Anomaly Crossing: New Horizons for Video Anomaly Detection as Cross-domain Few-shot Learning

Guangyu Sun[1*]   Zhang Liu[2*]   Lianggong Wen[2]   Jing Shi[1]   Chenliang Xu[1]

[1]University of Rochester   [2]Corning Inc.

gsun6@ur.rochester.edu, {j.shi,chenliang.xu}@rochester.edu

{LiuZ2, WenLB}@corning.com

Developer Category *Abstract*—Video anomaly detection aims to identify abnormal events that occur in videos. Since anomalous events are relatively rare, it is not feasible to collect a balanced dataset and train a binary classifier to solve the task. Thus, most previous approaches learn only from normal videos using unsupervised or semi-supervised methods. Obviously, they are limited in capturing and utilizing discriminative abnormal characteristics, which leads to compromised anomaly detection performance. In this paper, to address this issue, we propose a new learning paradigm by making full use of both normal and abnormal videos for video anomaly detection. In particular, we formulate a new learning task: cross-domain few-shot anomaly detection, which can transfer knowledge learned from numerous videos in the source domain to help solve few-shot abnormality detection in the target domain. Concretely, we leverage self-supervised training on the target normal videos to reduce the domain gap and devise a meta context perception module to explore the video context of the event in the few-shot setting. Our experiments show that our method significantly outperforms baseline methods on DoTA and UCF-Crime datasets, and the new task contributes to a more practical training paradigm for anomaly detection.

Index Terms—Anomaly Detection, Self-supervised Learning, Few-shot Learning, Domain Adaptation.

I. INTRODUCTION

Video anomaly detection [12], [32] has broad application potential in security [73], industry [4], and healthcare [16]. The goal is to identify if there is an abnormal event that happened in a given video. This goal can be formulated as a binary classification problem. However, one unique property of anomaly detection is that the distributions of normal and anomalous events are quite different, leading to a severe data imbalance. Training a naive binary classifier on such an imbalanced dataset will always steer the classifier to give a negative prediction, thus failing to detect the anomaly (Fig. 1 top). To tackle unbalanced data in video anomaly detection, different training paradigms are proposed. An intuitive solution is to re-sample or down-sample to get an equal amount of normal and abnormal samples. However, naive re-sampling or down-sampling is insufficient for the classifier to learn all anomaly patterns under extreme data imbalance. Another solution is to collect more anomalous data and construct a
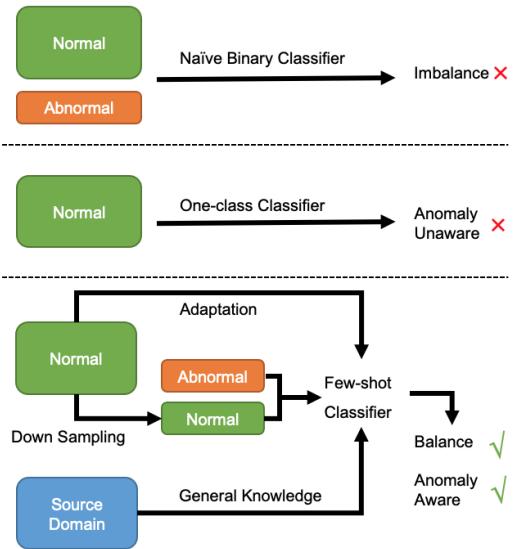

Fig. 1: Different training paradigms for video anomaly detection: Training a naive binary classifier will fail due to the severe data imbalance. Most existing methods focus on training a one-class classifier based on only normal samples, which is anomaly unaware. To leverage the abnormal samples to learn an anomaly-aware classifier in a balanced dataset, we propose a new pipeline to train a few-shot classifier.

balanced dataset [26], [29], [42], [45], [51], [73]. However, it is not possible to collect sufficient anomalous data in many real-world applications, *e.g.*, industrial production pipeline, as the defective rate can be low and the shutdown of machines suffers monetary loss.

Another line of work completely discards the abnormal data and uses only normal training samples since anomalous events are rare in practice, formulating the problem as single-class classification [14], [32] (Fig. 1 middle). Such a paradigm tries to approximate the distribution of normal samples and regards the out-of-distribution samples as anomalous; however, not all normal behaviors can be observed. Thus, some normal events may deviate from the approximated distribution, triggering false anomalies.

These two lines of works are either infeasible or neglect the discriminative abnormal characteristics, leading to com-

* Equal Contribution

promised anomaly detection performance. Therefore, a new training paradigm is needed to construct a balanced dataset and get an anomaly-aware classifier for anomaly detection.

We propose a new task *cross-domain few-shot video anomaly detection* (CD-FSVAD) for anomaly detection (Fig. 1 bottom). To address the insufficiency of abnormal samples, we resort to few-shot learning [18], [57] to better utilize the limited abnormal samples. However, recent few-shot learning largely relies on extensive annotated data for meta-learning, where base classes are sampled from the same domain as the novel classes [20], [57], clearly infeasible for anomaly detection. On the other hand, general knowledge may be feasibly gained from outside the in-domain data. Inspired by cross-domain few-shot learning (CD-FSL) [25], we expect to learn such general knowledge from another large-scale dataset as the source domain. In this new task, we can leverage a large amount of source-domain videos to help target-domain anomaly detection with only a few abnormal videos and abundant normal videos, which is a more realistic setting for traffic accident detection, surveillance, industrial production pipeline, *etc.*.

This new task is challenging owing to the large gap between the source and the target domain and the data imbalance between normal and abnormal samples. To tackle these challenges in CD-FSVAD, we devise a novel baseline for anomaly detection called *Anomaly Crossing*, where knowledge will be learned from the source domain, adapted from the normal samples, and finally fit the target domain through only a few abnormal samples. Two novel modules, Domain Adaptation Module (DAM) and Meta Context Perception Module (MCPM), are proposed in our pipeline. To reduce the domain gap, we devise the DAM that can use a large amount of target-domain normal videos. Then, inspired by the power of self-supervised learning (SSL) [3], given the backbone trained from source-domain videos, we fine-tune it via SSL using normal videos in the target domain to achieve the unsupervised domain adaptation. Furthermore, the video temporal context is crucial for anomaly detection, and we also expect to enable the model with adaptive context modeling ability in different novel scenes. However, the events in different scenes may need different temporal context modeling; for example, the pedestrian and vehicle surveillance has different motion patterns; thus, they need different context modeling. Therefore, the model should automatically adapt its contextual modeling to novel events. To achieve this goal, we propose the MCPM consisting of a learnable Graph Convolution Network (GCN) to perceive the meta-context that can be adapted to the target domain under a few-shot setting. We evaluate the performance of our pipeline on two different target domains with diverse scenes: a traffic dataset DoTA [66] and a surveillance dataset UCF-Crime [51]. Our pipeline consistently outperforms the comparison methods on both datasets, achieving 15% higher accuracy than the best-compared ones on the DoTA dataset.

