Domain Adaptation for Violence Detection in Camera Surveillance

Tim Lindenau (2023-82785) Stefan Djordje Tomić (2023-82383) Sangyoon Yu (2022-23102)
Jiayue Wang (2022-21806)

Abstract

Closed-circuit television (CCTV) has demonstrated its effectiveness in crime prevention, especially when it is actively monitored. However, the deployment of automatic violence detection for real-life CCTV footage encounters challenges due to the lack of camera surveillance datasets containing crimes and the domain shift caused by the installation environment and camera quality. To enhance violence detection capabilities, we explore an unsupervised domain adaptation approach to address domain discrepancy and employ a self-supervised learning approach to refine feature representation. In addition, we propose a novel framework to combine the two approaches with the same optimization objective. The proposed framework achieves an impressive improvement of 11% in the UCF-fighting dataset.

(a) (b) (c) (d)

Figure 1: Examples in the source domain ((a) and (b) from the hockey fighting dataset) and the target domain ((c) and (d) from the UCF-Crime dataset). (a) and (b) share very similar ice hockey rink backgrounds. The violent action in (c) is partially obscured by the car door. (d) is in black and white.

1. Introduction

Camera surveillance plays an important role in monitoring criminal activities and has been proven to be effective in crime prevention (Piza et al., 2019). With the widespread deployment of CCTV systems in public places, manually monitoring surveillance cameras in real-time becomes increasingly labor-intensive. As a result, there is a critical demand for lightweight violence detection systems that can be deployed on edge devices.

Compared with general video recognition, violence detection encounters greater challenges because of the lack of high-quality labeled data and large domain shifts. CCTV video clips often suffer from poor quality, such as low resolution or jitters, and their publication is subject to privacy concerns. In spite of the prevalence of surveillance cameras, there are remarkably few CCTV datasets containing fighting. This is primarily due to the low probability and diverse forms of violent incidents, making manually detecting and labeling violence in CCTV videos both labor-intensive and time-consuming. Furthermore, variations in deployment environments and camera quality of CCTV result in significant domain shifts between different surveillance cameras. Even if a model is well-trained for one setup, its performance may significantly deteriorate in others.

Our work endeavors to overcome domain shifts and improve the violence detection capability in CCTV footage. Since violence detection in CCTV suffers from a lack of available data, related datasets with violent scenes from movies or sports games are employed as the source domain. The labeled data in the source domain are then integrated with unlabeled, easily available CCTV footage to improve the performance of violence detection in CCTV videos (i.e., the target domain). As depicted in Figure 1, the videos in the source domain have very similar ice hockey rink backgrounds, while the videos in the target domain suffer from diverse backgrounds and image quality.

In this work, we investigate two separate approaches to enhance violence detection. First, we employ the discrepancy-based unsupervised domain adaptation (UDA) approach (Long et al., 2015; Vaserstein, 1969) with the goal of improving target domain performance by minimizing the domain discrepancy between the source and target domain. Second, a more informative feature representation for both domains is learned through Maximum Entropy Cod-

ing (Liu et al., 2022) in a self-supervised learning (SSL) manner. Since the strengths of UDA and SSL can complement the weaknesses of each other, we synergistically combine both approaches with the same optimization objective and achieve significant performance improvements. The contributions of this work are summarized as follows:

- We elucidate the inherent challenges of violence detection in surveillance cameras and target the unsupervised domain adaptation for CCTV footage.
- We adapt UDA and SSL approaches which are normally used in the image domain and adapt them to the video domain. In addition, we propose a novel framework by combining the two approaches with the same optimization objective.
- The UDA approach achieves an improvement of 7% and the SSL approach leads to an 8% enhancement. The combined approach further boosts performance by an additional 3%.

2. Related Work

2.1. CCTV Violence Detection

Although a wide range of approaches have been employed for violence detection in CCTV, to the best of our knowledge there has been no attempt at using self-supervised learning and but a single attempt utilizing UDA. Ciampi et al. (2023) successfully used UDA in order to improve the performance of violence detection in CCTV. They utilize Domain-Adversarial Neural Network and Minimum Class Confusion for UDA, though their model only considers still frames, thereby forfeiting all temporal information.

Most approaches in violence detection in CCTV use CNNs for feature extraction, though the more traditional approaches include the use of hand-crafted features such as Scale-Invariant Feature Transform (SIFT) or Speeded-up robust features (SURF) (Ullah et al., 2023). CNNs are often combined with LSTMs. Traditional classifiers, such as SVMs, Naïve Bayes, kNN and Decision Trees are often used.

