Ada-VAD: Domain Adaptable Video Anomaly Detection

Dongliang Guo* Yun Fu† Sheng Li‡

Abstract

Video anomaly detection (VAD) aims at identifying unusual behaviors from videos. Most of the existing video anomaly detection methods can achieve promising performance in the scenarios where training and test samples are drawn from the same distribution. In real-world situation, however, it is intractable to collect and label sufficient training video samples that cover many possible test scenarios, and existing methods demonstrate limited generalization ability. Focusing on this issue, we present the few-shot cross-domain video anomaly detection (FC-VAD) problem, which aims to adapt anomaly detection model to target samples, with access to only a few target video frames. To solve the FC-VAD problem, we propose an adaptive video anomaly detection framework named Ada-VAD, which contains a pretraining stage and an adaptation stage. In the pretraining stage, we synthesize abnormal samples and design a self-supervision based prediction task to pretrain a domain invariant model. In the adaptation stage, we adapt the pre-trained model to target domain with few-shot samples by mitigating the distribution shift with an adversarial training approach. We conduct extensive experiments on three benchmark datasets, and results show that our Ada-VAD approach outperforms the state-of-the-art VAD methods in most cases. Our code is available at `https://github.com/donglgcn/ADA-VAD`

Keywords: Domain Adaptation, Video Anomaly Detection, Self-supervised learning.

1 Introduction

Anomalies [37, 38, 10] has been extensively studied in the data mining field. Video anomaly detection (VAD), as a special setting of anomaly detection, has numerous real-world applications such as video surveillance. VAD aims to detect anomalous samples that deviate from the predefined normality during testing [40]. Most of the existing works apply conventional anomaly detection techniques to address the VAD problem [7, 19, 2, 17, 42, 35, 8]. Specifically, they learn normal patterns from a large number of normal training videos. During inference, anomalies can be detected because they deviate from normal patterns. According to different ways of learning normal patterns, existing VAD methods could be categorized as reconstruction-based methods [42], prediction-based methods [2, 17, 19], classification methods [8], and self-supervised learning methods [7, 35]. These methods have reached very promising accuracy if the training videos are sufficient and carefully selected by domain experts.

What is the insufficient data issue that hinders current VAD methods? In practice, it is very hard to collect a large number of training videos to cover a wide variety of test scenarios. Moreover, it is intractable to pick the normal videos from the endless videos. There are some possible ways to deal with the insufficient data issue, such as unsupervised learning [43] and few-shot learning methods [31, 22, 4], but they cannot obtain comparable results to traditional VAD methods.

What is the potential solution to the insufficient data issue? Unlike existing work, we aim to address the aforementioned data insufficient issue using domain adaptation. Assume that we have access to enough training samples in some datasets (i.e., source domains), which could be used to pre-train a VAD model. Collecting sufficient training videos from every test scenario (i.e., target domain) seems impractical, but it is feasible to capture a few normal video frames in the target domain. Thus, it is possible to adapt the pre-trained model to the target domain. Formally, we introduce the few-shot cross-domain video anomaly detection (FC-VAD) problem in this paper.

What are the challenges of FC-VAD? Adapting models across domains is a challenging task, mainly due to the distribution shift between the source domain and the target domain. In particular, most existing VAD methods focus on one scenario, i.e., training videos and test videos are from the same domain and follow the same distribution. These methods are likely to fail when they leverage the source domain pattern to identify video anomalies from other domains. In other words, due to the distribution shift, the normal patterns learned in the source domain are different from those in the target domain. We empirically demonstrate this phenomenon in Section. 3. Very few VAD methods [8, 22, 5, 31] attempt to be scene-agnostic. They are conceptually relevant to domain adaptation, but they mainly focus on generating a scene-insensitive model.

How does our method address these challenges? In this paper, we propose a novel adaptive video anomaly detection (Ada-VAD) framework to address the FC-VAD prob-

*University of Virginia, USA. Email: dongliang.guo@virginia.edu
†Northeastern University, USA. Email: yunfu@ece.neu.edu
‡University of Virginia, USA. Email: shengli@virginia.edu

lem. First, we design a predictive task to train a backbone model by exploiting the homology between the raw frame and the corresponding optical flow. Instead of constructing a simple task that may lead to trivial solutions [42, 17], our task is to predict several future frames and optical flows simultaneously. This simple yet effective strategy helps the model learn more semantic and dynamic information rather than static visual features. Second, to achieve a domain-invariant model, we design a pretraining method that leverages synthetic abnormal samples and minimizes mutual information between normal and abnormal samples. Instead of using some unrelated images as abnormal frames, we generate random noise to represent anomalies such as unseen objects, and we shuffle frame sequences to simulate erratic and sudden motions, akin to action anomalies. We further utilize mutual information to regulate latent embeddings. Finally, we design an adaptation model that aims to transfer our pre-trained model to the target domain using only few-shot target samples. We recognize that few-shot target samples may not adequately represent the underlying data distribution; therefore, directly fine-tuning the pre-trained model with these samples is not effective. To address this issue, we mix source samples with few-shot target samples to bridge the two domains and adopt adversarial learning to achieve domain-invariant latent embeddings.

