 ± 0.02
VGG16-MIL †	0.77 ± 0.01	0.80 ± 0.01	0.41 ± 0.02	0.05 ± 0.04	0.89 ± 0.01
ResNet50 † ‡ + DANN [37]	0.66 ± 0.02	0.71 ± 0.01	0.52 ± 0.03	0.13 ± 0.04	0.77 ± 0.01
ResNet50-MIL † + DANN [37]	0.70 ± 0.02	0.74 ± 0.01	0.52 ± 0.02	0.10 ± 0.01	0.81 ± 0.02
ResNet50 † ‡ + MCC [38]	0.68 ± 0.02	0.73 ± 0.02	0.51 ± 0.02	0.11 ± 0.02	0.79 ± 0.02
ResNet50-MIL † + MCC [38]	0.71 ± 0.02	0.76 ± 0.01	0.50 ± 0.01	0.10 ± 0.01	0.82 ± 0.01
VGG16 † ‡ + DANN [37]	0.76 ± 0.01	0.81 ± 0.02	0.40 ± 0.02	0.04 ± 0.01	0.89 ± 0.02
VGG16-MIL † + DANN [37]	0.79 ± 0.01	0.81 ± 0.01	0.39 ± 0.01	0.05 ± 0.02	0.90 ± 0.01
VGG16 † ‡ + MCC [38]	0.78 ± 0.01	0.82 ± 0.02	0.38 ± 0.02	0.05 ± 0.01	0.90 ± 0.01
VGG16-MIL† + MCC [38]	**0.81 ± 0.02**	0.82 ± 0.01	0.38 ± 0.01	0.04 ± 0.01	0.91 ± 0.01
(c) **Source Domain:** *Real-life Violence Situations* [17]					
Hanson et al. [20] †$	0.67 ± 0.04	0.65 ± 0.01	0.27 ± 0.14	0.38 ± 0.05	0.75 ± 0.08
Sudhakaran et al. [19] †$	0.53 ± 0.02	0.67 ± 0.03	0.91 ± 0.03	0.03 ± 0.02	0.81 ± 0.04
Akti et al. [24] †	0.77 ± 0.01	0.78 ± 0.01	0.28 ± 0.09	0.16 ± 0.09	0.87 ± 0.02
Li et al. [31] *$	0.54 ± 0.06	0.49 ± 0.05	0.37 ± 0.06	0.54 ± 0.07	0.56 ± 0.06
Soliman et al. [17] †	0.59 ± 0.01	0.68 ± 0.01	0.71 ± 0.02	0.10 ± 0.04	0.69 ± 0.02

Table 2 (continued)

Model	*Accuracy*↑	*F1*↑	*FA*↓	*MA*↓	*ROC AUC*↑
ResNet (2+1)D [21] *	0.79 ± 0.02	0.78 ± 0.01	0.14 ± 0.06	0.26 ± 0.01	0.86 ± 0.03
SlowFast [34] *	0.61 ± 0.01	0.71 ± 0.01	0.75 ± 0.02	0.03 ± 0.01	0.92 ± 0.01
VideoSwinTransformer [35] *	0.57 ± 0.01	0.69 ± 0.01	0.84 ± 0.01	0.02 ± 0.01	0.92 ± 0.01
ResNet50 [47] [14] † ‡	0.75 ± 0.02	0.78 ± 0.01	0.38 ± 0.09	0.10 ± 0.04	0.86 ± 0.01
VGG16 [48] [14] † ‡	0.78 ± 0.01	0.81 ± 0.01	0.35 ± 0.01	0.07 ± 0.01	0.89 ± 0.01
ResNet50-MIL †	0.79 ± 0.05	0.76 ± 0.07	0.13 ± 0.05	0.28 ± 0.06	0.90 ± 0.03
VGG16-MIL †	0.82 ± 0.01	0.81 ± 0.01	0.13 ± 0.01	0.22 ± 0.02	0.89 ± 0.01
ResNet50 † ‡ + DANN [37]	0.75 ± 0.01	0.79 ± 0.01	0.37 ± 0.04	0.12 ± 0.02	0.87 ± 0.01
ResNet50-MIL † + DANN [37]	0.79 ± 0.04	0.77 ± 0.05	0.11 ± 0.02	0.23 ± 0.03	0.90 ± 0.02
ResNet50 † ‡ + MCC [38]	0.77 ± 0.01	0.80 ± 0.02	0.36 ± 0.02	0.11 ± 0.01	0.89 ± 0.02
ResNet50-MIL † + MCC [38]	0.80 ± 0.02	0.79 ± 0.02	0.11 ± 0.02	0.20 ± 0.02	0.90 ± 0.01
VGG16 † ‡ + DANN [37]	0.79 ± 0.02	0.82 ± 0.02	0.34 ± 0.01	0.07 ± 0.01	0.90 ± 0.02
VGG16-MIL † + DANN [37]	0.82 ± 0.01	0.82 ± 0.02	0.12 ± 0.02	0.20 ± 0.02	0.91 ± 0.02
VGG16 † ‡ + MCC [38]	0.80 ± 0.01	0.83 ± 0.02	0.32 ± 0.01	0.07 ± 0.01	0.90 ± 0.01
VGG16-MIL† + MCC [38]	**0.85 ± 0.01**	0.84 ± 0.01	0.12 ± 0.02	0.17 ± 0.02	0.92 ± 0.01

