

# Semi-supervised Anomaly Detection for Weakly-annotated Videos

Khaled El-Tahan<sup>a</sup> and Marwan Torki<sup>b</sup>

Computer and Systems Engineering Department, Alexandria University, Egypt

**Keywords:** Semi-supervision, Pseudo Labels, Weak-supervision, Multiple Instance Learning, Anomaly Detection, Background Subtraction, Video Recognition.

**Abstract:** One of the significant challenges in surveillance anomaly detection research is the scarcity of surveillance datasets satisfying specific ethical and logistical requirements during the collection process. Weakly supervised models aim to solve those challenges by only weakly annotating surveillance videos and creating sophisticated learning techniques to optimize these models, such as Multiple Instance Learning (MIL), which maximizes the boundary between the most anomalous video clip and the least normal (false alarm) video clip using ranking loss. However, maximizing the boundary does not necessarily assign each clip its correct class. We propose a semi-supervision technique that creates pseudo labels for each correct class. Also, we investigate different video recognition models for better features representation. We evaluate our work on the UCF-Crime (Weakly Supervised) dataset and show that it almost outperforms all other approaches by only using the same simple baseline (multilayer perceptron neural network). Moreover, we incorporate different evaluation metrics to show that not only did our solution increase the AUC, but it also increased the top-1 accuracy drastically.

## 1 INTRODUCTION


Almost all public places now rely on surveillance cameras to increase public safety. However, the human need for surveillance analysis is very high in demand and very costly. The need for automatic surveillance anomaly detection systems is now higher than ever. The challenge with surveillance anomaly detection models is the dataset availability; many ethical and logistical requirements prevent us from collecting private surveillance videos. We have to rely on the publicly available videos, but even with that, those videos are long and diverse by nature. Long videos mean that we must do tons of work to annotate it thoroughly, and diverse videos suggest that many classical statistical-based vision algorithms cannot be applied.


Different methods were introduced to approach surveillance video anomaly detection. Solutions like (Liu and Ma, 2019) and (Landi et al., 2019) annotated different datasets and addressed the detection as a fully supervised problem. However, despite the good results achieved by fully supervised solutions, they require exhaustive human effort in the data annotation process. Other solutions like (Georgescu et al.,

2021) and (Cai et al., 2021) address the detection as an unsupervised problem. Although they technically solved the annotation difficulty, they achieved unappealing results. Also, since most videos on the internet have a title, there is no need to throw away that information; it is better to use the title as a means to create a weakly annotated dataset.

The work provided by (Sultani et al., 2018) addresses those problems by providing a weakly annotated surveillance videos dataset to avoid the hassle of dataset annotation and builds a Multiple Instance Learning (MIL) baseline to leverage the weakly annotated data. Given a video, we say it is weakly annotated when we know the label of this video, but we do not know the label of each clip (temporal segment) of the video. For an anomalous video, we know that the video contains an anomalous clip or more, but we do not know their exact temporal location. However, for the normal videos, there exist no anomalous segments at all; hence, normal videos are fully annotated by nature.

Despite promising results achieved by the weakly supervised models, it suffers from a critical problem due to its learning style. Relying on the Multiple Instance Learning and ranking loss optimizes the models by maximizing the boundary between the anomalous and normal clips. However, maximizing the

<sup>a</sup>  <https://orcid.org/0000-0001-8955-2319>

<sup>b</sup>  <https://orcid.org/0000-0002-6149-1718>

boundary between two different classes does not guarantee that each class will be assigned correctly.

**Our Contributions Can Be Summarized as Follows:**

- We present a novel semi-supervised solution on top of the weakly supervised model; we produce pseudo labels from a confident weakly supervised model and use them to guide the training in assigning each clip its correct class.
- We investigate different video recognition models for better video representations.
- We achieve results comparable to state-of-the-art despite using only the superficial (multilayer perceptron neural network) baseline provided in (Sultani et al., 2018).

The rest of this paper is structured as follows. Section 2 discusses the related work and the different approaches to resolving the annotation challenge. Section 3 presents our methods. Section 4 describes our experiments and the produced results. Finally, Section 5 concludes our writing and discusses future work.

