Consistency-based Self-supervised Learning for Temporal Anomaly Localization

Aniello Panariello[1], Angelo Porrello[1], Simone Calderara[1], and Rita Cucchiara[1]

AImageLab, University of Modena and Reggio Emilia, Italy
firstname.lastname@unimore.it

Abstract. This work tackles Weakly Supervised Anomaly detection, in which a predictor is allowed to learn not only from normal examples but also from a few labeled anomalies made available during training. In particular, we deal with the localization of anomalous activities within the video stream: this is a very challenging scenario, as training examples come only with video-level annotations (and not frame-level). Several recent works have proposed various regularization terms to address it *i.e.* by enforcing sparsity and smoothness constraints over the weakly-learned frame-level anomaly scores. In this work, we get inspired by recent advances within the field of self-supervised learning and ask the model to yield the same scores for different augmentations of the same video sequence. We show that enforcing such an alignment improves the performance of the model on XD-Violence.

Keywords: Video Anomaly Detection, Temporal Action Localization, Weakly Supervised, Self Supervised Learning

1 Introduction

The goal of Video Anomaly Detection is to detect and localize anomalous events occurring in a video stream. Usually, these events involve human actions such as abuse, falling, fighting, theft, etc. In the last decades, such a task has gained relevance thanks to its potential for a video surveillance pipeline. In light of the widespread presence of CCTV cameras, employing enough personnel to examine all the footage is indeed infeasible and, therefore, automatic tools must be exploited.

Traditional approaches leverage low level features [16] to deal with this task, computed either on visual cues [3, 22] or object trajectories [8, 35]. These techniques – which could suffer and be unreliable in some contexts [20, 32] – have been surpassed by reconstruction-based approaches [19]: these methods require a set of normal data to learn a model of regularity; afterward, an example can be assessed as "anomalous" if it substantially deviates from the learned model. In this regard, the architectural design often resorts to deep unsupervised autoencoders [45, 55], which may include several strategies for regularizing the latent space.

The major recent trend [42] regards the exploitation of weak supervision, which provides the learning phase also with examples from the anomalous classes. However, the annotations come in a video-level format: the learner does not know in which time steps the anomaly will show, but only if it does at least one time. For such a reason, the proposed techniques often fall into the framework of multiple instance learning [17, 43, 47], devising additional optimization constraints that mitigate the lack of frame-level annotations.

In this work, we attempt to complement the existing regularization techniques with an idea from the fields of self-supervised learning and consistency regularization [6, 15, 41, 48]. In this context, data augmentation is exploited to generate positive pairs, consisting of a couple of slightly different versions of the same example. Afterward, the network is trained to output very similar representations, in a self-supervised manner (no class labels are required). Similarly, we devise a data augmentation strategy tailored to sequences and encourage the network to assign the same frame-level anomaly scores for the elements of each positive pair. Such a strategy resembles the smoothness constraint often imposed in recent works; however, while those approaches focus on adjacent time steps, our approach can randomly span over longer temporal windows.

To evaluate the merits of our proposal, we conduct experiments on the XD-Violence [47] dataset. We found that the presence of our regularization term yields remarkable improvements, although being not yet enough to reach state-of-the-art approaches.

2 Related work

Video Anomaly Detection Traditionally researchers and practitioners exploit object trajectories [8, 34, 35], low-level handcrafted features [3, 22, 33] (such as Histogram of Gradients and Histogram of Flows), and connected component analysis cues [2, 4, 5]. These traditional approaches have proven to be effective on benchmark datasets, but turn out to be still ineffective when used on a real domain. This means that these methods do not adapt well to anomalies that have never been seen before. For this reason, the task of anomaly detection has recently been approached with deep neural networks.

Most recent works lean towards unsupervised methods [1, 19] usually based on deep autoencoders. These models leverage the reconstruction error to measure how much an incoming example conforms to the training set distribution. Remarkably, the authors of [29] resorted to a variation of this common paradigm: namely, they proposed an approach that learns the distribution of normal data by guessing the appearance of future frames. In this respect, the idea was to compare the actual frame and the predicted one: if their difference is high, then it is possible to assume that an anomalous event occurred.