Best results in terms of accuracy are marked in bold

* Pretrained on Kinetics-400 [49]. † Pretrained on ImageNet [10]. $ Input modality: frame-difference. ‡ Input modality: single-frame

overcomes all the other considered state-of-the-art methods present in the literature, and it gains about 35%, 8%, and 9% of accuracy when compared with the single-image classification-based method proposed in [14] and that constitutes the baseline of our previous work [16].

Regarding the Bus Violence target dataset, we illustrate the results in Table 3. In general, all the models exhibit very poor performances, pointing out more challenges in this recently established scenario compared with the Hockey Fight dataset. However, even in this case, our UDA scheme can mitigate the difficulties arising from the generalization capabilities. In this setting, the best performer in terms of accuracy is our modified ResNet50 architecture with our MIL-based technique and the MCC UDA module. Specifically, compared with the same architecture without UDA, we gain about 9%, 5%, and 14% in accuracy concerning the SCF, the RWF-2000, and the RLVS source domains, respectively. Furthermore, it is worth noting that, even in this case, we overcome the other considered state-of-the-art techniques, gaining about 11%, 13%, and 16% of accuracy when compared against our baseline in [16].

Taking into account missing alarms, it's noticeable that our UDA module can increase the performance compared with the same architecture without UDA, considering both the target domains. MAs are particularly critical in video violence detection as they signify instances of violent actions that occurred but went undetected. Approaches that struggle with this metric represent a significant limitation for violence detection systems; therefore, this represents an added value to our proposal.

Conclusion

In this research, we addressed the challenge of video violence detection within the context of limited data availability. The current state of deep learning solutions heavily relies on abundant labeled data for effective supervised learning. However, these models tend to struggle when applied to new, previously unseen scenarios that were not part of their training data. Consequently, a model trained on one domain, referred to as the source, often experiences a significant performance decline when deployed in another domain, known as the target. To address this issue, we introduced an unsupervised domain adaptation (UDA) approach for identifying violent and non-violent actions within trimmed videos. Our method combines supervised learning in the source domain with the utilization of an unlabeled target dataset. This combination aims to reduce the domain shift between the two datasets. Our proposed solution is based on single-image classification, where a simple multiple instance learning (MIL) approach is responsible for taking frames from video clips having the maximum classification score. The feature representations extracted from the target images are passed through a UDA module, in charge of making them domain-indiscriminate by minimizing the shift between the domains. To the best of our knowledge, this is the first attempt to employ a UDA

Table 3 Performance evaluation over the Bus Violence dataset [13]. We report the obtained results considering the Bus Violence benchmark as the target domain and three sets of clips for video violence detection in general contexts