In the broader field of anomaly detection the approaches are more diverse. Reconstruction-based methods such as GANs or autoencoders have been attempted, in addition to the prediction-based ViT (Şengönül et al., 2023).

Recent surveys have identified the need to develop small models capable of making real-time inferences as well as being deployed on resource-constrained environments (Ullah et al., 2023; Şengönül et al., 2023). This way, models could be deployed near edge devices, saving both bandwidth and power.

2.2. MoViNets

MoViNets is a family of lightweight and highly efficient video networks which have achieved state-of-the-art performance in several action-recognition benchmarks (Kondratyuk et al., 2021). The networks are pretrained on Kinetics-600, a large dataset of diverse human activity. MoViNets use fixed-size stream buffers to process frames, thus decoupling memory from clip duration. This allows for the processing of arbitrary-length videos in a memory-efficient manner, whilst utilizing temporal relations between frames beyond mere aggregation. The stream buffer enables the models to perform online inference.

2.3. Self-Supervised Learning

Supervised learning algorithms show great performance in various domains when trained on a large-scale dataset. However, data acquired from the real-world have abundant unlabeled data with a limited number of labeled data. Labeling data is costly and time-consuming. To address these limitations of supervised learning, various approaches have been introduced, such as active learning, unsupervised learning, semi-supervised learning, and self-supervised learning (SSL). Especially, SSL aims to learn rich features from large unlabeled data without human annotations. Normally, SSL is used for pretraining the network (encoder part), and then the model is transferred to downstream tasks via fine-tuning.

The most core aspect in SSL methods is which pretext task to choose. A pretext task is a hand-crafted learning objective in which supervision objectives can be achieved from the data itself. Pretext tasks can be classified into two approaches: 1) contrastive, and 2) non-contrastive. Contrastive methods, such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020), prevent a trivial solution by minimizing distance between the augmented views but maximizing distance between different images. Differently, in non-contrastive methods, such as SimSiam (Chen & He, 2021) and BYOL (Grill et al., 2020), use two branches of the model with shared weights (or momentum weight updates) to prevent collapse.

Maximum Entropy Coding Maximum Entropy Coding (MEC) (Liu et al., 2022) is an SSL method inspired from the maximum entropy principle in information theory. Its structure is similar to SimSiam (Chen & He, 2021), using two branches of the model with momentum-updated weights to prevent model collapse. Also, it maximizes the entropy in the representation space, extracting various information, by learning generalizable features from the domain. It is known that generalizable features learned with MEC are robustly transferred to different domain downstream tasks.

2.4. Unsupervised Domain Adaptation

In machine learning, it is generally assumed that training- and test data are independent and drawn from the same distribution (independent and identically distributed). Oftentimes this assumption does not hold leading to a significant drop in performance when a model trained on one domain (source) is tested on another domain (target).

Finetuning is one simple approach to improve target domain performance when labeled target domain data is available. In many real-world scenarios, this assumption is, however, too limiting as manually collecting and labeling training data for each new target domain is often difficult, sometimes even impossible. Unsupervised Domain Adaptation (UDA) is a different approach to improve target-domain performance. Requiring only labeled training data from the source domain, UDA improves target domain performance by minimizing the domain discrepancy between the source and target domain. The idea behind this is very comprehensible. Assuming that 1) high source-domain performance can be achieved through supervised learning and 2) the feature distributions of the source and target domain are similar, likely a classifier head trained only on the source domain will also perform well for the target domain.

For achieving the necessary domain alignment literature mostly knows two approaches. The first class – adversarial approaches – works by introducing an additional discriminator network tasked to distinguish source and target samples just by their features. The model is trained to make this task as difficult as possible, thus guaranteeing good feature alignment. A non-exhaustive list of relevant previous literature is Chen et al. (2019), Jamal et al. (2018), and Chen et al. (2022).

However, while adversarial approaches can be very effective they are often unstable and notoriously difficult to train. Discrepancy-based UDA methods overcome the stability issue by measuring domain discrepancy explicitly using mathematical metrics instead of solving a difficult min-max problem involving a discriminator. Common metrics used in discrepancy-based approaches are the Maximum Mean Discrepancy (MMD) (Long et al., 2015) or Wasserstein loss (Vaserstein, 1969).

3. Methods

This section introduces our two approaches to improving target domain performance, the first one being based on UDA and the second one SSL. As we later show in the experiments both approaches are complementary. Therefore, we end this section by introducing a completely new method, combining both approaches to gain significant performance improvements.