The contributions of this paper can be summarized as follows. Firstly, we introduce a more practical CD-FSVAD task to address the extreme data imbalance issue in anomaly detection and propose a novel pipeline Anomaly Crossing as a new training paradigm to tackle this task. Secondly,

we propose to leverage video self-supervised learning tasks on the target-domain normal samples to reduce the gap between source and target domain. Thirdly, we enable our Meta Context Perception Module to adapt its contextual modeling to different scenes automatically. Finally, Anomaly Crossing consistently outperforms the comparison methods on DoTA and UCF-Crime datasets.

II. RELATED WORK

Anomaly Detection. Anomaly detection is a challenging problem that has been studied for years [13], [18], [34], [36], [40], [53], [68], [72], [74]. Methods out of different perspectives are proposed. Among them the mainstream methods completely discard the anomalous data and try to learn a representation for only normal events due to the rarity of abnormal events and the intrinsic data imbalance. Then,are construction loss is used as anomaly score to identify abnormal events [17], [39], [68], [70], [74] or a one-class classier is used instead [14], [32], [50]. Another stream of works design models for video prediction learnings and detect anomaly based on difference between the predicted frame and the observed frame [47], [47], [71]. All these methods suffer limited accuracy especially for subtle abnormal cases which are similar with normal cases due to the ignorance of patterns in abnormal samples. However, such kind of cases are quite common in many applications such as in industry. Furthermore, these methods are lack of the capability to evolve when more abnormal samples are collected later on. To resolve these issues, several recent works try to leverage abnormal samples following weekly supervised setting [13], [19], [38], [40], [53]. However, information in abnormal samples is still not fully exploited due to the limitation of weekly supervised learning. Different with these methods we formalize anomaly detection as a few-shot classification problem under supervised learning setting. Meanwhile a Domain Adaptation Module (DAM) is designed to adapt knowledge learned from a source domain dataset (large) to the target domain (small) by leveraging excessive normal samples. Thus, information in both abnormal and normal samples is fully exploited, and a deep neural network (R(2+1)D) backbone is possible to be used to extract better spatial-temporal representations of video clips instead of shallow networks used in most of the existing works. Finally, a Meta Context Perception Module (MCPM) is designed to enhance the understanding of scene context which is critical for VAD especially for cases with complicated behaviors and patterns instead of a simple reconstruction loss or a FC classification head.

Few-shot Video Classification. Few-shot Video Classification. Few-shot learning is usually achieved by meta-training with a large amount of labeled data [20], [48], [58]. Based on such data, most of the methods are metric-based to extract a generalized video metric [41]. These metrics can be obtained by pooling [5], [10], [11], [35], [62], [75], adaptive fusion [8], [21], dynamic images [52], or attention [7], [69]. Significantly, with the help of progress in video representation learning, few-shot learning can be achieved even without meta-training. For example, after training a video encoder(e.g., 3D-CNN,
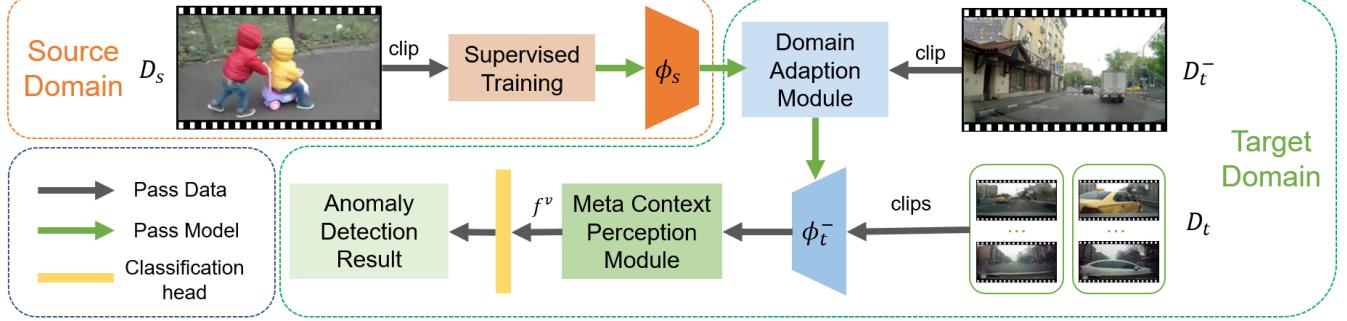
Fig. 2: The pipeline of Anomaly Crossing. The left of the first row shows a supervised training process to get the encoder ϕ_s in the source domain. The right of the first row shows the Domain Adaption Module (DAM) of the normal samples in the target domain to get the adapted encoder ϕ_t^-. The second row shows the meta-testing process in the target domain. Clips of each video in the support set are fed into ϕ_t^-, and the clip features will be passed to the Meta Context Perception Module (MCPM) to get the video feature f^v. Anomaly detection result will be output by a classification head given f^v. Blue and green arrows mean passing the data or entire model into the module.

R(2+1)D-CNN) on largescale data with supervision, few-shot learning can be achieved by simply learning a classifier using the few shots [62]. Other researchers focus on the feature extraction from video using Graph Neural Network (GNN) to get a better feature representation of the video [27] or further predict augmented video features [28]. However, in CD-FSVAD, the training dataset on the source domain has an extensive domain gap with the target domain. Therefore, based on our experiments, training a classifier on the target domain straightforwardly will not work well as before. As a result, our method includes two modules to mitigate the domain gap and achieve few-shot video classification in a cross-domain fashion.