## 2 RELATED WORK

We discuss here the related work for **anomaly detection in surveillance videos** in the first three subsections. And then, we use the fourth subsection to discuss semi-supervision related work on **image classification** since using semi-supervision in anomaly detection on weakly-supervised data is a novel approach.

### 2.1 Fully-supervised

Fully supervised solutions address anomaly detection in surveillance videos as a labeled classification problem by completely annotating all the videos. An example of those solutions is (Liu and Ma, 2019) which explores the background bias and then trains a region loss to drive the network to learn the anomalous region explicitly. Another example is (Landi et al., 2019) which studies the impact of considering spatiotemporal tubes instead of whole-frame video segments.

Despite achieving good results, those solutions require exhausting human resources and high costs for videos annotation and do not present a practical solution to utilize all the publicly available videos.

### 2.2 Unsupervised

Unsupervised anomaly detection in surveillance studies videos representations to create recognizable patterns without the need for labels at all. For instance, (Hasan et al., 2016) detect anomalies by creating generative models to learn the regular motion. Also, (Luo et al., 2017) proposes a Temporally-coherent Sparse Coding (TSC) to enforce alike adjacent frames be encoded with similar reconstruction coefficients. (Wang et al., 2018a) suggests a two-stage approach in which they estimate the normal events globally from the entire unlabeled videos and then feed the estimated normal clips into a one-class support vector machine to build a refined normality model. Other solutions like (Gong et al., 2019) uses autoencoders to produce higher reconstruction error for the abnormal inputs than the normal ones. (Morais et al., 2019) models the normal patterns of human movements in surveillance video using dynamic skeleton features to identify human-related anomalous segments. Moreover, (Park et al., 2020) uses a memory module with a different update design to record the prototypical patterns of normal data. (Wang et al., 2020) suggests a contrastive representation learning task to establish subcategories of normality as clusters. (Georgescu et al., 2021) approaches abnormal event detection in the video through self-supervised and multi-task learning at the object level instead of the frame level. Furthermore, (Cai et al., 2021) uses prior knowledge of appearance and motion signals to capture their correspondence in the high-level feature space.

Those solutions are creative, and they omit the need for labels at all. However, their performance is not appealing compared to other supervised and weakly supervised solutions. Furthermore, going unsupervised is a bit extreme since most publicly available surveillance videos on the internet have titles that can be utilized as a weak label.

### 2.3 Weakly-supervised

The UCF-Crime dataset and the baseline provided by (Sultani et al., 2018) are the origins for all weakly supervised anomaly detection in surveillance videos. (Zhu and Newsam, 2019) extends it by proposing an augmented temporal network to learn a motion-aware feature. (Zhang et al., 2019) defines an inner bag loss (IBL) for MIL to constrain the function space of the weakly supervised problem. Finally, (Feng et al., 2021) produces a multiple instance self-training framework (MIST) to refine task-specific discriminative representations with only video-level annotations.

## 2.4 Semi-supervised

Semi-supervised learning relies on labeled data to train the model, then uses the model to produce pseudo labels on the highly certain unlabeled data. This idea was first introduced on image classification by (Lee et al., 2013). Later approaches like (Berthelot et al., 2019b) works by guessing low-entropy labels for data-augmented unlabeled examples and mixing labeled and unlabeled data. (Berthelot et al., 2019a) improves (Berthelot et al., 2019b) by introducing distribution alignment and augmentation anchoring. Finally, (Sohn et al., 2020) first generates pseudo-labels using the model’s predictions on weakly-augmented unlabeled images and then trained to predict the pseudo-label when fed a strongly-augmented version of the same image.

Semi-supervised learning has been used only for semi-labeled datasets (i.e., part of the dataset is fully labeled, and the other part is unlabeled). Our work builds a novel semi-supervision scheme on top of the weakly supervised baseline (Sultani et al., 2018).

## 3 APPROACH

The proposed approach uses a mixup between multiple instance learning (weak supervision) and pseudo labels (semi-supervision) to optimize the model. The multiple instance learning is summarized in figure 1. The pseudo labels training is summarized in figure 2. The mixup between both approaches is illustrated in figure 3.