3.1. Unsupervised Domain Adaptation

Our chosen approach falls within the previously introduced class of discrepancy-based UDA methods. This particular class was selected over adversarial methods due to its inherently greater stability, offering a more reliable foundation for our analysis.

Figure 2 illustrates the architecture of our approach. During training the objective is twofold. First, The network is optimized for good classification performance on the source domain using the labeled videos and computing a classification loss $\mathcal{L}_{CL}$ such as the cross-entropy loss. Second, The network is optimized for domain alignment by comparing the features of the target and source domain using the domain alignment loss $\mathcal{L}_{DA}$. The overall training loss is a convex combination of both losses weighted by the hyperparameter λ, i.e.,

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{CL} + (1 - \lambda) \cdot \mathcal{L}_{DA}, \quad \lambda \in [0, 1]. \tag{1}$$

For the domain alignment loss $\mathcal{L}_{DA}$, multiple different losses are possible. We investigated our approach with two of the most common loss functions i.e., Maximum-Mean-Discrepancy (MMD) and Wasserstein Loss.

Maximum Mean Discrepancy: The MMD is a statistical measure used primarily in machine learning to compare the similarity between two distributions without making any assumptions about the form of those distributions. In more detail, MMD calculates the distance between the mean embeddings of features in a reproducing kernel hilbert space (RKHS) derived from two different data samples. As it is common we chose a Gaussian kernel function to induce the RKHS.

Wasserstein Distance: The Wasserstein Distance, also commonly known as the Earth Mover's Distance (EMD), offers a robust and geometrically intuitive way to quantify the dissimilarity between two probability distributions. Unlike the MMD, which computes the distance in terms of mean embeddings in an RKHS, the Wasserstein distance conceptualizes the distributions as two distinct masses of earth and measures the minimum 'work' required to transform one distribution into the other. This 'work' is quantified as the amount of distribution mass that needs to be moved, multiplied by the distance it is moved. The choice of weight function in the Wasserstein distance is crucial and we chose to utilize the squared loss, commonly also known as 2-Wasserstein distance.

3.2. Self-Supervised Learning

Our second approach, self-supervised learning (SSL), is used to train the network on the target domain. Since we have unlabeled target domain data, we utilize the SSL approach to train the feature extractor on the target domain

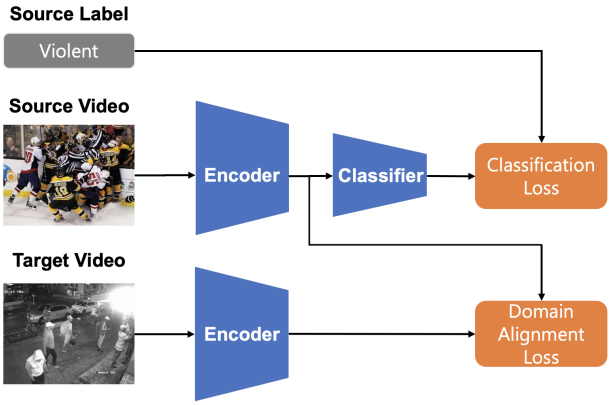

Figure 2: Architecture for discrepancy-based UDA. Target and source domain videos are fed to the network. The network is optimized for classification performance using the labeled source videos. Simultaneously, the network is optimized for domain alignment by comparing the features of the target and the source domain.

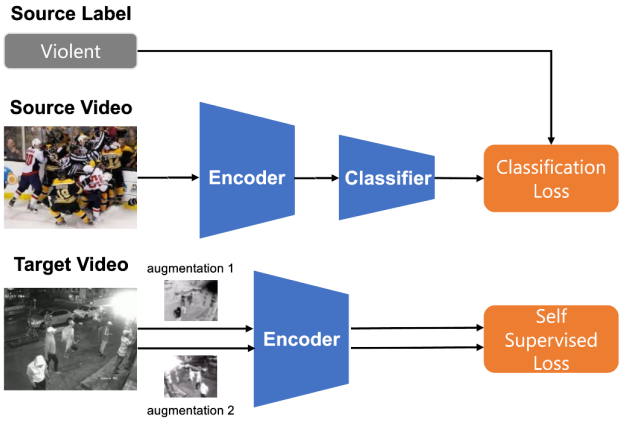

Figure 3: Architecture for MEC-based SSL. Target and source domain videos are fed into the network. Similar to UDA, first, the network is optimized for classification performance using the labeled source videos. Simultaneously, the network is optimized for self-supervised learning loss for augmented target domain samples.

to extract generalizable, rich features. Classifier layers are trained to map these features to labels using the labeled source domain data. Even though a domain gap exists, the features trained are highly generalizable on the target domain, so, even with error, the mapping of features and labels in the source domain can be applied to the target domain.