Cross-domain Few-shot Classification. Cross-domain Few-shot Classification. CD-FSL is recently proposed mainly in the image domain [15], [25], [64]. The goal is to strengthen the few-shot model generalizability when the base classes are sampled from the different domains as the novel classes. Most existing methods focus on the meta training process in the source domain, including ensemble modeling [37], meta fine-tuning [9], transductive multi-head model [30], representation fusion [1], and learned feature-wised transformation [54]. These methods are domain-agnostic so that they neglect the information on the target domain. Meanwhile, [46] is a concurrent work sharing the most similar spirit with our methods to leverage SSL on the target domain, which requires many unlabeled target-domain training data. In addition, however, we explore it in the anomaly detection task and research the more complex video modality since the video domain CD-FSL is nearly unexplored. [22] concentrates on CD-FSL for action recognition problems, which does not leverage the large-amount normal sample on the target domain as our method for domain adaptation. Compared with all the existing work, our method takes both source domain and target domain into considerations with the large-amount normal samples, while fills the blank of the study of CD-FSL on the video domain.

III. METHOD

Considering a common video anomaly detection task, *e.g.*, the traffic accident detection, there are sufficient normal samples but only a few abnormal samples available. Therefore, the goal is to learn a model to classify whether an input video contains an anomaly event or not. A naive binary classification model trained on all samples will fail due to extreme data imbalance. Down-sampling is a common way [2] for data balance, but simple down-sampling neglects the considerable normal data. To learn from fewer samples, we further propose to regard anomaly detection as a few-shot detection problem, formulated as a 2-way-K-shot classification problem. However, typical few-shot learning [20], [48], [58] requires a large dataset for meta-training whose classes are in the same domain of novel classes, which is infeasible for our anomaly detection problem since there are insufficient abnormal videos to support meta-training. To confront the few-shot challenge, we proposed to inflate our training data from a different domain where a large amount of videos are available, *e.g.*, human action videos [24], [31]. Nonetheless, directly applying the standard meta-learning approach on a target-independent domain is incapable of solving the target-domain problem; hence, we propose a new pipeline, Anomaly Crossing (Section III-B), as a baseline for this new formulation. Our Domain Adaptation Module (Section III-C) leverages the normal samples to mitigate the domain gap via self-supervised contrastive learning. Furthermore, concerning the impact of video context for anomaly detection, we propose a Meta Context Perception Module (Section III-D) to perceive the context-aware video feature.

A. Problem Formulation

In the perspective of few-shot learning (FSL), the video anomaly detection task can be formulated as follows. Here we refer to the anomaly detection dataset as the target domain and the other dataset (*e.g.* human action) as the source domain. Given a target-domain dataset of sufficient normal samples and K abnormal samples, we denote all normal samples as a sub-dataset D_t^-. Together with the abnormal samples, we down-

---

Algorithm 1 Training of the Anomaly Crossing Pipeline

1: **Training on the source domain:**
Learn a clip-level encoder ϕ_s on D_s.

2: **Domain Adaptation Module:**
Perform adaptation $\mathcal{A}$ on D_t^- and get a refined clip-level encoder ϕ_t^-.

3: **Meta Context Perception Module:** 1
  1) Sample n clips from each video $x_t \in X_t$ and extract corresponding clip-level features $f^c = \{\phi_t^-(x_{t,n}),$ for x_t in $X_t\}$ from ϕ_t^-.
  2) Extract corresponding video-level feature f^v of f^c from MCPM $\mathcal{P}$.

4: **Training:**
Pass f^v to the classification head c_t and train the parameters in MCPM and c_t by the loss between predicted anomaly detection result $\hat{Y}_t$ and ground truth Y_t in D_t.

---

sample D_t^- to the same size K to construct a 2-way-K-shot dataset $D_t = \{(x_t, y_t) \in X_t \times Y_t\}$ for few-shot learning. To enable the FSL, we include a source-domain dataset D_s composed of sufficient labeled samples $(x_s, y_s) \in X_s \times Y_s$ to learn general knowledge for classification. This setting corresponds to the cross-domain few-shot learning (CD-FSL), where the D_s and D_t have a large domain gap.

As a result, the Cross-Domain Few-shot Anomaly Detection (CD-FSVAD) task is formulated to learn a good classifier $X_t \rightarrow Y_t$ based on D_s, D_t^- and D_t.

B. The Anomaly Crossing Pipeline

Our Anomaly Crossing pipeline leverages both normal and abnormal samples on the target domain. First, as [62] indicates that meta-training is not a necessity if a video encoder is trained on large-scale supervised data, we also directly train a source-domain video clip encoder for a good initialization of the model backbone. Then, normal samples are leveraged for anomaly detection in a new horizon that adapts the backbone learned from source domain to the target domain instead of training from scratch. Finally, the down-sampled normal samples paired with the few abnormal samples are used to train only the classification head, as known as a meta-testing stage. The pipeline is shown in Fig. 2 and described in Algorithm 1.

Training on the source domain. We conduct supervised training on the source domain to learn a video encoder ϕ_s entailing the general knowledge for video classification. This process can be formalized as:

$$\min_{\phi_s, c_s} \mathcal{L}_s(c_s(\phi_s(X_s)), Y_s) \ , \tag{1}$$

where $\mathcal{L}_s$ is the loss function for supervised training on the source domain, and c_s is the classification head. An alternative is meta-training instead of standard training; however, our experiments show that these two training fashions are comparable in our cross-domain setting (Section IV-B). Despite that we learn from source-domain in a supervised way following the insight of [62], unsupervised learning fashion is also feasible for future exploration.

Fig. 3: Construction of samples for self-supervised learning. A spatial disturbance will be performed for a positive sample to gain a clip with a similar motion but a dissimilar scene. A temporal disturbance will be performed for the negative sample to gain a clip with a similar scene but dissimilar motion.

Domain adaptation. Domain adaptation enables the knowledge learned from the source domain to better assist the target domain for anomaly detection. Typical domain adaptation approaches require a large quantity of target domain samples for training [23], [44], which is infeasible for our case where our target domain D_t only contains few-shot abnormal samples. However, apart from D_t, we have numerous normal samples D_t^- in the target domain, which carries rich domain information that might benefit domain adaptation, which is a unique property for the anomaly detection task. Therefore, we devise a novel Domain Adaptation Module (DAM) to adapt the backbone parameter of ϕ_s to ϕ_t^- based on the normal samples D_t^-, so as to adapt representations on the source domain to the target domain. This process can be formalized as:

$$\phi_t^- = \mathcal{A}(\phi_s, D_t^-) \ , \tag{2}$$

where $\mathcal{A}$ refers to the DAM.