Model	*Accuracy*↑	*F1*↑	*FA*↓	*MA*↓	*ROC AUC*↑
(a) **Source Domain:** *Surveillance Camera Fight* [24]					
Hanson et al. [20] $^{\dagger\$}$	0.54 ± 0.02	0.19 ± 0.11	0.04 ± 0.03	0.89 ± 0.07	0.68 ± 0.02
Sudhakaran et al. [19] $^{\dagger\$}$	0.52 ± 0.01	0.27 ± 0.18	0.16 ± 0.17	0.79 ± 0.18	0.55 ± 0.02
Akti et al. [24] †	0.48 ± 0.03	0.31 ± 0.06	0.28 ± 0.07	0.73 ± 0.06	0.48 ± 0.03
Li et al. [31] $^{*\$}$	0.58 ± 0.01	0.66 ± 0.01	0.68 ± 0.02	0.11 ± 0.01	0.71 ± 0.01
Soliman et al. [17] †	0.50 ± 0.02	0.45 ± 0.03	0.40 ± 0.01	0.59 ± 0.03	0.52 ± 0.02
ResNet (2+1)D [21] *	0.52 ± 0.02	0.44 ± 0.06	0.52 ± 0.05	0.44 ± 0.07	0.54 ± 0.05
SlowFast [34] *	0.55 ± 0.03	0.40 ± 0.04	0.27 ± 0.05	0.62 ± 0.05	0.62 ± 0.03
VideoSwinTransformer [35] *	0.52 ± 0.01	0.65 ± 0.01	0.86 ± 0.01	0.10 ± 0.01	0.50 ± 0.01
ResNet50 [47] [14] †‡	0.54 ± 0.02	0.52 ± 0.06	0.44 ± 0.08	0.48 ± 0.08	0.55 ± 0.03
VGG16 [48] [14] †‡	0.51 ± 0.01	0.45 ± 0.07	0.39 ± 0.07	0.59 ± 0.08	0.51 ± 0.02
ResNet50-MIL †	0.55 ± 0.02	0.54 ± 0.02	0.40 ± 0.03	0.47 ± 0.03	0.60 ± 0.02
VGG16-MIL †	0.52 ± 0.01	0.46 ± 0.03	0.38 ± 0.01	0.57 ± 0.02	0.64 ± 0.01
ResNet50 †‡ + DANN [37]	0.55 ± 0.01	0.52 ± 0.04	0.44 ± 0.03	0.47 ± 0.06	0.56 ± 0.03
ResNet50-MIL † + DANN [37]	0.57 ± 0.02	0.54 ± 0.02	0.39 ± 0.02	0.46 ± 0.04	0.61 ± 0.02
ResNet50 †‡ + MCC [38]	0.58 ± 0.01	0.52 ± 0.03	0.45 ± 0.05	0.47 ± 0.04	0.63 ± 0.01
ResNet50-MIL† **+ MCC** [38]	**0.60 ± 0.01**	0.55 ± 0.02	0.38 ± 0.01	0.44 ± 0.01	0.68 ± 0.01
VGG16 †‡ + DANN [37]	0.53 ± 0.01	0.51 ± 0.04	0.49 ± 0.06	0.46 ± 0.05	0.51 ± 0.01
VGG16-MIL † + DANN [37]	0.54 ± 0.02	0.51 ± 0.02	0.39 ± 0.05	0.54 ± 0.04	0.53 ± 0.02
VGG16 †‡ + MCC [38]	0.53 ± 0.01	0.43 ± 0.01	0.28 ± 0.03	0.64 ± 0.01	0.52 ± 0.01
VGG16-MIL † + MCC [38]	0.56 ± 0.02	0.53 ± 0.02	0.37 ± 0.03	0.50 ± 0.03	0.57 ± 0.02
(b) **Source Domain:** *RWF-2000* [28]					
Hanson et al. [20] $^{\dagger\$}$	0.51 ± 0.01	0.07 ± 0.03	0.01 ± 0.01	0.96 ± 0.02	0.67 ± 0.05
Sudhakaran et al. [19] $^{\dagger\$}$	0.51 ± 0.01	0.08 ± 0.08	0.03 ± 0.03	0.95 ± 0.05	0.52 ± 0.02