### 3.1 Multiple Instance Learning

For multiple instance learning we use the baseline provided by (Sultani et al., 2018), with one difference, we use SlowFast (Feichtenhofer et al., 2019) instead of C3D (Tran et al., 2015) for surveillance video features extraction.

We start by splitting each surveillance video into a bag of temporal segments (clips) then apply feature extraction on those clips. We take pairs of bags (normal and anomalous) each as one training example. A ranking loss function would be a straightforward method to maximize the boundary between the normal and anomalous bags in each example, as in equation 1. (Sultani et al., 2018) uses the hinge function as a ranking loss function.

$$f(V_a) > f(V_n) \quad (1)$$

Where  $V_a$  and  $V_n$  represent anomalous and normal video segments,  $f(V_a)$  and  $f(V_n)$  represent the cor-

responding predicted scores produced from the multilayer perceptron classification neural network, respectively.

However, in the absence of video segment level annotations, it is not possible to use equation 1. Instead, (Sultani et al., 2018) proposes the following multiple instance ranking objective function:

$$\max_{i \in B_a} f(V_a^i) > \max_{i \in B_n} f(V_n^i) \quad (2)$$

Where  $B_a$  and  $B_n$  represent anomalous bag and normal bag, and  $max$  is taken over all videos segments in each bag.

The ranking loss in the hinge-loss formulation is therefore given as follows:

$$L(B_a, B_n) = \max(0, 1 - \max_{i \in B_a} f(V_a^i) + \max_{i \in B_n} f(V_n^i)) \quad (3)$$

However, since anomalous actions occur for a short time, few segments may contain anomalies, and since the video is a sequence of segments, the anomaly score should vary smoothly between adjacent video segments. Also, by incorporating sparsity and smoothness constraints on the instance scores, the loss function becomes:

$$L(B_a, B_n) = \max(0, 1 - \max_{i \in B_a} f(V_a^i) + \max_{i \in B_n} f(V_n^i)) + \lambda_1 \sum_i^{n-1} (f(V_a^i) - f(V_a^{i+1}))^2 + \lambda_2 \sum_i^n f(V_a^i) \quad (4)$$

Where  $\lambda_1$  and  $\lambda_2$  represent sparsity and smoothness constraints, respectively.

The final equation after adding the regularization is as follows:

$$L_{MIL} = L(B_a, B_n) + \|W\|_F \quad (5)$$

Where MIL stands for multiple instance learning.

### 3.2 Pseudo Labels

Semi-supervised solutions require semi-labeled datasets; however, we propose a new approach that relies on multiple instance learning with weakly annotated datasets instead of semi-labeled datasets.

We suggest a model similar to FixMatch (Sohn et al., 2020) in which we train our model on intervals of  $N$  epochs where  $N$  is a hyperparameter; we do, however, few modifications to fit our problem. First, since we do not have fully labeled examples, we use multiple instance learning instead in the first few epochs, as shown in figure 3. Once we are more certain about the model’s results, we start producing pseudo labels as summarized in figure 2. And finally, we use those pseudo labels to optimize the classification model actively.

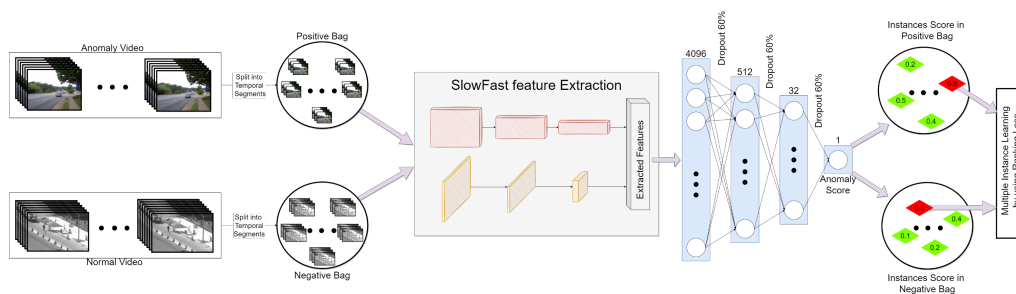


Figure 1: Demonstrates the method of multiple instance learning. We begin feeding the model with pair of videos. Each video is then divided into a predefined number of temporal segments (clips). Those segments are fed into a SlowFast (Feichtenhofer et al., 2019) feature extraction model to create two bags of clips features. Finally, we feed those bags into an MLP classification neural network and use a ranking loss to maximize the boundary between the most anomalous video clip and the least normal (false alarm) video clip.