Meta-testing on the target domain. Despite that ϕ_t^- involves the target domain information, it is unaware of the abnormal information. Therefore, we need a further fine-tuning of the model using D_t to capture the abnormal pattern, referred to as the meta-testing step in FSL. However, the typical meta-testing stage just adapts the last classification head with the frozen backbone, which is commonly a fully connected (FC) layer [6]. However, a simple tuning on the FC layer may not well capture the temporal dynamics of the abnormal pattern, which is critical for detecting video anomalies such as abnormal vehicle speed. Instead, we further leverage a Meta Context Perception Module (MCPM) that can refine the input video clip features into a spatio-temporal contextualized video feature, and the whole parameter of MCPM are tuned in this stage, as shown in Eq. (3). The parameter in MCPM ϕ_t constructs a mapping from X_t to $f^v(X_t)$], as shown in Eq. (4). Therefore, the objective of this process is as shown in Eq. (5).

$$f^c(x_t) = \{\phi_t^-(x_{t,i})\}, \ \ i = 1, 2, \ldots n \ , \tag{3}$$

$$\phi_t(X_t) = f^v(X_t) = \mathcal{P}(f^c(X_t)) \ , \tag{4}$$

$$\min_{\phi_t, c_t} \mathcal{L}_t(c_t(\phi_t(X_t)), Y_t) \ , \tag{5}$$

where $\mathcal{P}$ refers to the MCPM, $x_{t,i}$ refers i_{th} sampled clip in video $x_t \in X_t$, $\mathcal{L}_t$ is the loss function for classification on target domain, and c_t is the classification head.

C. Domain Adaptation Module

The goal of the Domain Adaption Module (DAM) is to adapt the learned knowledge from the source domain to the target domain. Since we only use the large number of normal samples in the target domain for adaptation, it is an unsupervised domain adaptation. Despite no label is available, recent progress on self-supervised learning [59], [65] on large scale video dataset has exhibited excellent representation and generalization capabilities. Inspired by this, a most straightforward idea is to re-train the ϕ_s on D_t^- using state-of-the-art self-supervised training. However, such an idea is intuitively risky because the knowledge learned from D_t^- might overwrite the modeling ability of ϕ_s due to well-known catastrophic forgetting [33]. Nonetheless, we find it works well in practice, and the experiment shows that the forgetting phenomenon does not occur (Sec. IV-C2). Therefore, we select a recent SOTA self-supervised learning algorithm [59] as our DAM. Note that other self-supervised approaches might also work, and we leave the selection as future work. The details for the DAM are as follows.

Sample Selections. Given a video clip c_0, we randomly crop three video clips c_1, c_2, c_3, and we apply different transformations to them to build the data triplet containing anchor, positive, and negative samples for the later contrastive learning. *Anchor*: We apply basic augmentations including random rotation, random cropping, color jittering to c_1 to get the anchor sample a. *Positive*: Every frame in the clip c_2 is applied with the same random warped to get the positive sample p, as shown in the right of Fig. 3. *Negative*: We randomly shift the starting time of c_3 while keeping its duration to construct the negative sample n, as shown in the left of Fig. 3.

Objective Function. Contrastive learning is employed to learn the representations by enhancing the affinity between the anchor and positive sample and the dissimilarity between the anchor and negative sample. Specifically, feed a, p, n into ϕ_s, we will get the clip feature z_a, z_p, z_n and the final InfoNCE [55] object function is written as:

$$\mathcal{L}_c = -\log \sum_{i=1}^{N} \frac{\exp(z_{a_i} \cdot z_{p_i})}{sim(z_{a_i}, z_{p_i}, z_{n_i}) + \sum_{j=0}^{K} \exp(z_{a_i} \cdot z_{a_j})} \ ,$$

where N is the numbers of video in D_t^-, K is the number of other samples, and $sim(z_{a_i}, z_{p_i}, z_{n_i}) = \exp(z_{a_i} \cdot z_{p_i}) + \exp(z_{a_i} \cdot z_{n_i})$. As a result, ϕ_s will be adapted to ϕ_t^- via DAM.

D. Meta Context Perception Module

The goal of MCPM is to update video clip features given D_t to get an updated video feature f^v entailing temporal information to better assist the anomaly detection on the target domain. Considering Graph Convolutions Networks (GCN) are widely used to capture the temporal information of video [60], [63], we apply GCN in MCPM. Meanwhile, the MCPM should also have the ability to be adapted to novel domains by a few shots of abnormal and normal samples.

In addition, we desire to handle both temporal and spatial information to serve anomaly detection better. Therefore, a Semantic-Temporal Graph Convolution Network
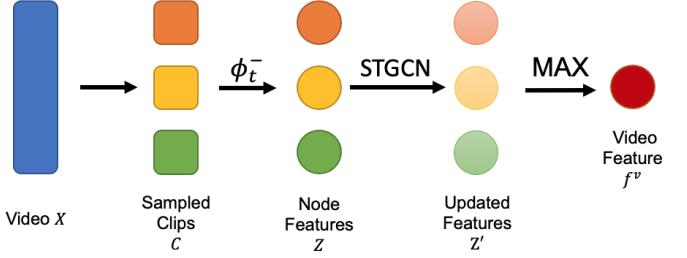

Fig. 4: Extract context-aware video feature in Meta Context Perception Module on the target domain. Sampled video clips will be fed into ϕ_t^- then updated as node features in STGCN. Finally, the video feature will be given by the max of all node features.

(STGCN) [63] is chosen to capture the spatio-temporal information.

1) Construction of STGCN: STGCN takes the clip feature as nodes and builds temporal edges $\mathcal{E}_t$ and semantic edges $\mathcal{E}_s$ based on temporal ordering and node similarity, respectively. For the video clip features fed into STGCN, we will sample L clips $C = \{c_1, c_2, \ldots, c_L\}$ with a fixed stride, then get the video clip features $z_l = \phi_t^-(c_i)$ by the video clip encoder ϕ_t^-. Then we will build a video graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{z_l\}_{l=1}^{L}$ and $\mathcal{E} = \mathcal{E}_t \cup \mathcal{E}_s$. The overview of this module is exposed in Fig. 4 and the details of the edges construction, the feature updating, and training of STGCN are described as follows:

Semantic Edges $\mathcal{E}_s$. Semantic edges are designed to connect nodes with similar semantic information. Each node is connected to its k nearest neighbors according to their feature distances.

Temporal Edges $\mathcal{E}_t$. Temporal edges are designed to connect adjacent nodes in temporal order. A forward edge and a backward will be built for each node except the first node (no backward edge) and the last node (no forward edge).