Akti et al. [24] †	0.52 ± 0.02	0.53 ± 0.03	0.49 ± 0.07	0.46 ± 0.04	0.50 ± 0.02
Li et al. [31] $^{*\$}$	0.55 ± 0.02	0.19 ± 0.02	0.01 ± 0.01	0.89 ± 0.02	0.85 ± 0.04
Soliman et al. [17] †	0.50 ± 0.02	0.02 ± 0.02	0.01 ± 0.01	0.99 ± 0.02	0.52 ± 0.03
ResNet (2+1)D [21] *	0.53 ± 0.03	0.43 ± 0.05	0.29 ± 0.01	0.64 ± 0.05	0.54 ± 0.03
SlowFast [34] *	0.53 ± 0.03	0.40 ± 0.08	0.26 ± 0.08	0.67 ± 0.07	0.55 ± 0.03
VideoSwinTransformer [35] *	0.53 ± 0.02	0.52 ± 0.04	0.45 ± 0.04	0.49 ± 0.08	0.57 ± 0.03
ResNet50 [47] [14] †‡	0.54 ± 0.01	0.49 ± 0.04	0.34 ± 0.05	0.56 ± 0.06	0.58 ± 0.01
VGG16 [48] [14] †‡	0.54 ± 0.01	0.41 ± 0.03	0.25 ± 0.06	0.67 ± 0.04	0.54 ± 0.01
ResNet50-MIL †	0.56 ± 0.02	0.51 ± 0.04	0.32 ± 0.02	0.54 ± 0.02	0.63 ± 0.02
VGG16-MIL †	0.55 ± 0.01	0.42 ± 0.01	0.23 ± 0.01	0.65 ± 0.02	0.59 ± 0.03
ResNet50 †‡ + DANN [37]	0.55 ± 0.01	0.52 ± 0.01	0.33 ± 0.01	0.50 ± 0.01	0.60 ± 0.01
ResNet50-MIL † + DANN [37]	0.58 ± 0.01	0.52 ± 0.02	0.30 ± 0.02	0.52 ± 0.01	0.66 ± 0.02
ResNet50 †‡ + MCC [38]	0.56 ± 0.01	0.59 ± 0.02	0.49 ± 0.05	0.37 ± 0.05	0.62 ± 0.02
ResNet50-MIL† **+ MCC** [38]	**0.61 ± 0.02**	0.61 ± 0.01	0.47 ± 0.03	0.35 ± 0.02	0.67 ± 0.01
VGG16 †‡ + DANN [37]	0.55 ± 0.02	0.52 ± 0.03	0.24 ± 0.02	0.65 ± 0.02	0.55 ± 0.01
VGG16-MIL † + DANN [37]	0.57 ± 0.02	0.44 ± 0.02	0.21 ± 0.02	0.62 ± 0.03	0.58 ± 0.02
VGG16 †‡ + MCC [38]	0.55 ± 0.01	0.51 ± 0.02	0.20 ± 0.05	0.69 ± 0.06	0.55 ± 0.01
VGG16-MIL † + MCC [38]	0.60 ± 0.03	0.48 ± 0.02	0.20 ± 0.02	0.58 ± 0.02	0.62 ± 0.03
(c) **Source Domain:** *Real-life Violence Situations* [17]					
Hanson et al. [20] $^{\dagger\$}$	0.58 ± 0.02	0.49 ± 0.07	0.26 ± 0.07	0.57 ± 0.08	0.61 *pm* 0.03
Sudhakaran et al. [19] $^{\dagger\$}$	0.52 ± 0.01	0.45 ± 0.02	0.35 ± 0.04	0.61 ± 0.04	0.55 ± 0.02
Akti et al. [24] †	0.52 ± 0.01	0.39 ± 0.01	0.27 ± 0.04	0.68 ± 0.03	0.55 ± 0.01
Li et al. [31] $^{*\$}$	0.51 ± 0.03	0.43 ± 0.02	0.35 ± 0.05	0.62 ± 0.06	0.50 ± 0.03
Soliman et al. [17] †	0.55 ± 0.03	0.53 ± 0.03	0.43 ± 0.05	0.45 ± 0.03	0.58 ± 0.02