In summary, our major contributions are as follows: (1) We empirically demonstrate that a domain gap can lead to performance degradation in video anomaly detection, and formally introduce the few-shot cross-domain video anomaly detection (FC-VAD) problem. (2) We propose a self-supervised learning method to pre-train a generalized model by utilizing data augmentation and mutual information. We further propose the Ada-VAD framework that adapts the model to target domain with few samples.(3) Extensive experiments on three datasets and cross-domain evaluations validate the effectiveness of our methods.

2 Related Work

Video Anomaly Detection. Video anomaly detection is a typical outlier detection task. It has been studied for a long time and a lot of solutions [36, 30, 46, 29, 4, 5, 2, 8, 7, 19, 42, 35, 22] have been proposed to address the problem. Although video anomaly detection develops some new settings like weakly supervised video anomaly detection [33, 15, 39, 6], supervised video anomaly detection [16, 31] and online video anomaly detection [14], the most realistic and popular setting is unsupervised or one-class video anomaly detection. Existing video anomaly detection methods can be roughly classified into four categories. They are density based approaches [13, 34, 27], reconstruction based approaches [42, 26, 32, 2, 42, 19], classification based approaches [8], and self-supervised learning approaches [35, 7, 3, 28]. Density based approaches aim to quantify the normal distribution, so that those abnormal samples which are deviated from the normal distribution can be recognized. Classification based approaches convert outlier detection problem to a binary classification problem. For example, Georgescu et al. [8] construct background-agnostic abnormal samples to let the model distinguish between normal and abnormal samples. Self-supervised learning approaches utilize some pre-text tasks to differentiate normal and abnormal samples. For example, Wang et al. [35] design a jigsaw quiz with an assumption that model can solve the puzzle in normal videos but has a high probability of failing in abnormal samples. Reconstruction based approaches are widely used. They have the intuition that the model that is trained only on normal videos can overfit normal samples, so the reconstruction error will be very high when applying to abnormal videos. Our proposed method falls into this category. However, we notice that reconstruction based approaches could learn a trivial solution especially when training data are not sufficient or has a serious homogenization issue. We address the problem by increasing the difficulty of prediction task and learning cross-modal information.

Few-shot Domain Adaptation. Few-shot domain adaptation is a special setting of domain adaptation [48, 47], and its main challenge lies in that data in two domains are sampled from different distributions, and there are only a few samples from the target domain. Particularly, few-shot domain adaptation fits the video anomaly detection task because, in real-world situation, sufficient training data from the target domain are always very hard to collect due to time and difficulties in labeling. Although a handful of works have also pointed this issue, this direction is still largely underexplored. Sun et al. [31] study the problem in a supervised video anomaly detection task. They use a meta-learning technique to transfer the model to the target domain. Our work is different from their work for the following reasons. First, we focus on a more realistic setting, which is unsupervised video anomaly detection. Second, we achieve knowledge transfer from a completely different perspective. Scene-agnostic methods [8, 22, 5] also try to solve the insufficient data issue by learning a scene-insensitive model. For example, Georgescu et al. [8] leverage background-agnostic abnormal samples that can be applied to any scenes such that the model can distinguish between normal and abnormal for every scene. Lu et al. [22] use the meta-learning approach to make the reconstruction model easily adapt to various scenes. However, our Ada-VAD is largely different from them for two key reasons. Previous work, typically meta-learning, needs a great amount of training data from multiple domains. However, our work can be applied to either small datasets (e.g., UCSD Ped2 [24]) or large datasets (e.g., ShanghaiTech [23]). Furthermore, our method has two stages that can learn a domain generalized model and then adapt to the target domain using few-shot target samples.