Node Feature Updating. Taking into consideration that temporal edges are in a linear structure, we use three 1d-convolution layers to update the node features from the temporal edges Z^t. To update the node feature from the semantic edges Z^s, we use a single-layer edge convolution [61] $Z^s = ([Z^T, AZ^T - Z^T]W)^T$ as our graph convolution operation. $Z = \{z_1, z_2, \ldots, z_L\}$ are the features of all nodes in the graph, W is the trainable weight of three 2d-convolution layers, A is the adjacency matrix of the graph without self-loops, and $[\cdot, \cdot]$ represents the matrix concatenation of columns. The final updated node feature is given by a fusion of Z, Z^t and Z^s: $Z' = \mathrm{ReLU}(Z + Z^t + Z^s)$.

E. Training of STGCN

With the updated node features Z', the video feature $f^v = \max(Z')$ will be computed as the maximum of all node features. The feature will be classified by a classification head c_t to get the predicted class, as shown in the second row of Fig. 2. Based on the 2-way K-shot dataset D_t, we use the cross-

TABLE I: **2-way 5-shot result of different anomaly types in DoTA.** ST refers to "Start-stop or stationary", AH refers to "Moving ahead or waiting" *etc.*. The full explanation of the abbreviation is shown in Appendix B. The result under each type means the meta-testing set will be constructed only by this type. For example, "All Types" means the meta-testing set will be constructed among all anomaly types.

Methods	AH VP	ST OO-LEFT	LA OO-RIGHT	TC VO	OC UK	All Types
Deep SVDD [49]	0.49 0.50	0.55 0.55	0.49 0.52	0.53 0.56	0.57 0.53	0.52
Liu *et al.* [37]	0.66±0.09 0.65±0.09	0.55±0.09 0.75±0.09	0.61±0.10 0.73±0.08	0.67±0.09 0.64±0.10	0.66±0.10 0.59±0.09	0.66±0.10
Xian *et al.* [62]	0.66±0.10 0.65±0.09	0.53±0.08 0.72±0.09	0.61±0.10 0.70±0.10	0.65±0.10 0.60±0.09	0.63±0.10 0.59±0.10	0.65±0.09
Ours	**0.84±0.07** **0.79±0.07**	**0.75±0.09** **0.84±0.07**	**0.80±0.08** **0.80±0.09**	**0.82±0.08** **0.82±0.09**	**0.83±0.07** **0.69±0.10**	**0.81±0.08**

TABLE II: 2-way 5-shot result of different anomaly types in UCF-Crime. A full table with all different types is in Appendix B.

Methods	Arson Road Accidents	Burglary Shooting	Explosion Shoplifting	All Types
Deep SVDD [49]	0.59 0.57	0.55 0.52	0.59 0.63	0.54
Liu *et al.* [37]	0.70±0.09 0.61±0.11	0.59±0.10 0.56±0.10	0.65±0.10 0.65±0.10	0.56±0.09
Xian *et al.* [62]	0.80±0.08 0.68±0.09	0.62±0.08 0.60±0.09	0.73±0.09 0.70±0.09	0.62±0.10
Ours	**0.86±0.07** **0.75±0.09**	**0.69±0.09** **0.72±0.09**	**0.78±0.09** **0.77±0.09**	**0.66±0.10**

entropy loss to train W in STGCN and the classification head c_t.

IV. EXPERIMENT

A. Experimental Setting

Datasets. IG-65M is a large dataset collected from Instagram, which contains more than 65 million videos. Pretrained backbones are released instead of the dataset itself. Therefore, We choose IG-65M [24] as our source domain. To test our method in different task types with diverse scenes, we choose a traffic dataset (Detection of Traffic Anomaly (DoTA) [66]) and a surveillance dataset (UCF-Crime [51]) as our target domains: Detection of Traffic Anomaly (DoTA) is a traffic dataset containing 4,677 videos with temporal, spatial, and categorical annotations. Since the anomaly start and the end time are annotated, we capture the frames before the anomaly start time as the normal sample while capturing the frames between the anomaly start and end time as the abnormal sample. UCF-Crime consists of 1,900 long and untrimmed real-world surveillance videos, with 13 real anomalies such as fighting, road accident, burglary, robbery, and normal activities. In order to get enough samples for evaluation, we augmented UCF-Crime by annotating more temporal segmentation.

Networks and Baselines. We use 34-layer R(2+1)D IG-65M pretrained backbone as our feature encoder on the source domain. The clip length is 8, and the frame size is 224×224. For each method, training will be based on this pretrained backbone instead of training from scratch. We choose a one-class classification method Deep SVDD [49], a few-show classification method [62] and a cross-domain few-shot classification method on image domain [37] as our baselines. Deep SVDD [49] leverage only normal videos to learn a hypersphere in feature space built by ϕ_s. Liu *et al.* [37]

TABLE III: Evaluation on vary number of shots

DAM	MCPM	5 shot	10 shot	15 shot	20 shot
✓	✗	0.76±0.09	0.78±0.08	0.79±0.07	0.81±0.07
✓	✓	0.81±0.08	0.83±0.07	0.84±0.07	0.86±0.07

applied meta-training on the source domain with batch spectral regularization. On the other hand, Xian *et al.* [62] assumes that a model pretrained in a large and general dataset can get adequate knowledge to achieve few-shot learning by training a linear classifier without updating the model's parameters in the test phase. We choose 3DFSV as an implementation of this method.

Domain adaptation module settings. For the self-supervised learning task, we set the batch size as 44, the size of the memory bank as 4200, the learning rate as 0.003, and train the network for 200 epochs. For DoTA and UCF-Crime, 4136 and 950 samples with lengths more than 20 frames are used for DAM correspondingly.

Meta context perception module settings. The STGCN is implemented following the structure GCNeXt as in Xu *et al.* [63]. The video clip length is 8 frames, and the stride size is 4 frames. The minimum video length is set as 20 frames, and the maximum length is 124 frames. The graph output dimension is 32 with 4 paths. The graph is trained with a learning rate of 0.001 and a batch size of 4 for 60 epochs.

Evaluation settings. We will employ a 2-way-5-shot-15-query meta-testing on the target domain based on the support set S_t and the query set Q_t, which means there are 2×5 samples in the support set for training and 2×15 samples in the query set for evaluation in each iteration of testing. Since the performance is sensitive to the sampling of S_t and Q_t, we report the average accuracy and standard deviation among 200 iterations of testing with randomly sampled S_t and Q_T. The performance on the meta-test set sampled from all different anomaly types will be the primary metric for evaluation. In addition, to explore the performance for a specific anomaly type, we build additional meta-test sets sampled from each anomaly type. We used the same abbreviation of anomaly types as [66] for DoTA. Note that for Deep SVDD, we did not employ the meta-testing, so the result is among the entire test set; therefore, the standard deviation will not be reported here.