While it is easy to label whole videos (*e.g.*, anomaly present or not), the availability of fine-grained labeled anomalous videos is often scarce. For this reason, weakly supervised methods have seen a great advance. Among these works, Sultani *et al.* [42] introduced a new pattern for video anomaly detection termed

Multiple Instance Learning (MIL), upon which most of the subsequent weakly supervised methods developed. In this scenario, a positive video (*i.e.* containing an anomaly) and a negative one are taken into account at each iteration. These videos are usually split into segments; the model assigns them an anomaly score through a sequence of convolutional layers followed by an MLP. The scores are collected into a positive bag and a negative bag: as in the former, there is at least an anomaly (while in the latter there are no anomalies) the minimum score from the positive bag has to be as far as possible from the maximum score of the negative bag. To impose this constraint, the objective function contains a MIL ranking loss as well as smoothness and sparsity constraints. Zhou *et al.* [55] introduced the attention mechanism combined with the MIL approach, to improve the localization of the anomalies.

Temporal Action Localization The majority of the works dealing with Temporal Action Localization are either fully supervised or weakly supervised. The former ones can be broadly categorized into one stage [27, 30] and two stages methods [26, 49, 53]. For the first type, in works such as [27], action boundaries and labels are predicted simultaneously; in [30] this concept is developed by exploiting Gaussian kernels to dynamically optimize the temporal scale of each action proposal. On the other hand, two-stage methods initially generate temporal action proposals and then classify them. A manner to produce proposals is by exploiting the anchor mechanism [13,18,50], sliding window [40] or combining confident starting and ending frames of an action [26, 28].

Weakly supervised methods have been pioneered by UntrimmedNet [44], STPN [36] and AutoLoc [39], in which the action instances are localized by applying a threshold on the class activation sequence. This paradigm has been recently brought into the video anomaly detection settings in [31, 46].

3 Proposed Method

In the following, we present the proposed model and its training objective. The third part is dedicated to explaining the process of proposal generation, which consists of a post-processing step grouping adjacent similar scores into contiguous discrete intervals.

3.1 Model

In our set-up, each video is split into segments of 16 frames, with no overlap between consecutive segments. To extract video-level features, some works [12, 52] opted for combining 2D-CNNs and Recurrent Neural Networks; instead, we give each segment to a pre-trained I3D network [10]. Indeed, the authors of [24] have shown that the I3D features can prove to be effective for video-anomaly detection, even when a shallow classifier as XGBoost [9, 14] is employed for later classification. In our case, the I3D network is pre-trained on the Kinetics [21] dataset and not fine-tuned on our target data.

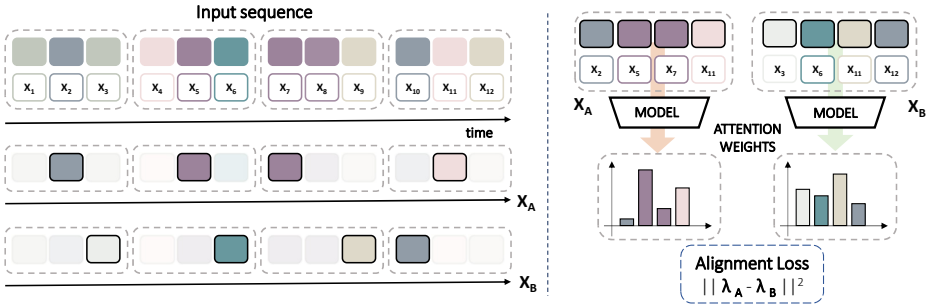

Fig. 1. Overview of the proposed framework. (left) Augmentation function sampling two slightly different sequences out of a single one. The original sequence gets split into windows and for each, we randomly sample a single feature vector. This is done twice to obtain two sequences X^A and X^B. (right) Both sequences are separately fed to the model, obtaining two sequences of attention weights λ_A and λ_B, pulled closer together by the alignment loss.

This way, each example is represented by a variable-length sequence of T feature vectors $\mathcal{X} = (x_1, x_2, \ldots, x_T)$. During training, examples come with a label y indicating whether an anomalous event appears in that sequence at least one time. Hence, given a training set of examples $\{(\mathcal{X}_j, y_j)\}_{j=1}^{\mathcal{N}}$ forged in the above-mentioned manner, we seek to train a neural network $f(\cdot;\theta)$ that solves such a task with the lowest empirical error:

$$\min_{\theta} \quad \frac{1}{\mathcal{N}} \sum_{i=1}^{N} \mathrm{BCE}(f(\mathcal{X}_i;\theta), y_i) \tag{1}$$

where $\mathrm{BCE}(\cdot,\cdot)$ stands for the binary cross entropy loss function. For the architectural design of $f(\cdot;\theta)$, we took inspiration from [36]. It consists of two main parts, discussed in the following paragraphs: namely, the computation of attention coefficients and the creation of an aggregate video-level representation.