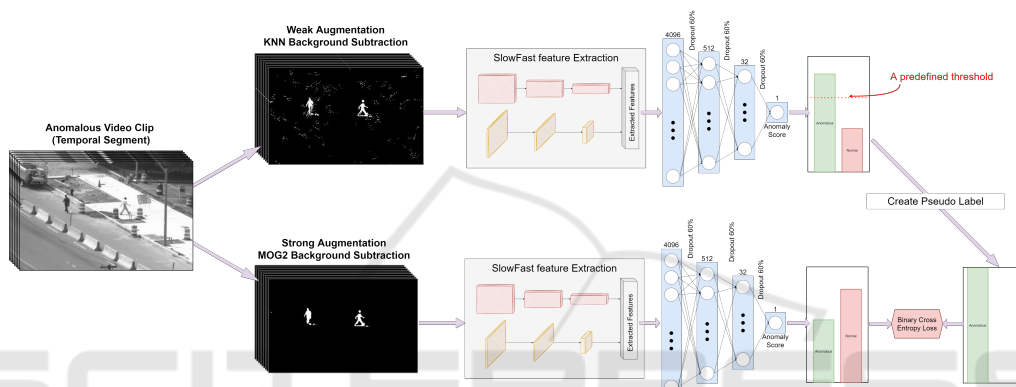


Figure 2: Illustrates the semi-supervision process. Given an anomalous surveillance video clip, we create two augmented copies of it. The first copy is weakly augmented by applying KNN background subtraction, and the second is strongly augmented by using MOG2 background subtraction. We then feed the two clips into the feature extraction module, and after that, we predict the probabilities of output classes for each clip (normal or anomalous). Suppose the anomalous class probability in the weak augmentation prediction is above a predefined threshold. In that case, a pseudo label is created and used to optimize the model with the strong augmentation prediction via binary cross-entropy.



Figure 3: Overall model training, we stack the training into intervals; each interval consists of N epochs where N is a hyperparameter. In the first interval, we only use multiple instance training until we are more certain about the model's prediction; we then increase the number of semi-supervised epochs per interval.

### 3.2.1 Anomalous Pseudo Labels

For anomalous surveillance video clips (temporal segments), we produce two copies, one is strongly aug-

mented, and the other is weakly augmented, just like in FixMatch (Sohn et al., 2020). Unlike in FixMatch, we do not use RandAugment nor CTAugment for augmentation. According to (Liu and Ma, 2019), surveil-

lance anomaly detection models are biased towards the background, which affects their performance. Inspired by this observation, we use background subtraction algorithms for augmentation. Using background subtraction meets our requirements to create augmented versions for the pseudo labels and reduces the background bias as a plus. We use MOG2 background subtraction for strong augmentation and KNN background subtraction for weak augmentation.

Both augmentations (weak and strong) are fed into the feature subtraction module, and an anomaly score is predicted for each of them; if the anomaly score for the weak augmentation is above a predefined threshold, then a pseudo label is produced and is used as a label to train the model with the strong augmentation input via binary cross-entropy loss. If otherwise (the score is below the threshold), no pseudo label is produced, and the training example is omitted.

### 3.2.2 Normal Pseudo Labels

Since normal videos are fully annotated (we know the label of each temporal segment), a pseudo label creation is not needed because we always have a label. However, we have to use normal labels with the same amount as the anomalous pseudo labels to maintain output classes distribution and prevent output bias.