Table 3 (continued)

Model	*Accuracy*↑	*F1*↑	*FA*↓	*MA*↓	*ROC AUC*↑
ResNet (2+1)D [21] *	0.51 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.99 ± 0.01	0.57 ± 0.08
SlowFast [34] *	0.51 ± 0.01	0.02 ± 0.02	0.01 ± 0.01	0.99 ± 0.01	0.54 ± 0.04
VideoSwinTransformer [35] *	0.51 ± 0.03	0.30 ± 0.07	0.22 ± 0.06	0.76 ± 0.07	0.53 ± 0.03
ResNet50 [47] [14] † ‡	0.54 ± 0.01	0.49 ± 0.03	0.38 ± 0.08	0.54 ± 0.06	0.56 ± 0.01
VGG16 [48] [14] † ‡	0.53 ± 0.01	0.54 ± 0.02	0.35 ± 0.09	0.56 ± 0.08	0.58 ± 0.02
ResNet50-MIL †	0.55 ± 0.02	0.50 ± 0.02	0.37 ± 0.06	0.53 ± 0.05	0.58 ± 0.02
VGG16-MIL †	0.58 ± 0.04	0.56 ± 0.02	0.33 ± 0.04	0.55 ± 0.04	0.60 ± 0.02
ResNet50 † ‡ + DANN [37]	0.57 ± 0.01	0.49 ± 0.03	0.25 ± 0.04	0.59 ± 0.03	0.57 ± 0.02
ResNet50-MIL † + DANN [37]	0.57 ± 0.02	0.53 ± 0.02	0.37 ± 0.04	0.48 ± 0.03	0.61 ± 0.02
ResNet50 † ‡ + MCC [38]	0.61 ± 0.01	0.54 ± 0.09	0.32 ± 0.15	0.51 ± 0.13	0.61 ± 0.01
ResNet50-MIL† + MCC [38]	**0.63 ± 0.01**	0.59 ± 0.03	0.30 ± 0.03	0.47 ± 0.02	0.64 ± 0.02
VGG16 † ‡ + DANN [37]	0.54 ± 0.01	0.54 ± 0.03	0.40 ± 0.05	0.49 ± 0.03	0.58 ± 0.03
VGG16-MIL † + DANN [37]	0.59 ± 0.02	0.56 ± 0.02	0.32 ± 0.03	0.52 ± 0.02	0.61 ± 0.02
VGG16 † ‡ + MCC [38]	0.57 ± 0.01	0.54 ± 0.04	0.36 ± 0.08	0.50 ± 0.08	0.59 ± 0.01
VGG16-MIL † + MCC [38]	0.60 ± 0.02	0.58 ± 0.03	0.32 ± 0.02	0.50 ± 0.02	0.63 ± 0.02

Best results in terms of accuracy are marked in bold

* Pretrained on Kinetics-400 [49]. † Pretrained on ImageNet [10]. $^{\$}$ Input modality: frame-difference. ‡ Input modality: single-frame

framework for video violence detection. Our experiments used three source datasets comprising videos depicting violent and non-violent scenes in various general contexts. On the other hand, the target domains consisted of collections of clips capturing violent and non-violent actions in very specific environments, such as hockey matches and public transport. The obtained results indicate that our UDA scheme can enhance the generalization capabilities of the models considered by mitigating the domain gap.

Author Contributions Luca Ciampi: conceptualization, methodology, investigation, experimentation, writing-original draft preparation, writing-review and editing; Carlos Santiago: conceptualization, writing-review and editing; Fabrizio Falchi: writing-review and editing, funding acquisition; Claudio Gennaro: writing-review and editing, supervision; Giuseppe Amato: supervision, funding acquisition.

Funding Open access funding provided by ISTI - PISA within the CRUI-CARE Agreement. 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; SUN – Social and hUman ceNtered XR (EC, Horizon Europe No. 101092612). The study design, analysis, decision to publish, and manuscript preparation were conducted without any involvement from the funding sources.

Availability of Data and Materials All data is publicly available.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Research involving Human Participants and/or Animals Neither humans nor animals have been involved in this research.

Informed Consent Neither humans nor animals have been involved in this research.

References

1. Erakin ME, Demir U, Ekenel HK. On recognizing occluded faces in the wild. In: 2021 IEEE International Conference of the Biometrics Special Interest Group (BIOSIG) 2021; https://doi.org/10.1109/biosig52210.2021.9548293.
2. Li L, Mu X, Li S, Peng H. A review of face recognition technology. IEEE Access. 2020;8:139110–20. https://doi.org/10.1109/ACCESS.2020.3011028.