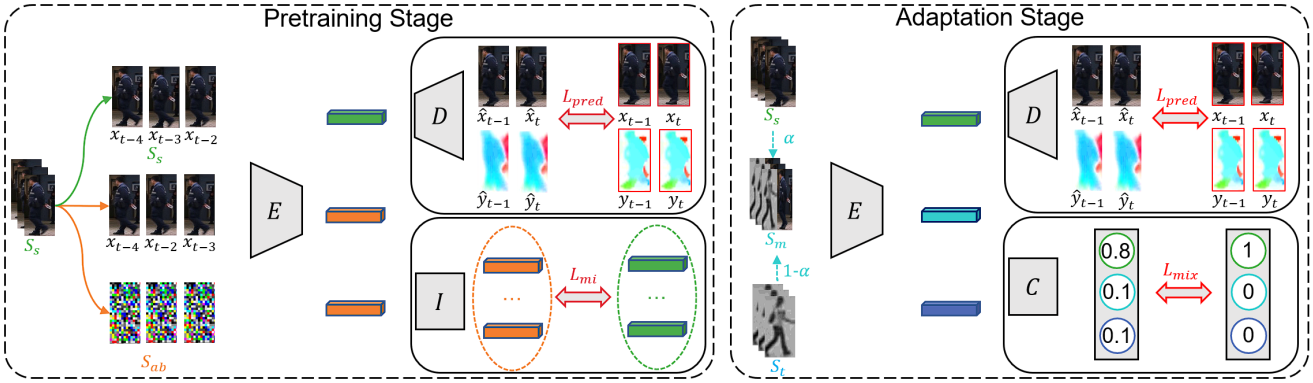


Figure 1: Overview of our Ada-VAD framework. Ada-VAD framework mainly consists of two key parts, namely pretraining stage and adaptation stage. In the pretraining stage, we force the model to predict well on source normal samples, and meanwhile, we synthesize abnormal samples to separate normal embeddings and abnormal embeddings by minimizing their mutual information. In the adaptation stage, we first generate mixup samples by mixing source samples and target samples in a ratio α. Then we optimize the model to predict well on all three kinds of training samples and restrict their embeddings in a similar feature space in an adversarial way, such that the model can be well adapted to target domain.

Table 1: Comparison of anomaly detection performance (AUC scores) between different SOTA methods with different training domains. Target domains are shown in bold. The best performance occurs always when training and testing are in the same domain.

	Frame Pred [17]	HF-VAD [19]	Jigsaw [35]
ped-**ped**	**95.39**	**99.30**	**98.89**
ave-**ped**	86.96	90.84	81.98
shtech-**ped**	89.49	85.52	95.25
ave-**ave**	**85.06**	**91.10**	**92.12**
ped-**ave**	83.97	83.65	74.72
shtech-**ave**	82.25	88.36	80.42

3 Motivation and Problem Definition

In this section, we first discuss the drawbacks of existing video anomaly detection models. Then we introduce the *few-shot cross domain video anomaly detection* problem.

Self-supervised learning methods are widely used in unsupervised video anomaly detection methods. All of these methods have an intuition that training on normal data would lead to high performance on normal data, whereas they have poor performance on abnormal data. Thus, based on the performance gap, they can identify abnormal frames in test videos. For example, the prediction-based network $f_\theta(\cdot)$ tries to predict the future frame x_t given a sequence of previous frames $x_{1:t-1}$. The difference between the predicted frame $\hat{x}_t$ and the ground truth frame x_t will indicate whether one is normal or not. Since during training, the model $f_\theta(\cdot)$ can only access normal training data, it can perform well in predicting future normal frames, but it is hard to predict abnormal frames.

In practice, a critical issue is that it is very hard to collect and label a large amount of video data in every test scenario. Moreover, existing work cannot perform well when the training data are insufficient. It is also impossible to adapt their model across datasets (i.e., domains), since the model will simply regard target domain as abnormal samples. We evaluate three state-of-the-art video anomaly detection methods (i.e., Frame Pred [17], HF-VAD [19] and Jigsaw [35]) under different settings on three datasets, including UCSD Ped2 (Ped) [24], CUHK Avenue (Ave) [20], and ShanghaiTech Campus (Shtech) [23]. The results are shown in Table 1. Specifically, in order to find out whether domain shift would cause a performance drop, we train them on different training datasets and test them on each target dataset. We have two observations from the results. First, existing models achieve best performance when the training and test videos are from the same domain. Second, for cross-domain evaluation, the performance in the target domain is dramatically deteriorated. For example, the Jigsaw method [35] loses 17% in the Avenue-to-Ped scenario and 18% in the Ped-to-Avenue scenario.

Based on these observations, we can draw the following conclusions. First, domain gaps exist in various video datasets, and such a gap impedes most of existing video anomaly detection methods and causes significant performance drop. In addition, existing works are unable to deal with the domain gap issue when the target training videos are insufficient. We believe that it is because existing work mainly focuses on overfitting training video patterns while sacrificing the generalization ability. Specifically, a model is unlikely to learn an accurate normal pattern due to the lack of target training samples.

Although it is costly to collect particular training videos

Algorithm 1 Training of Ada-VAD Framework

Input: S_s, S_t

Output: $f_\theta(\cdot), f_\phi(\cdot)$, which predicts future frames and optical flows, respectively

{Pretraining stage}
Synthesize abnormal samples S_{ab} given S_s
Initialize $f_\theta(\cdot), f_\phi(\cdot), I_\theta(\cdot), I_\phi(\cdot)$
Optimize $I_\theta(\cdot), I_\phi(\cdot)$ by loss (4.3) given S_s, S_{ab}
Optimize $f_\theta(\cdot), f_\phi(\cdot)$ by loss (4.4) given S_s