Figure 3 illustrates the architecture of our approach. SSL is mostly used as pretraining and transferred to a downstream task after pretraining in a two-stage manner. However, we simultaneously apply source domain classification loss $\mathcal{L}_{CL}$ and target domain SSL loss $\mathcal{L}_{SSL}$ together, similar to how we applied UDA. Thus, the training objective is twofold. This design choice is due to the reason that the downstream task and SSL pretraining are done on different domains, so when trained solely on the target domain and fine-tuned on the source domain, the domain gap disturbs the knowledge transfer. Similar to UDA, the overall training loss is a convex combination of two losses by the hyperparameter λ, i.e.,

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{CL} + (1 - \lambda) \cdot \mathcal{L}_{SSL}, \quad \lambda \in [0, 1]. \tag{2}$$

There are various SSL methods that can be applied. Especially, two approaches, contrastive and non-contrastive learning, both have their own advantages. Even though contrastive learning methods show great performance, the need for a large batch size is difficult to apply in the video domain due to memory issues. In this work, we applied Maximum-Entropy Coding (MEC) (Liu et al., 2022), a method that combines both contrastive and non-contrastive learning and can be considered a generalized method for SSL.

Maximum Entropy Coding: MEC enables the network to learn generalizable representations via maximizing the information in the representation space. Famous SSL methods such as SimSiam (Chen & He, 2021) and SimCLR (Chen et al., 2020) can be expressed into an MEC objective. MEC has two networks updated, one with the MEC loss and one with a momentum update similar to SimSiam. Two different types of augmented videos are fed into the network, and the entropy between different videos is maximized, while between the same video but different augmentations is minimized.

3.3. UDA and SSL are Complementary

The interesting part of the UDA and SSL approaches that we applied to this problem is that they are complementary to each other.

UDA: *Strengths.* Can align source and target domain representations to reduce the domain gap, and representation can be applied in both domains. *Weaknesses.* Difficult to learn rich features and might only learn simple features.

SSL: *Strengths.* Can learn rich and generalizable features. *Weaknesses.* The domain gap between source and target domains can lead to classifiers learned on the source domain not being applicable to the target domain.

Since the two approaches are complementary, we propose combining the two approaches together to further improve performance. We combine the losses by simply adding all into a single loss weighted by hyperparameters, i.e.,

$$\mathcal{L} = \lambda_{CL} \cdot \mathcal{L}_{CL} + \lambda_{UDA} \cdot \mathcal{L}_{UDA} + \lambda_{SSL} \cdot \mathcal{L}_{SSL}. \tag{3}$$

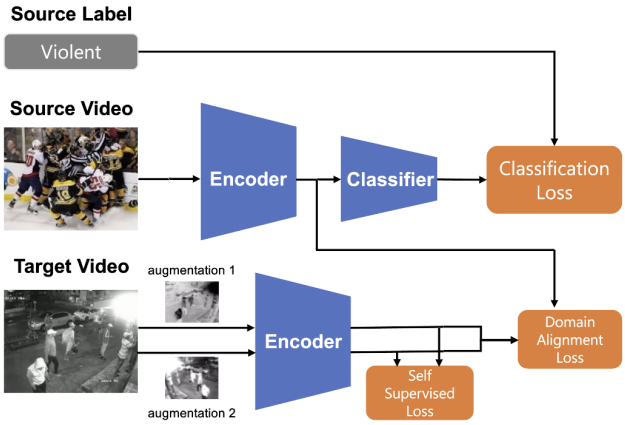

Figure 4: Architecture combining UDA and SSL approaches. Target and source domain videos are fed into the network. Both UDA and SSL losses are applied.

4. Experiments

4.1. Datasets

We utilize the hockey fight dataset (Bermejo Nievas et al., 2011) as the source domain. It consists of 1000 video clips collected from hockey games of the National Hockey League. The "fight" and "non-fight" classes each account for half of the clips. Each clip contains 50 frames with a resolution of 720×576 pixels and is manually labeled as "fight" or "non-fight". The dataset is randomly split into training, validation, and test sets with a ratio of 8:1:1. All the video clips have very similar backgrounds and stable image quality.