B. Evaluation Result

The evaluation result on DoTA and UCF-Crime is shown in Table I and Table II. Among all of the results, Deep SVDD

[49] only leads to the lowest performance, indicating that simply leveraging normal videos as semi-supervised training is inadequate. Liu *et al.* [37] allows the meta-training in the source domain, which does not show a significant advantage over the supervised classification pretrained on source domain as Xian *et al.* [62] on DoTA and get even worse performance on UCF-Crime. Therefore, our method simply uses the classification pretrained on the source domain and adapts to the normal samples, leading to much better performance. In our method, we gain general knowledge from the source domain, and adapt the learned knowledge by the normal samples, and fit to the target domain by leveraging the abnormal samples. As a result, anomaly detection is implemented in a cross-domain fashion.

Compared with DoTA, the result of UCF-Crime is less effective. Here are some possible explanations. At the data level, the number of normal samples for the DAM in UCF-Crime is less, but the scenes are more diverse than DoTA, which influenced the efficiency of the DAM. Besides, in UCF-Crime, the anomaly sometimes happens in a small scene region, which will not cause enough global temporal and spatial information disturbance. It will be hard to detect, which might be a limitation of our methods even if it still outperforms the comparison methods.

C. Ablation Study

1) Impact of Different Shots: The ability of growth with more shots provided is an important metric to evaluate a method for few-shot learning. Therefore, we test the performance of our method in the different stages with a different number of shots provided in the support set. The result is shown in Fig. III. With more shots provided, the performance of our method will increase as expected.

2) Impact of Different Modules: In order to discover the impact of each module, we test the performance of all anomaly types without specific modules on DoTA. A brief result is shown in Table IV, and the full results of the ablation study are shown in Appendix C.

Pretrain on Source Domain. Without the pretraining on the source domain, the performance will extremely drop. One possible explanation is that the DAM will be lack corresponding general knowledge learned from the source domain; therefore, even two modules are applied, the performance will drop, which emphasizes the feasibility of cross-domain task.

Domain Adaptation Module. The accuracy is improved after adaptation. It shows that our DA module managed to transfer the adapted domain knowledge from the source domain to the target domain.

Meta Context Perception Module. The performance with only MCPM for all anomaly types is not improved compared with the pretrained model, but the performance for specific anomaly types is improved. (The results are included in Appendix C.1). However, if the representation for modeling the meta context is adapted, the learned meta context will benefit the anomaly detection thus get better performance compared with DAM only model. Meanwhile, compared with the video feature computed directly by the maximum of all

TABLE IV: Evaluation on impacts of each module

Source Domain	Domain Adaptation	Meta Context Perception	Accuracy
✓	✗	✗	0.65±0.09
✗	✓	✓	0.54±0.11
✓	✓	✗	0.76±0.09
✓	✗	✓	0.65±0.10
✓	✓	✓	**0.81±0.08**

Fig. 5: t-SNE plots for features in each stage of Anomaly Crossing under a 5-shot setting on DoTA. Blue and red points refer to the normal and abnormal samples correspondingly.

video clip features, MCPM also gets the better performance, as shown in Appendix C.2, which proves that MCPM can extract better video features under the few-shot setting.

V. DISCUSSION

A. Why do the modules work?

The classification accuracy is highly dependent on the grouping of features on the target domain. We hypothesize that our modules can adapt the model into a better representation, i.e., the grouping of features is more separable. To verify that our modules give a better representation of the target domain, we plot the t-SNE [56] of test data to visualize the features exacted by encoders in each stage. The plot is shown in Fig. 5 for the result among all anomaly types in DoTA. The remaining result for each specific anomaly type and the results on UCF-Crime is shown in Appendix D. The features extracted by ϕ_s are less separable, which verifies that simply supervised training cannot give a good representation when there is a domain gap. After DAM, the encoder ϕ_t^- gains the knowledge from the normal samples, leading to more separable features. However, ϕ_t^- does not gain knowledge directly from the abnormal samples. In MCPM, the video encoder ϕ_t leverages as few as 5 shots and temporal information on the target domain to further get more separable features, verifying our hypothesis.

B. Extreme situations for DAM

In reality, some more extreme situations exist that there will not even be sufficient normal samples (*e.g.*, Medical). A solution is to use a similar dataset without annotations to substitute the normal samples for DAM. We want to evaluate the robustness of our method under such extreme situations, so we test the performance using a different but similar dataset BDD100K [67], which is a driving video dataset with 100K videos, to substitute the normal samples on DoTA. We will use 10000 samples in BDD100K for DAM. The result is shown in Table V. The performance of DAM only model on BDD100K can still yield a pleasing outcome. However, the performance after adding MCPM is worse than directly using the normal

TABLE V: Evaluation on an extreme situation that normal samples are insufficient

DAM Dataset	Domain Adaptation	Meta Context Perception	Accuracy
BDD100K	✓	✗	0.75±0.04
BDD100K	✓	✓	0.77±0.09
DoTA	✓	✗	0.76±0.09
DoTA	✓	✓	0.81±0.08

samples on the target domain. One possible explanation is that the video context will be more brutal to perceive with the substitute adaption to gain the same performance for video clips, but it will be more complicated to learn the meta context based on domain knowledge learned from a similar dataset.

C. Potential Negative Impacts and Limitation

Potential Negative Societal Impacts. If this work is applied to real scenarios, being too confident about the result might cause unpredictable losses (*e.g.*, a malfunction of a crucial machine were not correctly detected). To mitigate such negative impacts, please consider the result rationally and take precautions against inaccurate results.

Limitations. The performance of our method is sensitive to the sampling of the few shots on the target domain. We simply perform a multi-time sampling to mitigate the impact of different sampling, which will slightly differ between the statistical and the real performance.

VI. CONCLUSION

We propose the Cross-domain Few-shot Anomaly Detection task to address real-world problems. We devise a new pipeline – Anomaly Crossing – to handle video anomaly detection and prove the effectiveness of self-supervised learning and context perception in this task. Anomaly Crossing outperforms existing methods significantly in DoTA and UCF-Crime datasets under two different settings. This task proves that the knowledge from different domains and tasks, e.g., action recognition, can be transferred to the current task, which builds a bridge from different tasks and enhances the significance of video representation learning by expanding the downstreaming tasks.