Attention coefficients The aim of this module is to assign a weight $\lambda_t \in [0,1]$ to each element of the input sequence. As explained in the following, these weights will identify the most salient segments of the video *i.e.*, the likelihood of having observed an abnormal event within each segment. The module initially performs a masked temporal 1D convolution [25] to allow each feature vector to encode information from the past. Such a transformation – which does not alter the number of input feature vectors – is followed by two fully connected layers activated by ReLU functions, except for the last layer where a sigmoid function is employed.

Video-level representation Once we have guessed the attention values, we exploit them to aggregate the input feature vectors. Such an operation – which

resembles a temporal weighted average pooling presented in [37] – produces a single feature vector with unchanged dimensionality; formally:

$$\mathbf{x} = \sum_{t=1}^{T} \lambda_t x_t. \tag{2}$$

We finally feed it to a classifier $g(\cdot)$, composed of two fully connected layers. The final output represents the guess of the network for the value of y.

3.2 Training Objective

As mentioned before, we train our network in a weakly supervised fashion *i.e.*, only video level labels are provided to the learner. However, to provide a stronger training signal and to encourage attention coefficients to highlight salient events, we follow recent works [7, 51] and use some additional regularization terms that encode a prior knowledge we retain about the dynamics of abnormality.

Often, the presence of anomalous activities is characterized by the properties of sparseness and smoothness. Namely, anomalies appear rarely (*i.e.*, normal events dominate) and transitions between the two modalities usually occur throughout multiple frames. Such a peculiarity represents a prior we would like to enforce over the scores learned by the model: the majority of them should be close to zero and have similar values for video segments. In formal terms, the first constraint can be injected by penalizing the l_1 norm [11] of the attention weights, as follows:

$$\mathcal{L}_{sp} = \|\lambda\|_1, \tag{3}$$

while the second one can be carried out by imposing adjacent coefficients to vary as little as possible:

$$\mathcal{L}_{sm} = \sum_{t=1}^{T-1} (\lambda_t - \lambda_{t+1})^2. \tag{4}$$

Alignment loss Our main contribution consists in adding a consistency-based regularization term to the overall objective function. Overall, the idea is to generate two slightly different sequences out of a single one and, then, to encourage the model to produce the same attention coefficients for the two inputs.

To do so, we introduce a data augmentation function, shown in Fig. 1, that allows us to forge different versions $\mathcal{X}_A$ and $\mathcal{X}_B$ from the same example $\mathcal{X}$. In more detail, we split each sequence $(x_1, x_2, \ldots, x_T)$ into fixed-size blocks, whose length L is a hyperparameter we always set to 3; afterward, we randomly choose a feature vector within each block.

Once the variants $\mathcal{X}_A$ and $\mathcal{X}_B$ have been created, we ask the network to minimize the following objective function:

$$\mathcal{L}_a = \sum_{t=1}^{T} (\lambda_t^A - \lambda_t^B)^2, \tag{5}$$

where λ^A and λ^B are respectively the attention coefficients computed by the network for $\mathcal{X}_A$ and $\mathcal{X}_B$. With this additional regularization term, we seek to enforce that not only adjacent time-steps should have the same weight, but also those lying within a wider temporal horizon.

Overall objective Finally, the objective function will be:

$$\mathcal{L} \equiv \mathcal{L}_{cl} + \alpha\mathcal{L}_{sp} + \beta\mathcal{L}_{sm} + \gamma\mathcal{L}_{a}, \tag{6}$$

where the parameters α, β, and γ give different weights to each loss component and $\mathcal{L}_{cl}$ is the binary cross entropy loss function.

3.3 Temporal Proposal

During inference, we refine the anomaly scores by applying a post-processing step. Usually, two segments considered paramount by the network are interleaved by "holes", mostly due to noisy acquisitions or poor representations. The purpose of this phase is therefore to merge temporally close detections in a single retrieved candidate. In particular, as done in [54], we initially take out from the candidate set all those time-steps whose corresponding attention scores are lower than a certain threshold (in our experiments, < 0.35). The remaining non-zero scores will be used to generate the temporal proposal.