To adapt the model to the surveillance scenario, we employ the UCF-Crime Dataset (Sultani et al., 2019) as the target domain. The UCF-Crime dataset comprises long untrimmed surveillance videos with normal activities and 13 categories of abnormal activity such as "fighting", "stealing", and "road accidents". In this study, we mainly focus on the category "fighting", as strictly all videos contain interpersonal violence. We construct a subset of the UCF-Crime dataset as UCF-fighting with 50 videos of "fighting" and 50 videos of "normal" activity. It's worth noting that the videos suffer from various camera perspectives, low image quality, occasional occlusions, and transient violence in a long video.

4.2. Experimental Settings

In this study, we employ the MoViNet-A1-Base model as the encoder. It is pre-trained on Kinetics-600 (Carreira & Zisserman, 2017), achieving top-1 accuracy of 76.69% and top-5 accuracy of 93.40%. Adam algorithm is used as an optimizer with initial learning to be $5e-5$ and halved every two epochs. All the models are trained for 5 epochs with a batch size of 32. The videos are split into clips with 16 frames and are augmented through horizontal flipping and random cropping during training. Besides, videos in the UCF-fighting dataset are untrimmed, the videos are resampled with a frame rate to be 5.

4.3. Unsupervised Domain Adaptation

The results for UDA are shown in the upper half of Table 1, where we show the performance of a model pre-trained only on the source domain and the performance after applying UDA using either the MMD loss or Wasserstein loss on the UCF dataset. For the weighting factor λ we used a value of 0.6 in both cases. The results demonstrate that Wasserstein-distanced-based UDA is effective and consistently improves performance across train, target, and validation sets. MMD-based UDA, on the other hand, was not as effective, and while performance on the train and validation set could be improved, performance on the test-set dropped significantly. This drop in performance indicates that the MMD-based approach did not manage to create an alignment that generalizes well but rather overfitted the train and validation set. Our observation that the Wasserstein distance is more effective in aligning domains aligns with and has already been made within previous work. For example, in generative neural networks, the Wasserstein distance is generally preferred because it is more effective at comparing distributions with little overlap and provides more stable gradients, thus simplifying optimization (Arjovsky et al., 2017).

Focusing back on Wasserstein-based UDA, the careful reader might have noticed that performance improvements on the test, and validation set, were significantly higher than on the test set, i.e., 11% and 15% on the train and validation set compared to 7% on the test set. From this observation, one might conclude that our approach overfits the train and validation set. However, we do not believe that is the case as the test set already started with a better performance on the source-only trained baseline. Therefore, instead of overfitting, our UDA method equalizes the performance difference across dataset splits with all splits performing similarly well after aligning domains, thus further proving the effectiveness of our approach.

As a final ablation, we investigated the effect of the layout of the latent space on model performance. The latent space of the original MoViNet has the shape of $w \times h \times t$, where t is the temporal dimension. Before the classification head, the space is then average-pooled across time to dimensions $w \times h$. When optimizing for the best hyperparameters, we investigated whether aligning the latent space before or after temporal pooling is more effective, with the conclusion that aligning the average pooled space is significantly more effective. We hypothesize this is the case for two distinct reasons. First, removing the temporal dimension reduces

Table 1: Accuracy of different methods across various experiments. The best results are marked in bold.

Method	Source Val	Source Test	Target Train	Target Val	Target Test
Baseline (Source only)	**0.99**	**0.99**	0.59	0.57	0.70
UDA (Wasserstein Distance)	0.97	0.98	**0.70**	0.72	0.77
UDA (MMD)	0.96	0.98	0.66	0.71	0.58
SSL (MEC, two stage learning)	**0.99**	**0.99**	0.60	0.63	0.66
SSL (MEC, combination loss)	0.97	0.97	0.66	0.76	0.78
UDA (Wasserstein Distance) + SSL (MEC)	0.93	0.95	**0.70**	**0.79**	**0.81**

the complexity of the latent space and therefore makes it easier to align domains. Second, we believe as all videos in our datasets are uniformly violent or non-violent, the loss of temporal information does not come at too much of a cost and is easily traded off by the improved alignment.