{Adaptation stage}
Synthesize mix-up samples S_m with ratio α given S_s, S_t.
Load pre-trained model $f_\theta(\cdot), f_\phi(\cdot)$
Initialize category classifier $C_\theta(\cdot), C_\phi(\cdot)$ for embeddings generated by $f_\theta(\cdot), f_\phi(\cdot)$ respectively
Optimize $C_\theta(\cdot), C_\phi(\cdot)$ by loss (4.6)
Optimize $f_\theta(\cdot), f_\phi(\cdot)$ by loss (4.7) given S_s, S_t, S_m
return $f_\theta(\cdot), f_\phi(\cdot)$

for a single domain, it is reasonable to capture a few normal frames in that domain, leading to a few-shot learning problem. In addition, because normal events are almost the same in surveillance videos, like human walking, the abundant normal samples in the source domain can provide normal information which could be adopted for pretraining a base model for video anomaly detection. Furthermore, by exploiting the few-shot frames in target domain, we can adapt the model pre-trained in the source domain to the target domain. In this paper, we address the challenge of insufficient training data by cross-domain learning, i.e. leveraging sufficient single source domain training samples and adapting the model to target domain. Formally, we can define the few-shot cross-domain video anomaly detection as follows:

Few-shot Cross-domain Video Anomaly Detection (FC-VAD). *Given training samples S_s from a single source domain and few-shot target training samples S_t. Learn a model $f(\cdot)$ that can detect anomalies in the target domain test samples, S'_t, i.e.,*

$$F(S_s, S_t) = f(\cdot), \tag{3.1}$$

where F is the training method to learn an anomaly detector model $f(\cdot)$.

4 Methodology

4.1 Overview. As illustrated in Figure 1, our Ada-VAD framework consists of two stages, that is, the pretraining stage and the adaptation stage. We first use the off-the-shelf object detector [1] and the optical flow generator [12] to generate spacial and temporal cubes (STC) and the optical flow cubes (OFC), respectively. During pretraining stage, our objective is to learn a prediction model that can not only predict future frames, but also separate latent embeddings of normal and abnormal samples. After the pretraining stage, we will pass the prediction model forward to the adaptation stage for model fine-tuning. In this stage, we use source samples, few-shot target samples, and our generated mixup samples to further improve the prediction ability on target domain samples. Meanwhile, we adopt an adversarial learning approach that assists the model to learn domain-invariant embeddings. Afterwards, the prediction model can be adapted from the source domain to target domain.

In the following sections, we introduce the details of prediction model, pretraining stage, and adaptation stage. Finally, we show how to infer anomalies in test scenarios.

4.2 Future Predictor. Predicting future frames or optical flow [17, 2] is a popular approach in video anomaly detection. It tries to predict the future frame x_t given previous frames $x_{1:t-1}$, i.e. $p(x_t|x_{1:t-1})$, or predict the future optical flow (as an auxiliary branch), i.e. $p(y_t|y_{1-t})$, where y represents optical flow.

However, we find that such a prediction task is too easy for a deep learning model to fit and may easily lead to a trivial solution. We observe that training videos are usually at 30 Hz, which means that there is a very short gap between $x_{1:t-1}$ and x_t. Thus, the trained models tend to false alarm the normal videos with obvious actions but ignore anomalies with small movements. We solve this issue by two methods. First, to avoid the trivial solution, we increase the difficulty of the prediction task. Specifically, model $f_\theta(\cdot)$ is required to predict two frames, i.e., $\hat{x}_{t-1,t} = f_\theta(x_{1:t-2})$. In this way, the model is encouraged to learn more temporal information, which also makes the trivial solution hard to converge. Second, to help the model learn more temporal and spatial information instead of focusing on the foreground of video frames, we create a new task that predicts future optical flow directly from their raw frames, i.e., $\hat{y}_{t-1,t} = f_\phi(x_{1:t-2})$. Since the optical flow is generated from raw frames, it is reasonable to assume that $x_{1:t-2}$ has enough information to generate $\hat{y}_{t-1,t}$. This task can facilitate learning semantic and temporal information.

We adopt the U-Net [18] as our prediction model $f(\cdot)$, which has an encoder E and a decoder D, i.e., $f(\cdot) = D(E(\cdot))$. Future object frames prediction and future optical flow prediction have their own prediction model $f_\theta(\cdot)$ and $f_\phi(\cdot)$, respectively. $H^x = E_\theta(x_{1:t-2})$ represents the embeddings to predict the raw frame. $H^y = E_\phi(y_{1:t-2})$ represents the embeddings to predict the future optical flow.

Finally, based on the two tasks, we have the following prediction loss:

$$L_{pred} = ||\hat{x}_{i-1,i} - x_{i-1,i}||^2 + ||\hat{y}_{i-1,i} - y_{i-1,i}||^2, \tag{4.2}$$

where $\hat{x}$ and x refer to the predicted frame and the raw frame,

respectively. $\hat{y}$ and y refer to the predicted optical flow and the ground truth optical flow, respectively.