### 3.2.3 Mathematical Formulation

The binary cross-entropy loss for normal labels by the semi-supervision process is defined as follows:

$$L_{PL-Normal} = \sum_{i \in B_n} BCE(1, f(V_n^i)) \quad (6)$$

Where BCE means binary cross-entropy and PL-Normal means the normal part of the pseudo labels method, note that while unneeded in equation 6, we use the concept of bags (i.e.,  $B_n$ ) to be able to add this equation with the multiple instance learning equations later.

After adding the anomalous pseudo labels loss:

$$L_{PL} = \sum_{i \in B_n} (f_{WeakAug}(V_a^i) > th) * (BCE(1, f(V_n^i)) + BCE(0, f_{StrongAug}(V_a^i))) \quad (7)$$

Where PL means pseudo labels,  $f_{StrongAug}(V_a^i)$  means the anomalous output of the strong augmentation input, and  $f_{WeakAug}(V_a^i)$  means the anomalous output of the weak augmentation input, and  $th$  is a hyperparameter with a value ( $0 \leq th \leq 1$ ) that we use as a predefined threshold.

Note that the condition ( $f_{WeakAug}(V_a^i) > th$ ) equals 1 if satisfied and 0 otherwise. This condition is multiplied to equation 6 too, to prevent classes distribution bias.

From equation 5 and equation 7 the overall loss function becomes:

$$L = \alpha * L_{MIL} + \beta * L_{PL} \quad (8)$$

Where  $\alpha$  and  $\beta$  are hyperparameters representing the multiple instance learning percentage and the pseudo labels training percentage, respectively, and ( $\alpha + \beta = 1$ ).

Finally, hyperparameters  $th$ ,  $\alpha$ , and  $\beta$  are established via grid search (LaValle et al., 2004).

## 4 EVALUATION

In this section we present our experimental setup and results.

### 4.1 Experimental Setup

We describe our feature extraction, classification model and evaluation metrics used.

#### 4.1.1 Feature Extraction

We extract the visual features from the surveillance videos using the different methods as follows:

- We first fix the frame rate to 32 fps. As shown in SlowFast (Feichtenhofer et al., 2019), having fps as a power of 2 delivers better time performance for the feature extraction process.
- We then apply the background subtraction if required.
- We resize all frames to 240 x 320 for all of them except for X3D, in which we resize to 312 x 416.
- For C3D (Tran et al., 2015), I3D (Carreira and Zisserman, 2017), I3D NLN (Wang et al., 2018b) and, SlowFast (Feichtenhofer et al., 2019): We use the 8x8 R80 architecture.
- For X3D (Feichtenhofer, 2020): We use the X3D-Large architecture.

#### 4.1.2 Classification Model

For the classification model, we use the simple multi-layer perceptron neural network baseline provided by (Sultani et al., 2018).



Table 1: AUC Comparisons with different experiments on top of the (Sultani et al., 2018) baseline. The second column shows the AUC of the baseline (MIL) with other feature extraction models, third and fourth columns show the AUC of the baseline with various features extraction models while feeding the input augmented by KNN and MOG2 background subtraction algorithms, respectively. The fifth column shows the AUC of the semi-supervised on top of weakly annotated data solution (Multiple Instance Learning + Pseudo Labels) with different feature extraction models.

Feature Extraction Module	AUC			
	MIL	MIL + KNN BS	MIL + MOG2 BS	MIL + PL
C3D (Tran et al., 2015)	75.41	74.20	71.14	78.60
I3D (Carreira and Zisserman, 2017)	76.84	73.45	70.84	79.31
I3D with NLN (Wang et al., 2018b)	77.32	74.01	72.76	79.89
X3D (Feichtenhofer, 2020)	76.98	73.86	72.49	80.06
SlowFast (Feichtenhofer et al., 2019)	79.37	77.70	73.04	<b>81.24</b>

The classification model consists of three fully connected layers. The first layer has 4096 units, followed by 512, 32, then 1 unit. Between every two layers, we apply 60% dropout for training.

We change, however, the loss sparsity and smoothness hyperparameters almost per each experiment. We will provide the specific configuration with the public source code.