To generate the proposals, we do not use the rough coefficients, but instead a more refined version. In particular, we compute a 1-d activation map in the temporal domain, called Temporal Class Activation Map (T-CAM) [54], which indicates the relevance of the segment t in the prediction of one of the two classes involved (*normal vs anomalous*). Each value a_t of such activation map is computed as $a_t = g(x_t)$, *i.e.*, the guess of the classifier $g(\cdot)$ (introduced in Sec. 3.1) if masking the contributions of all time-steps except the t-th one. Furthermore, we extract the Weighted T-CAM, which combines the attention weight and the T-CAM activation values, *i.e.*, $\psi_t = \lambda_t \cdot a_t$. This operation let us emphasize the most important features for generating the proposal.

The last operation involves interpolating the weighted scores in the temporal axis and taking the bounding box that covers the largest connected component [36] to generate the final proposal [54]. The anomaly score for each proposal is then computed as:

$$\sum_{t=t_{start}}^{t_{end}} = \lambda_t \cdot \frac{a_t}{t_{end} - t_{start} - 1}, \tag{7}$$

where t_{start} and t_{end} represent the beginning and the ending of a single proposal.

4 Experiments

We conduct our experiments on the **XD-Violence Dataset** [47], a multi-modal dataset that contains scenes from different sources such as movies, sports, games,

	Video Level		Segment Level		Frame Level Proposal		Frame Level	
Align Loss	AUC%	AP%	AUC%	AP%	AUC%	AP%	AUC%	AP%
-	**97.91**	**98.36**	84.39	66.75	85.14	68.01	84.57	65.96
✓	97.79	98.28	**85.49**	**66.87**	**90.23**	**71.68**	**85.65**	**66.05**

Table 1. For different levels and metrics, results of our model with and without the proposed align loss. There is an improvement on almost all metrics when leveraging the proposed term.

Supervision	Method	AP%
Unsupervised	SVM baseline	50.78
	OCSVM [38]	27.25
	Hasan et al. [19]	30.77
Weakly Supervised	Ours (no align loss)	68.01
	Ours	71.68
	Sultani et al. [42]	75.68
	Wu et al. [47]	75.41
	RTFM [43]	77.81

Table 2. Comparison with recent works. We report the frame-level AP score on XD-Violence for both unsupervised and weakly-supervised methods. All the competitors exploit the I3D network for extracting features from RGB frames.

news, and live scenes. It holds a great variety concerning the devices used for video acquisition; indeed, the examples were captured by CCTV cameras, handheld cameras, or car driving recorders. There are 4754 videos for a total of 217 hours: among all these, 2405 are violent videos and 2349 are non-violent. The training set consists of 3954 examples; the test set, instead, features 800 ones, split into 500 violent and 300 non-violent videos. The dataset comes with multiple modalities such as RGB, optical flow, and audio; however, we restrict our analysis only to the RGB input domain.

XD-Violence also comes with segment-level ground truth labels; however, we only use video-level annotation during training, conforming to the weakly supervised setting. Differently, we exploit both the segment-level and frame-level annotations during the test phase, thus evaluating the model's capabilities for fine-grained localization.

Metrics We used the most popular metrics in anomaly detection settings, namely the Area Under Receiver Operating Characteristic Curve (AUC) and the Area Under Precision-Recall Curve (AP). While the AUC tends to be optimistic in the case of unbalanced datasets, the AP gives a more accurate evaluation of these scenarios.