4.3 Pretraining Stage. In our pretraining stage, we design two modules to distinguish between normal sample and anomaly samples.

First, we synthesize some scene-agnostic anomalies because the abnormal samples can help the model build the negative pairs and help the generator produce semantic embeddings. Inspired by [8], We generated Gaussian noise with the same input dimension as the noisy negative samples. Because using noise as negative sample is too easy for model to identify and will lead to degradation, beyond that, we regard shuffled frames as anomaly since shuffled frames always show odd motions and are intuitive anomaly by human cognition. Our goal is not to synthesize realistic anomalies, but to leverage noisy frames that are dissimilar to normal frames to better estimate the distributions of normal frames. Consequently, our model is able to detect anomalies based on the new insights on normal data distributions.

In addition, we employ a mutual information based criterion [11] to enlarge the distance of embeddings between normal and abnormal samples. Within source normal samples, we aim to enlarge their mutual information. Between normal and abnormal samples, we aim to separate their latent embeddings. Like [11], we use a neural network to estimate the lower bound of two distributions. In our framework, we use two mutual information estimators, $I_\theta(H_i^x, H_j^x)$ and $I_\phi(H_i^y, H_j^y)$, to predict mutual information in frame embeddings and optical flow embeddings, respectively. We try to maximize mutual information among normal samples, while minimizing mutual information between normal and abnormal samples, such that the latent embeddings of normal and abnormal samples will be well separated. In summary, our mutual information loss can be formulated as:

$$L_{MI} = \frac{1}{N} \sum_{i,j \in S_s, k \in S_{ab}} \Big(- \frac{I_\theta(H_i^x, H_j^x) - I_\theta(H_i^x, H_k^x)}{I_\theta(H_i^x, H_j^x) + I_\theta(H_i^x, H_k^x)} - \frac{I_\phi(H_i^y, H_j^y) - I_\phi(H_i^y, H_k^y)}{I_\phi(H_i^y, H_j^y) + I_\phi(H_i^y, H_k^y)} \Big), \tag{4.3}$$

where S_s is the normal sample set in source domain, and S_{ab} is the synthetic abnormal sample set. $I(H_i, H_j)$ represents the mutual information between two source normal samples, and $I(H_i, H_k)$ represents the mutual information between a source normal sample and a synthetic abnormal sample, and N is the number of total combination pairs. Finally, our loss in the pretraining stage can be written as:

$$L_{pretrain} = L_{pred} + \lambda_1 \cdot L_{MI}, \tag{4.4}$$

where λ_1 is the hyperparameter for adjusting the loss of mutual information and the loss of prediction.

4.4 Adaptation Stage. Our adaptation stage is used for adapting the pre-trained model to target domain. We observe that if we directly use few-shot target training samples to fine-tune the model, the performance do not improve. This is because few-shot samples can barley represent the target data distribution. Thus, we adopt the mixup technique [44] which is an effective data augmentation method. In our setting, we apply it to video sequences. We design a dynamic mixup mechanism with its corresponding mixup ratio prediction network. The dynamic mix-up can be formulated as:

$$S_m = \alpha \cdot S_s + (1 - \alpha) \cdot S_t, 0 \leq \alpha \leq 1, \tag{4.5}$$

where S_m represents the mix-up samples, and α represents the mix-up ratio. In this way, it bridges the data distribution from the source domain distribution to the target distribution, so the future prediction model can be transferred to the target domain in a steady way. The mix-up samples and target samples are all normal samples and will feed into our future predictor.

Futhermore, we regard α as a soft label, which will be converted to represent three categories, including the source sample class ($\alpha = 0$), mix-up sample class ($0 \leq \alpha \leq 1$), and target sample class ($\alpha = 1$). An MLP based classifier $C(\cdot)$ is trained to predict which category an embedding belongs to, and meanwhile, the encoder E tries to fool the category classifier to make a wrong classification. This adversarial training process facilitates the model to generate domain-invariant embeddings that can help generate target domain future frames based on the pre-trained model. This minimax game of adversarial training can be written as:

$$\begin{aligned} \min_E \max_C L_{mix}(E, C) = {} & \mathbb{E}_{h \sim p(h|s \in S_s)} \log(C(h)) \\ & + \mathbb{E}_{h \sim p(h|s \in S_m)} \log(C(h)) \\ & + \mathbb{E}_{h \sim p(h|s \in S_t)} \log(C(h)), \end{aligned} \tag{4.6}$$

where $h \sim p(h|s \in S)$ represents sampling an instance within a particular category S.