#### 4.1.3 Evaluation Metrics

We follow previous (Sultani et al., 2018) in using the receive operating characteristic (ROC) curve and its corresponding area under the curve (AUC). However, we also evaluate our experiments using top accuracy for only the anomalous class.

## 4.2 Results and Comparisons

We present different set of experimental results. First, we show the role of feature extraction and its impact on the AUC. Next, we show that adding pseudo labels improves the Multiple Instance Learning framework as well as anomalous video classification. We also show the comparative evaluation against state-of-the-art methods. Finally, we show qualitative results on the detection we obtain against the baseline of (Sultani et al., 2018).

#### 4.2.1 Effect of Feature Extraction Method

The results shown in table 1 shows superior performance for SlowFast (Feichtenhofer et al., 2019) as a feature extractor comparing to the other methods. The 4% enhancement in performance to (Sultani et al., 2018) by using (Feichtenhofer et al., 2019) alone beats other sophisticated solutions like (Zhu and Newsam, 2019) and (Zhu and Newsam, 2019) which suggests that the surveillance anomaly detection models are struggling with the video representations.

#### 4.2.2 Effect of Background Subtraction and Pseudo Labeling

In table 1, we show that background subtraction using KNN or MOG on top of the multiple instance learning baseline (Sultani et al., 2018) lowered the AUC. This supports the findings of (Liu and Ma, 2019) that anomaly detection models on surveillance video are biased towards the background.

Also, table 1 shows that using the pseudo labels on top of the multiple instance learning increases consistently the AUC around 2% to 3%, which makes our solution surpasses the baseline (Sultani et al., 2018) with around 6% in AUC.

This result confirms our intuition that using pseudo labels will enhance the performance of the model of use.

#### 4.2.3 Effect of the Pseudo Labeling on Anomalous Class Accuracy

The most exciting results are anomalous class accuracy shown in table 2. We observe drastic improvement in anomalous class accuracy. This suggests that maximizing the boundaries between the anomalous and normal bags via ranking loss and multiple instance learning do not necessarily assign each segment its correct class. Forcing the model to increase the separation using the pseudo labels achieved two goals. The first goal is to increase the separation between normal and anomalous videos. The second goal is to assign each video the correct class label.

#### 4.2.4 Comparative Evaluation

We show a comparative evaluation against the state-of-the-art methods is anomaly detection. Our work beats most of the state-of-the-art results shown in table 3. However, we achieve comparable results to the best model (Feng et al., 2021). We achieved that without modifying the simple baseline provided by (Sultani et al., 2018). We just use more helpful videos

Table 2: Accuracy for the anomalous class at threshold=0.5 for both the weakly supervised solution and the semi-supervised on top of weak supervision solution; this accuracy is measured for different feature extraction models.

Feature Extraction Module	Anomalous class accuracy	
	MIL without PL	MIL + PL
C3D (Tran et al., 2015)	18.04	42.18
I3D (Carreira and Zisserman, 2017)	21.57	52.07
I3D with NLN (Wang et al., 2018b)	19.71	48.29
X3D (Feichtenhofer, 2020)	24.36	57.42
SlowFast (Feichtenhofer et al., 2019)	23.01	61.78