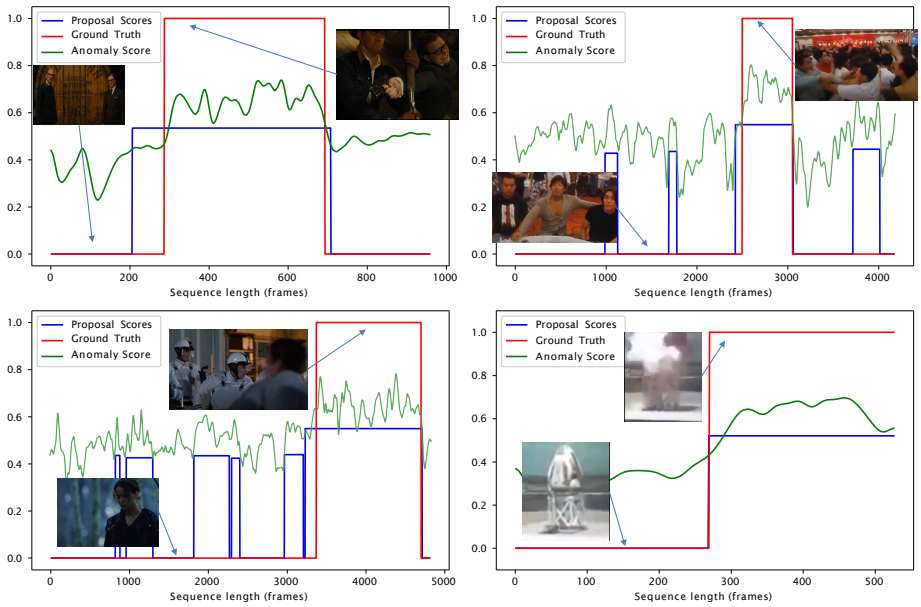

Fig. 2. Qualitative examples of the capabilities of our model to perform anomaly localization. The temporal proposal scores are indicated with a blue line, while the weighted T-CAM scores and the ground truth are shown in green and red respectively.

We assess the performance of the classification head by providing these two metrics at the video level; therefore, we use the attention scores, their interpolation, as well as the temporal proposals to assess the other grains.

Training Details We use the Adam optimizer [23] with a learning rate of 10^{-4} for the first 10 epochs and 10^{-5} for the remaining 40 epochs. The hyperparameters for the loss components are set to: $\alpha = 2 \times 10^{-8}$, $\beta = 0.002$, $\gamma = 0.5$. The threshold for taking out the low weighted T-CAM scores is set to 0.35; the batch size equals 8.

Results Tab 1 reports the comparison between the baseline approach with and without the proposed alignment objective. It can be observed that its addition leads to a remarkable improvement in most of the metrics. In particular, we have a gain of about 1% in AUC for segment and frame level metrics, while the AP remains almost the same. The greatest improvement subsists for the temporal proposal metric, where we gain 5 points in AUC and around 4 points in AP. The video level metrics remain approximately the same but still very high.

When comparing our approach with other recent works (see Tab 2), it can be seen that it outperforms the unsupervised state-of-the-art methods; however, it is in turn surpassed by the weakly supervised ones. We conjecture that such

a gap is mainly due to the bag representations inherent in these approaches, which could confer superior robustness; therefore, we leave to future works the extension of our idea to these methods.

Qualitative analysis Fig 2 presents several qualitative results, showing two fight scenes in the first row, and a riot and an explosion in the second one. We notice that the scores rightly increase when an anomalous action begins. Unfortunately, they remain always close to the uncertainty regime (assuming a score around 0.5) and never tend towards discrete decisions. In future works, we are going to address also this issue.

Looking at the original videos, we could also explain why the scores yielded by the model are noisy and subject to local fluctuations. Indeed, as sudden characters or camera' movements are likely to intervene, they are mistaken for real anomalies by the model, which is susceptible to these visual discontinuities due to the lack of segment-level annotations during training. Differently, when relying on the entire video sequence, the model can conversely recognize them correctly.

5 Conclusions

This work proposes a novel strategy to learn effective frame-level scores in weakly supervised settings when only video-level annotations are made available to the learner. Our proposal – which builds upon recent advances in the field of self-supervised learning – relies on maximizing the alignment between the attention weights of two different augmentations of the same input sequence. We show that a base network equipped also with other common regularization strategies (*e.g.* sparsity and smoothness) brings even more improvements. We found that it is not enough to achieve state-of-the-art performance; however, we leave aside for future works a comprehensive investigation of its applicability to more advanced architectures.

Acknowledgments. This work has been supported in part by the InSecTT project, funded by the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement 876038. The Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and AU, SWE, SPA, IT, FR, POR, IRE, FIN, SLO, PO, NED and TUR. The document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

References

1. Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2019)