Combining the two strategies, our loss for the Adapation Stage can be formulate as:

$$L_{ada} = L_{pred} - \lambda_2 \cdot L_{mix}. \tag{4.7}$$

Finally, the overall loss function of our Ada-VAD framework is written as:

$$L = L_{pretrain} + L_{ada}. \tag{4.8}$$

4.5 Anomaly Detection. The training process is summarized in Algorithm 1. At the test stage, we use f_θ and f_ϕ to calculate anomaly scores based on: (1) future object frame prediction error T_x, and (2) future optical flow prediction error T_y. In detail, we calculate the anomaly score by using the weighted summation of two prediction errors as:

$$T = w_x \cdot \frac{T_x - \mu_x}{\sigma_x} + w_y \cdot \frac{T_y - \mu_y}{\sigma_y}, \tag{4.9}$$

Table 2: Comparison of anomaly detection performance among our backbone architecture and existing state-of-the-art in cross-domain setting (i.e., **without** accessing target domain training samples). We report the AUC (%) of different methods on UCSD Ped2 (Ped), CUHK Avenue (Ave) and Shanghai Tech (Shtech) datasets. Target domains are shown in bold. The best and the second best results are shown in **bold** and underline, respectively.

	Ave-**Ped**	Shtech-**Ped**	Shtech-**Ave**	Ped-**Ave**	Ave-**Shtech**	Ped-**Shtech**
Frame Pred. [17]	86.96	89.49	82.28	**83.97**	73.76	71.72
HF-VAD [19]	<u>90.84</u>	85.52	**88.36**	<u>83.65</u>	76.44	74.88
Background [8]	87.00	90.60	83.60	-	76.30	-
Jigsaw [35]	81.98	<u>95.25</u>	80.42	74.72	**79.43**	68.86
Pretraining Stage (Ours)	**98.41**	**97.53**	<u>83.62</u>	82.08	<u>76.77</u>	**75.03**

where $\mu_x, \sigma_x, \mu_y, \sigma_y$ are the means and standard deviations of the object frame prediction errors and the optical flow prediction errors. w_x and w_y are weights of two errors. Notably, the statistics of $\mu_x, \sigma_x, \mu_y, \sigma_y$ are computed from source training videos only.

5 Experiments

5.1 Datasets. We conduct experiments on three popular benchmarks. (1). **UCSD Ped2 (Ped)** [24]. Ped2 contains 16 training videos and 12 test videos. They have a resolution of 240 $\times$ 360 pixels in gray scale. (2). **CUHK Avenue (Ave)** [20]. Avenue has 16 training videos and 21 testing videos with 47 abnormal events. They have a resolution of 360 $\times$ 640 RGB pixels. (3). **ShanghaiTech Campus (Shtech)** [23]. Shanghaitech dataset have 330 training videos and 107 testing videos captured by 13 different cameras. It has 107 kinds of anomalies. Each video has a resolution of 480 $\times$ 856 RGB pixels.

Table 3: Comparison of our pretraining stage model with SOTA baselines in single domain video anomaly detection setting. We calculate AUC(%) on three benchmark datasets. The best and the second best results are shown in **bold** and underline, respectively.

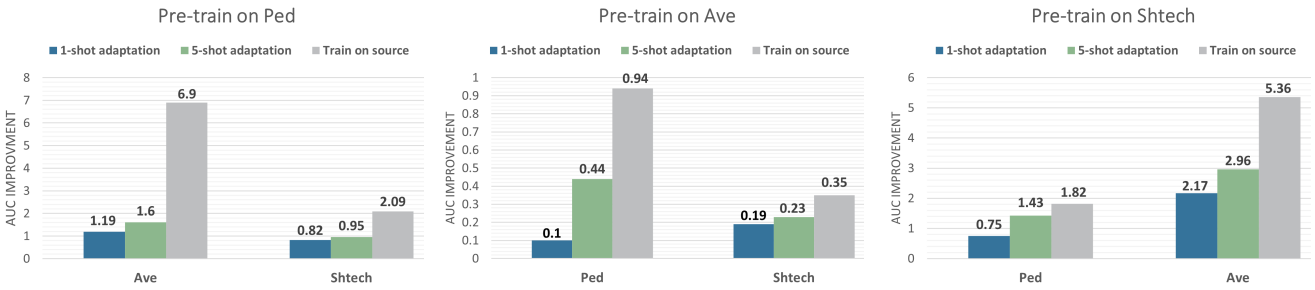
Method	**Ped2**	**Ave**	**Shtech**
MNAD-R [26]	90.2	82.8	69.8
Mem-AE [9]	94.1	83.3	71.2
Conv-VRNN [21]	96.1	85.8	-
MNAD-P [26]	97.0	88.5	70.5
ST-AE [45]	91.2	80.9	-
AMC [25]	96.2	86.9	-
ANO-PCN [41]	96.8	86.2	73.6
VEC [42]	97.3	90.2	74.8
HF-VAD [19]	<u>99.3</u>	<u>91.1</u>	76.2
Jigsaw [35]	98.89	**92.12**	**84.24**
Pretraining Stage (Ours)	**99.35**	88.98	<u>77.12</u>