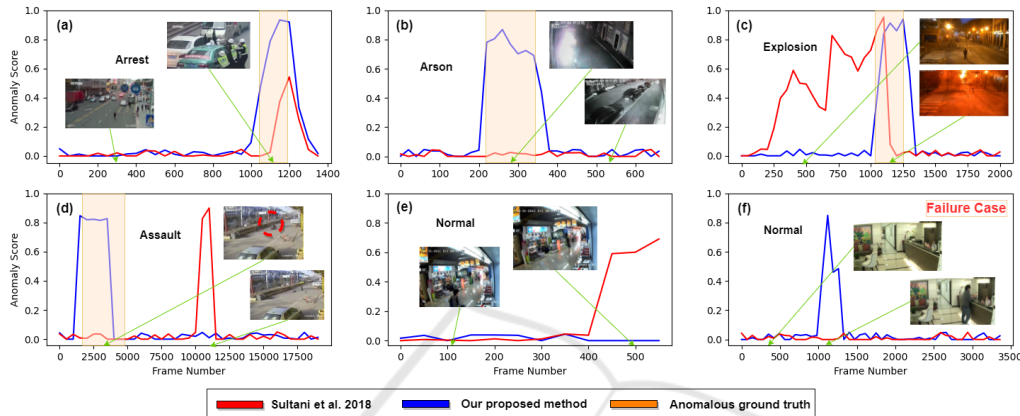


Figure 4: Qualitative results of our work (blue) method on testing videos compared against (Sultani et al., 2018) (red). The colored window represents the ground truth anomalous clip. (a), (b), (c), and (d) show videos containing anomalous segments, while (e) and (f) are normal videos. Our method beats (Sultani et al., 2018) in (a), (b), (c), (d), and (e), however it fails in (f).

representations (SlowFast) and more reliable training methodology (Pseudo Labels), which suggests future opportunities for better results.

Table 3: AUC Comparisons between state-of-the-art methods that rely on weakly supervised data.

Method	Reported AUC
(Sultani et al., 2018)	75.41
(Zhu and Newsam, 2019)	79.00
(Zhang et al., 2019)	78.66
(Feng et al., 2021)	<b>82.30</b>
Ours (MIL + PL)	81.24

#### 4.2.5 Qualitative Results

Finally, in figure 4 we show qualitative results of our work against (Sultani et al., 2018). (a)-(d) show anomalous videos in which our method beats (Sultani et al., 2018) and produces more accurate results. Also, in (e), we show that in normal videos, our method is more robust compared to (Sultani et al., 2018). We include (f) to represent a few cases in which our method fails; we believe that in this specific case, our method was unable to identify that this is an ordinary visit to a clinic due to background bias removal.

## 5 CONCLUSIONS

We propose a novel semi-supervision anomaly detection method on surveillance videos. Our novel method leverages pseudo labels produced by multiple instance learning for weakly supervised datasets. We also exploit the idea of background bias in surveillance anomaly detection to build a more robust pseudo labels augmentation. We use those pseudo labels for better guidance in the training process and achieve results comparable to the state-of-the-art while using a simple multilayer perceptron neural network.

As for future work, we intend to distill the produced pseudo labels by investigating different randomized augmentations techniques; we also plan to incorporate the semi-supervision method on the other state-of-the-art sophisticated classification models to outperform them and demonstrate the generality of our approach.

## ACKNOWLEDGEMENTS

This work is funded by the Science and Technology Development Fund STDF (Egypt); Project id: 42519 - “Automatic Video Surveillance System for Crowd Scenes”.

## REFERENCES

- Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. (2019a). Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring.
- Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. (2019b). Mixmatch: A holistic approach to semi-supervised learning.
- Cai, R., Zhang, H., Liu, W., Gao, S., and Hao, Z. (2021). Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 938–946.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.
- Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213.
- Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.
- Feng, J.-C., Hong, F.-T., and Zheng, W.-S. (2021). Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14009–14018.
- Georgescu, M.-I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., and Shah, M. (2021). Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12742–12752.
- Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. v. d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714.
- Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742.
- Landi, F., Snoek, C. G., and Cucchiara, R. (2019). Anomaly locality in video surveillance.
- LaValle, S. M., Branicky, M. S., and Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, 23(7-8):673–692.
- Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.
- Liu, K. and Ma, H. (2019). Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1490–1499.
- Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349.
- Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., and Venkatesh, S. (2019). Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12004.
- Park, H., Noh, J., and Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381.
- Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence.
- Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.
- Wang, S., Zeng, Y., Liu, Q., Zhu, C., Zhu, E., and Yin, J. (2018a). Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 636–644.
- Wang, X., Girshick, R., Gupta, A., and He, K. (2018b). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.
- Wang, Z., Zou, Y., and Zhang, Z. (2020). Cluster attention contrast for video anomaly detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2463–2471.
- Zhang, J., Qing, L., and Miao, J. (2019). Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE.
- Zhu, Y. and Newsam, S. (2019). Motion-aware feature for improved video anomaly detection.