REFERENCES

[1] Thomas Adler, Johannes Brandstetter, Michael Widrich, Andreas Mayr, David Kreil, Michael Kopp, Günter Klambauer, and Sepp Hochreiter. Cross-domain few-shot learning by representation fusion, 2021.

[2] Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*. 2017.

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

[4] Deegan J Atha and Mohammad R Jahanshahi. Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Structural Health Monitoring*, 2018.

[5] Rami Ben-Ari, Mor Shpigel, Ophir Azulai, Udi Barzelay, and Daniel Rotman. Taen: Temporal aware embedding network for few-shot action recognition. *arXiv preprint arXiv:2004.10141*, 2020.

[6] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[7] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.

[8] Yang Bo, Yangdi Lu, and Wenbo He. Few-shot learning of video action recognition only based on video contents. In *WACV*, 2020.

[9] John Cai and Sheng Mei Shen. Cross-domain few-shot learning with meta fine-tuning. *arXiv preprint arXiv:2005.10544*, 2020.

[10] Congqi Cao, Yajuan Li, Qinyi Lv, Peng Wang, and Yanning Zhang. Few-shot action recognition with implicit temporal alignment and pair similarity optimization. *arXiv preprint arXiv:2010.06215*, 2020.

[11] Chris Careaga, Brian Hutchinson, Nathan Hodas, and Lawrence Phillips. Metric-based few-shot learning for video action recognition. *arXiv preprint arXiv:1909.09602*, 2019.

[12] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[13] Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia*, 24:4067–4076, 2022.

[14] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *CVPR*, 2015.

[15] Ziqiu Chi, Zhe Wang, Mengping Yang, Dongdong Li, and Wenli Du. Learning to capture the query distribution for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4163–4173, 2022.

[16] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017.

[17] Zhiwen Fang, Joey Tianyi Zhou, Yang Xiao, Yanan Li, and Feng Yang. Multi-encoder towards effective anomaly detection in videos. *IEEE Transactions on Multimedia*, 23:4106–4116, 2021.

[18] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 2006.

[19] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection, 2021.

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. PMLR, 2017.

[21] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *MM*, 2020.

[22] Zan Gao, Leming Guo, Weili Guan, An-An Liu, Tongwei Ren, and Shengyong Chen. A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2. *IEEE Transactions on Image Processing*, 2020.

[23] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, 2017.

[24] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019.

[25] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *ECCV*. Springer, 2020.

[26] Chengkun He, Jie Shao, and Jiayu Sun. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools and Applications*, 2018.

[27] Yufan Hu, Junyu Gao, and Changsheng Xu. Learning dual-pooling graph neural networks for few-shot video classification. *IEEE Transactions on Multimedia*, 23:4285–4296, 2021.

[28] Yufan Hu, Junyu Gao, and Changsheng Xu. Learning scene-aware spatio-temporal gnns for few-shot early action prediction. *IEEE Transactions on Multimedia*, pages 1–1, 2022.

[29] Jing Huo, Yang Gao, Wanqi Yang, and Hujun Yin. Abnormal event detection via multi-instance dictionary learning. In *International conference on intelligent data engineering and automated learning*, 2012.

[30] Jianan Jiang, Zhenpeng Li, Yuhong Guo, and Jieping Ye. A transductive multi-head model for cross-domain few-shot learning. *arXiv preprint arXiv:2006.11384*, 2020.

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[32] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 2018.

[33] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017.

[34] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*. IEEE,

2009.

[35] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *ICCV Workshops*, 2019.

[36] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[37] Bingyu Liu, Zhen Zhao, Zhenpeng Li, Jianan Jiang, Yuhong Guo, Haifeng Shen, and Jieping Ye. Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. *arXiv preprint arXiv:2005.08463*, 2020.

[38] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, and Shenghua Gao. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, 2019.

[39] Yang Liu, Jing Liu, Jieyu Lin, Mengyang Zhao, and Liang Song. Appearance-motion united auto-encoder framework for video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(5):2498–2502, 2022.

[40] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. Collaborative normality learning framework for weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(5):2508–2512, 2022.

[41] Su Lu, Han-Jia Ye, and De-Chuan Zhan. Few-shot action recognition with compromised metric via optimal transport, 2021.

[42] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 2021.

[43] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020.

[44] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017.

[45] Jaeyoo Park, Junha Kim, and Bohyung Han. Learning to adapt to unseen abnormal activities under weak supervision. In *ACCV*, 2020.

[46] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. *arXiv preprint arXiv:2010.07734*, 2020.

[47] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection, 2020.

[48] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[49] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer Dy and Andreas Krause, editors, *ICML*, 2018.

[50] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection, 2018.

[51] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos, 2019.

[52] Shaoqing Tan and Ruoyu Yang. Learning similarity: Feature-aligning network for few-shot action recognition. In *IJCNN*, 2019.

[53] 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, 2021.

[54] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020.

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[57] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[58] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2016.

[59] Jinpeng Wang, Yuting Gao, Ke Li, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. *arXiv preprint arXiv:2009.05757*, 2020.

[60] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs, 2018.

[61] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 2019.

[62] Yongqin Xian, Bruno Korbar, Matthijs Douze, Bernt Schiele, Zeynep Akata, and Lorenzo Torresani. Generalized many-way few-shot video classification. In *ECCV*, 2020.

[63] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020.

[64] Xuemiao Xu, Hai He, Huaidong Zhang, Yangyang Xu, and Shengfeng He. Unsupervised domain adaptation via importance sampling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4688–4699, 2020.

[65] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervisedspatio-temporal representation learning, 2020.

[66] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. When, where, and what? a new dataset for anomaly detection in driving videos. *arXiv preprint arXiv:2004.03044*, 2020.

[67] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning, 2020.

[68] Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021.

[69] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *ECCV*, 2020.

[70] Sijia Zhang, Maoguo Gong, Yu Xie, A. K. Qin, Hao Li, Yuan Gao, and Yew-Soon Ong. Influence-aware attention networks for anomaly detection in surveillance videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5427–5437, 2022.

[71] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multispace for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3694–3706, 2021.

[72] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 2011.

[73] 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 Conference on Computer Vision and Pattern Recognition*, 2019.

[74] Joey Tianyi Zhou, Le Zhang, Zhiwen Fang, Jiawei Du, Xi Peng, and Yang Xiao. Attention-driven loss for anomaly detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4639–4647, 2020.