5.2 Results and Discussions

5.2.1 Cross Domain Results. To validate the effectiveness of the pretraining stage in the proposed Ada-VAD framework, we test our method in a cross-domain manner. To have a fair comparison, we compare our pretraining stage model with other baselines in a setting in which all methods can only access source training samples for model training. Our model is trained on one source dataset and tested on the other two target datasets separately. For example, if Avenue is used as the training dataset, then Ped and Shtech are the target test datasets. In this manner, we conduct experiments in 6 different settings, which cover all possible combinations of source and target domains. We show the results in Table 2. From the table, we observe that our pre-trained model outperforms the SOTA methods such as Jigsaw in most cases, and it obtains the second-best result in the Shtech-Ave setting. Generally, the performance of our method is quite consistent, and it achieves comparable performance than the SOTA baselines. According to the results, we can claim that the pretraining stage of our Ada-VAD framework can produce a model with better generalization ability.

5.2.2 Few-shot Domain Adaptation Results. In the adaptation stage of our Ada-VAD framework, we can further utilize the few-shot target training samples to improve model training. Intuitively, with the introduction of target domain, the performance can be further improved. We test on three datasets and evaluate on all 6 cross-domain settings, which covers all possible combinations. We also conduct experiments on our pretraining stage model, 1-shot adaptation stage model, 5-shot adaptation stage model, and our pretraining stage model in the single-domain setting (i.e., training and test videos are from the same dataset). The results are the AUC improvement based on the pretraining stage model. We separately use each dataset as the source domain, regard the other two datasets as target domain datasets, and test on them independently.

Results are shown in Figure 2. From the three figures, we have several observations and inferences. (1) Training and test only on source domain samples (i.e., the single domain setting) have the most performance improvement



(a) Source domain is Ped dataset. Target domains are Ave and Shtech datasets.

(b) Source domain is Ave dataset. Target domains are Ped and Shtech datasets.

(c) Source domain is Shtech dataset. Target domains are Ped and Ave datasets.

Figure 2: We compared the AUC score improvement of our pretraining stage model with the 1-shot adaptation model, the 5-shot adaptation model, and the single-domain model. We test on three datasets and evaluate on all 6 cross-domain settings, which covers all possible combinations. The results are consistent. Training in the source domain can achieve the highest performance, and with the number of target samples increasing, the adaptation stage model can improve performance.

and can reach the highest performance. It is reasonable since without domain gap model can easily find normal pattern. It proves the effectiveness of our pretraining stage model on single domain scenarios. (2) The consistent trend of results shows that the performance of 1-shot adaptation stage model is better than the pretraining stage model. It is because 1-shot target sample and synthetic mixup samples can better adapt the pretraining stage model to target domain. This observation also validates the effectiveness of the adaptation stage in our Ada-VAD framework. (3) The performance of 5-shot adaptation stage model is better than the 1-shot based model. It is reasonable because with more target samples come in, the adaptation model can have a better estimation of target domain distributions. However, we notice that the improvement from 1-shot to 5-shot is slightly smaller than that from pre-trained model to 1-shot model. We infer that giving 1-shot target sample is sufficient to roughly construct target distribution so given more target samples can only refine the target distribution which leads to marginal improvements. But still the performance is getting closer to the pretraining stage model trained on the same domain as testing. It also proves the effectiveness of our adaptation stage. Thus, with the consistent and comprehensive results on three benchmark datasets, we can claim that our Ada-VAD framework can effectively address the FC-VAD problem.

5.2.3 Source Domain Evaluation. We also conduct experiments on the traditional single domain video anomaly detection setting and compare our method with the SOTA approaches [26, 9, 21, 45, 25, 41, 42, 19, 35]. We train our model on each of the three datasets and test it on the same dataset. The results on three benchmark datasets are shown in Table 3. As can be seen, our pretraining stage model achieves comparable performance to the SOTA methods. For example, our model achieves the best results on the

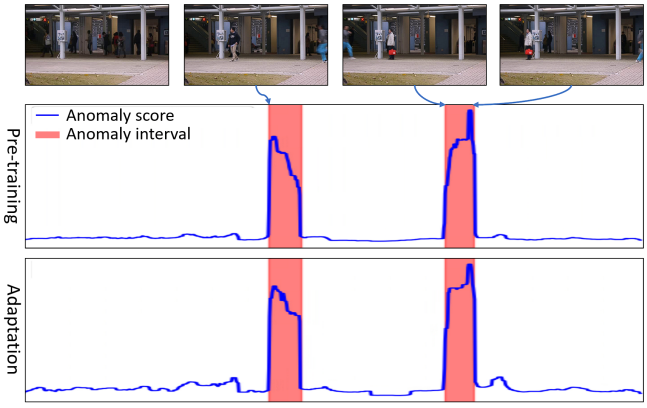


Figure 3: An example of anomaly detection curve on Avenue dataset. From top to bottom, we show the sampled video frames, our pretraining stage results, adaptation stage results. The pretraining stage model is trained and tested on Avenue, while adaptation stage result is trained on Ped and tested on Avenue. Larger values in curve indicate higher possibility to be anomaly.