2. Amraee, S., Vafaei, A., Jamshidi, K., Adibi, P.: Anomaly detection and localization in crowded scenes using connected component analysis. Multimedia Tools and Applications (2018)
3. Benezeth, Y., Jodoin, P.M., Saligrama, V., Rosenberger, C.: Abnormal events detection based on spatio-temporal co-occurences. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. IEEE (2009)
4. Bolelli, F., Allegretti, S., Baraldi, L., Grana, C.: Spaghetti labeling: Directed acyclic graphs for block-based connected components labeling. IEEE Transactions on Image Processing (2019)
5. Bolelli, F., Allegretti, S., Grana, C.: One dag to rule them all. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
6. Boschini, M., Buzzega, P., Bonicelli, L., Porrello, A., Calderara, S.: Continual semi-supervised learning through contrastive interpolation consistency. arXiv preprint arXiv:2108.06552 (2021)
7. Cai, R., Zhang, H., Liu, W., Gao, S., Hao, Z.: Appearance-motion memory consistency network for video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)
8. Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., Tishby, N.: Detecting anomalies in people's trajectories using spectral graph analysis. Computer Vision and Image Understanding (2011)
9. Candeloro, L., Ippoliti, C., Iapaolo, F., Monaco, F., Morelli, D., Cuccu, R., Fronte, P., Calderara, S., Vincenzi, S., Porrello, A., et al.: Predicting wnv circulation in italy using earth observation data and extreme gradient boosting model. Remote Sensing (2020)
10. Carreira, J., 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 (2017)
11. Cascianelli, S., Costante, G., Crocetti, F., Ricci, E., Valigi, P., Luca Fravolini, M.: Data-based design of robust fault detection and isolation residuals via lasso optimization and bayesian filtering. Asian Journal of Control (2021)
12. Cascianelli, S., Costante, G., Devo, A., Ciarfuglia, T.A., Valigi, P., Fravolini, M.L.: The role of the input in natural language video description. IEEE Transactions on Multimedia (2019)
13. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2018)
14. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (2016)
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (2020)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Ieee (2005)
17. Feng, J.C., Hong, F.T., Zheng, W.S.: Mist: Multiple instance self-training framework for video anomaly detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2021)

18. Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R.: Turn tap: Temporal unit regression network for temporal action proposals. In: IEEE International Conference on Computer Vision (2017)
19. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
20. Hu, X., Hu, S., Huang, Y., Zhang, H., Wu, H.: Video anomaly detection using deep incremental slow feature analysis network. IET Computer Vision (2016)
21. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
22. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. IEEE (2009)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2014)
24. Koshti, D., Kamoji, S., Kalnad, N., Sreekumar, S., Bhujbal, S.: Video anomaly detection using inflated 3d convolution network. In: International Conference on Inventive Computation Technologies (ICICT). IEEE (2020)
25. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2017)
26. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: IEEE International Conference on Computer Vision (2019)
27. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM international conference on Multimedia (2017)
28. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (2018)
29. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2018)
30. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2019)
31. Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J.: Localizing anomalies from weakly-labeled videos. IEEE transactions on image processing (2021)
32. Medel, J.R., Savakis, A.: Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv preprint arXiv:1612.00390 (2016)
33. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. IEEE (2009)
34. Monti, A., Porrello, A., Calderara, S., Coscia, P., Ballan, L., Cucchiara, R.: How many observations are enough? knowledge distillation for trajectory forecasting. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2022)
35. Morris, B.T., Trivedi, M.M.: Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)

36. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2018)
37. Porrello, A., Vincenzi, S., Buzzega, P., Calderara, S., Conte, A., Ippoliti, C., Candeloro, L., Di Lorenzo, A., Dondona, A.C.: Spotting insects from satellites: modeling the presence of culicoides imicola through deep cnns. In: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE (2019)
38. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In: Advances in Neural Information Processing Systems (1999)
39. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European Conference on Computer Vision (2018)
40. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
41. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems (2020)
42. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2018)
43. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: IEEE International Conference on Computer Vision (2021)
44. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2017)
45. Wang, X., Che, Z., Jiang, B., Xiao, N., Yang, K., Tang, J., Ye, J., Wang, J., Qi, Q.: Robust unsupervised video anomaly detection by multipath frame prediction. IEEE transactions on neural networks and learning systems (2021)
46. Wu, J., Zhang, W., Li, G., Wu, W., Tan, X., Li, Y., Ding, E., Lin, L.: Weakly-supervised spatio-temporal anomaly detection in surveillance video. In: International Joint Conferences on Artificial Intelligence (2021)
47. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: Proceedings of the European Conference on Computer Vision (2020)
48. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems (2020)
49. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: IEEE International Conference on Computer Vision (2017)
50. Yang, L., Peng, H., Zhang, D., Fu, J., Han, J.: Revisiting anchor mechanisms for temporal action localization. IEEE Transactions on Image Processing (2020)
51. Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., Kloft, M.: Cloze test helps: Effective video anomaly detection via learning to complete video events. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 583–591 (2020)

52. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2015)
53. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: IEEE International Conference on Computer Vision (2019)
54. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
55. Zhu, Y., Newsam, S.: Motion-aware feature for improved video anomaly detection. In: British Machine Vision Conference (2019)