4.4. Self-Supervised Learning

The results for SSL are shown in Table 1. Maximum entropy coding was applied, with $\epsilon_{MEC} = 16$. As previously explained in Section 3.2, we implemented two types of architecture: two-stage learning (SSL pretraining, CL fine-tuning) and single-stage learning with a combination loss. The results are very different; two-stage learning did not improve performance on the target domain. This is because the domain gap makes the extractor fit only the source domain at fine-tuning. We can see this by the source domain performance being 99%, while the target domain performance is low. However, when the SSL and classification loss are combined and trained simultaneously, the representations are able to learn rich features from the target domain that are also good for the source domain classification task at the same time. As a result, the target domain performance improved across all target domain data splits, showing an 8% increase in the target test dataset.

4.5. UDA + SSL

We combined UDA and SSL, with both the best-performing options: Wasserstein distance and combination loss. As explained in Section 3.3, the strengths and weaknesses of the two methods were complemented when the two methods were combined. As we assumed, combining both UDA and SSL further improved performance over UDA and SSL independently applied. An additional 3-4% performance increase was observed, showing that each method is complementary, and the weaknesses we assumed for each method were accurate.

5. Future Work

Compared to static images, videos incorporate an additional temporal dimension, offering abundant information for video understanding. In our approach, the temporal information is extracted through 3D convolutional neural networks in the MoViNet backbone. While UDA and SSL concentrate on enhancing the latent representation space, they do not extensively emphasize the temporal features. We hypothesize that strengthening the utilization of temporal information in UDA and SSL could further benefit anomaly detection. We suggest this as a direction for future work to enhance our existing findings.

6. Conclusion

This work illuminates the inherent challenges of violence detection in surveillance videos. We employ UDA and SSL approaches to address the lack of high-quality labeled camera surveillance data and the domain shift between fighting in sports games and surveillance videos. The UDA approach tackles domain discrepancy, while SSL enhances feature representation. Additionally, we propose a novel framework to integrate the two approaches with the same optimization objective. It effectively overcomes the domain shifts and significantly improves performance without using labeled data in the target domain.

6.1. Roles of Team Members

- *Tim*: 1) Research, especially focusing on unsupervised domain adaptation 2) Implementation of UDA 3) code refactor to align all experimental setups
- *Stefan*: 1) Preliminary research on project topic 2) Processing UCF-Crime dataset 3) Implementation of training/testing setup in Tensorflow
- *Sangyoon*: 1) Proposed and implemented SSL using MEC and combined SSL with UDA into a single loss and tested the results. 2) Recorded final presentation.
- *Jiayue*: 1) Building up the PyTorch-based training and evaluation codes. 2) Implementation of data augmentation and cross-dataset evaluation. 3) Attempt to design an attention module to enhance the utilization of temporal information based on (Woo et al., 2018; Chen et al., 2019), but the effect was not very stable.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., and Sukthankar, R. Violence detection in video using computer vision techniques. In Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., and Kropatsch, W. (eds.), *Computer Analysis of Images and Patterns*, pp. 332–339, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23678-5.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Chen, M.-H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., and Zheng, J. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6321–6330, 2019.

Chen, P., Gao, Y., and Ma, A. J. Multi-level attentive adversarial learning with temporal dilation for unsupervised video domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1259–1268, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Ciampi, L., Santiago, C., Costeira, J., Falchi, F., Gennaro, C., and 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*, pp. 37–46, Prague, Czech Republic, 2023. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-642-2. doi: 10.5220/0011965300003497.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Jamal, A., Namboodiri, V. P., Deodhare, D., and Venkatesh, K. Deep domain adaptation in action space. In *BMVC*, volume 2, pp. 5, 2018.

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., and Gong, B. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16020–16030, June 2021.

Liu, X., Wang, Z., Li, Y.-L., and Wang, S. Self-supervised learning via maximum entropy coding. *Advances in Neural Information Processing Systems*, 35:34091–34105, 2022.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Piza, E. L., Welsh, B. C., Farrington, D. P., and Thomas, A. L. Cctv surveillance for crime prevention. *Criminology & Public Policy*, 18(1):135–159, 2019. doi: https://doi.org/10.1111/1745-9133.12419. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9133.12419.

Şengönül, E., Samet, R., Abu Al-Haija, Q., Alqahtani, A., Alturki, B., and Alsulami, A. A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. *Applied Sciences*, 13(8):4956, January 2023. ISSN 2076-3417. doi: 10.3390/app13084956.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos, 2019.

Ullah, F. U. M., Obaidat, M. S., Ullah, A., Muhammad, K., Hijji, M., and Baik, S. W. A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos. *ACM Computing Surveys*, 55(10):200:1–200:44, February 2023. ISSN 0360-0300. doi: 10.1145/3561971.

Vaserstein, L. N. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.