Ped dataset. Since our model focus more on the generalization ability, the results on single domain evaluation is acceptable. In summary, we can conclude that our method achieves anomaly detection ability as well as the domain generalization ability.

5.2.4 Qualitative Analysis. We give an example of evaluation on the Avenue dataset in Figure 3. The anomaly curve indicates the anomaly scores of all test frames of a video sequentially. All anomaly scores are calculated and fused in the same way. Particularly, the pretraining stage model is trained and tested on the same domain, but the adaptation stage model is trained on the Ped dataset and 1-shot Avenue training sample, and then tested on the Avenue dataset. As

can be seen, although there are some fluctuations in normal interval, our adaptation model can achieve almost the same results as the single-domain pretraining model. It means that even with only 1-shot target sample, our Ada-VAD model can still adapt well to target domain and achieve comparable performance with the single-domain model.

6 Conclusions

In this paper, we study a realistic issue, i.e., insufficient training data, and present a new setting named few-shot cross-domain VAD (FC-VAD). To address this issue, we design a novel framework named Ada-VAD that improves the domain generalization ability and can adapt the model to target domain with few-shot samples. Experimental results fully support our assumptions and validate the effectiveness of our framework compared with the state-of-the-art methods.

7 Acknowledgement

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-23-1-0290.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[2] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. Comprehensive regularization in a bi-directional predictive network for video anomaly detection. 36:230–238, Jun. 2022.

[3] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European conference on computer vision*, pages 334–349. Springer, 2016.

[4] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 934–935, 2020.

[5] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 254–255, 2020.

[6] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. 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, 2021.

[7] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12742–12752, June 2021.

[8] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[9] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 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, 2019.

[10] Zihan Guan, Mengxuan Hu, Zhongliang Zhou, Jielu Zhang, Sheng Li, and Ninghao Liu. Badsam: Exploring security vulnerabilities of sam via backdoor attacks. *arXiv preprint arXiv:2305.03289*, 2023.

[11] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[13] Christophe Leys, Olivier Klein, Yves Dominicy, and Christophe Ley. Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of experimental social psychology*, 74:150–156, 2018.

[14] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li. Video anomaly detection with compact feature sets for online performance. *IEEE Transactions on Image Processing*, 26(7):3463–3478, 2017.

[15] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *Proceedings of the AAAI, Virtual*, 24, 2022.

[16] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1490–1499, 2019.

[17] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[18] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[19] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13588–13597, October 2021.

[20] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[21] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and

Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16Th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[22] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *European Conference on Computer Vision*, pages 125–141. Springer, 2020.

[23] Weixin Luo, Wen Liu, and Shenghua Gao. 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, 2017.

[24] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[25] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019.

[26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[27] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2598–2607, 2020.

[28] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13576–13586, 2022.

[29] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2188–2196. PMLR, 2021.

[30] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[31] Guangyu Sun, Zhang Liu, Lianggong Wen, Jing Shi, and Chenliang Xu. Anomaly crossing: A new method for video anomaly detection as cross-domain few-shot learning, 2021.

[32] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020.

[33] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021.

[34] Melissa Turcotte, Juston Moore, Nick Heard, and Aaron McPhall. Poisson factorization for peer-based anomaly detection. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 208–210. IEEE, 2016.

[35] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, 2022.

[36] Yizhou Wang, Dongliang Guo, Sheng Li, and Yun Fu. Towards explainable visual anomaly detection. *arXiv preprint arXiv:2302.06670*, 2023.

[37] Yizhou Wang, Can Qin, Yue Bai, Yi Xu, Xu Ma, and Yun Fu. Making reconstruction-based method great again for video anomaly detection. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1215–1220. IEEE, 2022.

[38] Yizhou Wang, Ruiyi Zhang, Haoliang Wang, Uttaran Bhattacharya, Yun Fu, and Gang Wu. Vaquita: Enhancing alignment in llm-assisted video understanding. *arXiv preprint arXiv:2312.02310*, 2023.

[39] Jie Wu, Wei Zhang, Guanbin Li, Wenhao Wu, Xiao Tan, Yingying Li, Errui Ding, and Liang Lin. Weakly-supervised spatio-temporal anomaly detection in surveillance video. *arXiv preprint arXiv:2108.03825*, 2021.

[40] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[41] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813, 2019.

[42] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020.

[43] M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14724–14734, 2022.

[44] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[45] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017.

[46] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019.

[47] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[48] Ronghang Zhu, Dongliang Guo, Daiqing Qi, Zhixuan Chu, Xiang Yu, and Sheng Li. Trustworthy representation learning across domains. *arXiv preprint arXiv:2308.12315*, 2023.