3. Avvenuti M, Bongiovanni M, Ciampi L, Falchi F, Gennaro C, Messina N. A spatio- temporal attentive network for video-based crowd counting. In: 2022 IEEE Symposium on Computers and Communications (ISCC), 2022;1–6. https://doi.org/10.1109/ISCC55528.2022.9913019
4. Di Benedetto M, Carrara F, Ciampi L, Falchi F, Gennaro C, Amato G. An embedded toolset for human activity monitoring in critical environments. Expert Syst Appl. 2022;199: 117125. https://doi.org/10.1016/j.eswa.2022.117125.
5. Ciampi L, Gennaro C, Carrara F, Falchi F, Vairo C, Amato G. Multi-camera vehicle counting using edge-ai. Expert Syst Appl. 2022;207: 117929. https://doi.org/10.1016/j.eswa.2022.117929.
6. Amato G, Ciampi L, Falchi F, Gennaro C. Counting vehicles with deep learning in onboard uav imagery. In: 2019 IEEE Symposium on Computers and Communications (ISCC). 2019;1–6. https://doi.org/10.1109/ISCC47284.2019.8969620.
7. Ciampi L, Messina N, Falchi F, Gennaro C, Amato G. Virtual to real adaptation of pedestrian detectors. Sensors. 2020;20(18):5250. https://doi.org/10.3390/s20185250.
8. Kim B, Yuvaraj N, SriPreethaa KR, Santhosh R, Sabari A. Enhanced pedestrian detection using optimized deep convolution neural network for smart building surveillance. Soft Comput. 2020;24(22):17081–92. https://doi.org/10.1007/s00500-020-04999-1.
9. Huo X, Xie L, Hu H, Zhou W, Li H, Tian Q. Domain-agnostic prior for transfer semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022;7065–75. https://doi.org/10.1109/CVPR52688.2022.00694.
10. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009;248–255. https://doi.org/10.1109/CVPR.2009.5206848
11. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014, pp. 740–755. Springer, Cham 2014;
12. Torralba A, Efros AA. Unbiased look at dataset bias. In: CVPR 2011, 2011;1521–1528. https://doi.org/10.1109/CVPR.2011.5995347
13. Ciampi L, Foszner P, Messina N, Staniszewski M, Gennaro C, Falchi F, Serao G, Cogiel M, Golba D, Szczesna A, Amato G. Bus violence: An open benchmark for video violence detection on public transport. Sensors. 2022;22(21):8345. https://doi.org/10.3390/s22218345.
14. Akti S, Ofli F, Imran M, Ekenel HK. Fight detection from still images in the wild. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022, 2022;550–559. https://doi.org/10.1109/WACVW54805.2022.00061 .
15. Bermejo Nievas E, Deniz Suarez O, Bueno García G, Sukthankar R. Violence detection in video using computer vision techniques. In: Computer Analysis of Images and Patterns, pp. 332–339. Springer, Berlin, Heidelberg 2011'. https://doi.org/10.1007/978-3-642-23678-5_39
16. Ciampi L, Santiago C, Costeira J, Falchi F, Gennaro C, Amato G. Unsupervised Domain Adaptation for Video Violence Detection in the Wild. In: Proceedings of the 3rd International Conference on Image Processing and Vision Engineering - IMPROVE, pp. 37–46. SciTePress, 2023; https://doi.org/10.5220/0011965300003497 . INSTICC
17. Soliman MM, Kamal MH, El-Massih Nashed MA, Mostafa YM, Chawky BS, Khattab D. Violence recognition from videos using deep learning techniques. In: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019;80–85. https://doi.org/10.1109/ICICIS46948.2019.9014714
18. Asad M, Yang Z, Khan Z, Yang J, He X. Feature fusion based deep spatiotemporal model for violence detection in videos. In: Neural Information Processing, pp. 405–417. Springer, Cham 2019;
19. Sudhakaran S, Lanz O. 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), pp. 2017;1–6. https://doi.org/10.1109/AVSS.2017.8078468
20. Hanson A, PNVR K, Krishnagopal S, Davis L. Bidirectional convolutional lstm for the detection of violence in videos. In: Computer Vision – ECCV 2018 Workshops, pp. 280–295. Springer, Cham 2019;. https://doi.org/10.1007/978-3-030-11012-3_24
21. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2018;6450–6459 https://doi.org/10.1109/CVPR.2018.00675
22. Sharma M, Baghel R. Video surveillance for violence detection using deep learning. In: Advances in Data Science and Management, pp. 411–420. Springer, Singapore 2020;. https://doi.org/10.1007/978-981-15-0978-0_40
23. Mugunga I, Dong J, Rigall E, Guo S, Madessa AH, Nawaz HS. A frame-based feature model for violence detection from surveillance cameras using convlstm network. In: 2021 6th International Conference on Image, Vision and Computing (ICIVC), pp. 2021; https://doi.org/10.1109/ICIVC52351.2021.9526948
24. Akti S, Tataroglu GA, Ekenel HK. Vision-based fight detection from surveillance cameras. In: IEEE Ninth International Conference on Image Processing Theory, Tools and Applications, IPTA 2019, Istanbul, Turkey, November 6-9, 2019, pp. 2019;1–6. https://doi.org/10.1109/IPTA.2019.8936070 .
25. Gnouma M, Ejbali R, Zaied M. A two-stream abnormal detection using a cascade of extreme learning machines and stacked auto encoder. Multimedia Tools and Applications. 2023. https://doi.org/10.1007/s11042-023-15060-2.
26. Ullah W, Hussain T, Ullah FUM, Lee MY, Baik SW. Transcnn: Hybrid cnn and transformer mechanism for surveillance anomaly detection. Eng Appl Artif Intell. 2023;123: 106173. https://doi.org/10.1016/j.engappai.2023.106173.
27. Wu J-C, Hsieh H-Y, Chen D-J, Fuh C-S, Liu T-L. Self-supervised sparse representation for video anomaly detection. In: Computer Vision – ECCV 2022, pp. 729–745. Springer, Cham 2022. https://doi.org/10.1007/978-3-031-19778-9_42
28. Cheng M, Cai K, Li M. Rwf-2000: An open large scale video database for violence detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2021;4183–4190. https://doi.org/10.1109/ICPR48806.2021.9412502
29. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
30. Shi X, Chen Z, Wang H, Yeung D, Wong W, Woo W. 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, pp. 2015;802–810. https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html
31. Li J, Jiang X, Sun T, Xu K. Efficient violence detection using 3d convolutional neural networks. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 2019;1–8. https://doi.org/10.1109/AVSS.2019.8909883
32. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: 2015

IEEE International Conference on Computer Vision (ICCV), pp. 2015;4489–4497. https://doi.org/10.1109/ICCV.2015.510

33. Feichtenhofer C, Pinz A, Wildes RP. 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, pp. 2016;3468–3476. https://proceedings.neurips.cc/paper/2016/hash/3e7e0224018ab3cf51abb96464d518cd-Abstract.html
34. Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2019;6201–6210. https://doi.org/10.1109/ICCV.2019.00630
35. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H. Video swin transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2022;3192–3201. https://doi.org/10.1109/CVPR52688.2022.00320
36. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2021;9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
37. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V. Domain-Adversarial Training of Neural Networks, pp. 189–209. Springer, Cham 2017; https://doi.org/10.1007/978-3-319-58347-1_10 .
38. Jin Y, Wang X, Long M, Wang J. Minimum class confusion for versatile domain adaptation. In: Computer Vision – ECCV 2020, pp. 464–480. Springer, Cham 2020;https://doi.org/10.1007/978-3-030-58589-1_28
39. Zhang Y, David P, Gong B. Curriculum domain adaptation for semantic segmentation of urban scenes. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2017;2039–2049. https://doi.org/10.1109/ICCV.2017.223
40. Hong W, Wang Z, Yang M, Yuan J. Conditional generative adversarial network for structured domain adaptation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2018;1335–1344 . https://doi.org/10.1109/CVPR.2018.00145
41. Chen Y, Li W, Chen X, Van Gool L. 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), pp. 2019;1841–1850. https://doi.org/10.1109/CVPR.2019.00194
42. Ciampi L, Santiago C, Costeira JP, Gennaro, C, Amato G. Domain Adaptation for Traffic Density Estimation. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP, pp. 2021;185–195. https://doi.org/10.5220/0010303401850195 . INSTICC
43. Ciampi L, Santiago C, Costeira JP, Gennaro C, Amato G. 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. CEUR Workshop Proceedings, vol. 2659, pp. 2020;82–85. https://ceur-ws.org/Vol-2659/ciampi.pdf
44. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59. https://doi.org/10.1109/TKDE.2009.191.
45. Csurka G. In: Csurka, G. (ed.) A Comprehensive Survey on Domain Adaptation for Visual Applications, pp. 1–35. Springer, Cham 2017. https://doi.org/10.1007/978-3-319-58347-1_1 .
46. Carbonneau M-A, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recogn. 2018;77:329–53. https://doi.org/10.1016/j.patcog.2017.10.009.
47. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016. https://doi.org/10.1109/cvpr.2016.90
48. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings 2015.
49. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A. The kinetics human action video dataset. CoRR arXiv:abs/1705.06950 2017.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.