[75] Xiatian Zhu, Antoine Toisoul, Juan-Manuel Perez-Rua, Li Zhang, Brais Martinez, and Tao Xiang. Few-shot action recognition with prototype-centered attentive learning. *arXiv preprint arXiv:2101.08085*, 2021.

APPENDIX

This section provides the implementation details for our experiments, including baselines, Domain Adaptation Module, and Meta Context Perception Module, as a supplementary for Section 4.1. **Our codes are released at https://github.com/likeyhnbm/AnomalyCrossing.**

A. Experiment Environment

The experiments are performed on a Linux operation system with 4×Tesla P100 GPUs. If no parameter-updating of the backbone is needed, we will only use one GPU for most cases.

B. Datasets

- DoTA [66]: We capture the frames before the anomaly start time as the normal video and the frames between the anomaly start time and end time as the abnormal video for each video in DoTA. The temporal annotation is gained from the official documents of the dataset. We only select 10 ego-involved anomaly types in DoTA during the evaluation, considering the performance of non-ego types is not as significant as ego-involved ones, and we regard them as our future work.
- UCF-Crime [51]: We use the videos in "normal" category provided in the dataset as the normal videos and capture the abnormal videos based on the temporal annotations in the official documents of the dataset. Considering that the annotated videos in UCF-Crime are not sufficient for evaluation, we annotated more videos to get a more stable evaluation result based on a larger test set.

C. Sampling Details

When extracting frames from a video, the frame sample rate is set as 10. A video clip will contain L frames, where L is set according to the architecture of the video clip encoder. In our default setting, we will sample a video clip to represent the video where it is sampled. Considering the fairness of evaluation when constructing the meta-testing set, for each video in the support set, we will randomly sample a video clip to adapt the model, but for each video in the query set, we will sample the middlemost video clip to give the result.

D. Selection of Backbone

For the supervised training on the source domain, we evaluate the performance of different backbones with different video clip lengths on the target domain to select the best video clip encoder ϕ_s. We build a model with each backbone followed with a classification head and apply a meta-testing on the model to evaluate the performance [1], i.e., freeze the backbone, train the classification head on the support set and test the result on the query set. We choose cosine distance [43] as the classification head, and the frame size is set as 224×224. The result of 2-way 5-shot 15-query for 200 runs is shown in Table VI. The detailed architecture of each backbone is shown as follows:

- IG65M [24]: 34-layer R(2+1)D encoder, pretrained on IG-65M.
- Kinetics [31]: 34-layer R(2+1)D encoder, pretrained on Kinetics.
- XDC [3]: 18-layer R(2+1)D encoder, XDC pretrained on IG-Kinetics [24], [31].

Therefore, we choose IG65M with a video clip length of 8 as our basic video clip encoder ϕ_s for further adaptation in each method.

E. Baselines

1) Deep SVDD [49]: We implement Deep SVDD by adapting a PyTorch implementation [2] to our datasets. For the training, we will sample the middlemost video clip of each normal video. Then, we will feed these video clips as input and ϕ_s as the network to train the Deep SVDD. Then, we will fine-tune the hyperparameter r to get the maximum accuracy on the validation set. We choose a different setting for the validation set based on the amount of annotated samples. For DoTA, we use a copy of the train set as the validation set. For UCF-Crime, we split half of the test set as the validation set. Then, normal and abnormal samples are kept as the same number for the testing to maintain the data balance.

---

[1]Codes are adapted from https://github.com/artest08/LateTemporalModeling3DCNN
[2]https://github.com/lukasruff/Deep-SVDD-PyTorch

TABLE VI: **2-way 5-shot result of different backbones.** The number in the brackets refers to the video clip length.

Backbones	AH VP	ST OO-LEFT	LA OO-RIGHT	TC VO	OC UK	All Types
IG65M(8) [24]	0.66±0.10 **0.65±0.09**	0.53±0.08 **0.72±0.09**	**0.61±0.10** **0.70±0.10**	**0.65±0.10** 0.60±0.09	**0.63±0.10** **0.59±0.10**	**0.65±0.09**
IG65M(16) [24]	**0.66±0.09** 0.60±0.10	**0.57±0.08** 0.71±0.10	0.60±0.09 0.69±0.10	0.63±0.09 0.59±0.09	0.61±0.09 0.56±0.09	0.60±0.09
IG65M(32) [24]	0.63±0.10 0.55±0.10	0.57±0.09 0.72±0.10	0.59±0.09 0.68±0.10	0.61±0.10 0.57±0.09	0.60±0.10 0.54±0.09	0.61±0.09
Kinetics(8) [31]	0.59±0.10 0.67±0.08	0.51±0.08 0.68±0.09	0.57±0.10 0.68±0.10	0.62±0.10 **0.63±0.08**	0.61±0.09 0.57±0.09	0.61±0.10
Kinetics(16) [31]	0.58±0.10 0.59±0.10	0.54±0.09 0.68±0.11	0.56±0.08 0.65±0.11	0.59±0.10 0.59±0.10	0.57±0.10 0.56±0.09	0.59±0.10
Kinetics(32) [31]	0.61±0.10 0.54±0.09	0.54±0.09 0.54±0.13	0.57±0.09 0.65±0.11	0.59±0.10 0.59±0.10	0.58±0.09 0.57±0.09	0.58±0.10
XDC(8) [3]	0.59±0.10 0.53±0.09	0.52±0.09 0.61±0.10	0.55±0.10 0.60±0.10	0.56±0.09 0.52±0.09	0.54±0.09 0.53±0.09	0.56±0.09
XDC(16) [3]	0.64±0.11 0.59±0.09	0.53±0.09 0.69±0.09	0.60±0.10 0.65±0.10	0.63±0.10 0.54±0.10	0.61±0.11 0.57±0.09	0.62±0.11
XDC(32) [3]	0.65±0.10 0.61±0.10	0.57±0.09 0.68±0.10	**0.61±0.10** 0.63±0.10	0.62±0.10 0.61±0.09	0.61±0.10 0.54±0.09	0.60±0.10

This method is not in a few-shot setting, so the evaluation of this method will be slightly different from others. However, the metric we use for evaluation is the classification accuracy on the unseen data, which is consistent with the few-shot setting. Therefore, we believe the evaluation of Deep SVDD can be comparable with others.

2) Liu et al. [37]: We implement Liu *et al.*'s method based on its official implementation [3]. We apply the meta-training with Batch Spectral Regularization (BSR) to the IG-65M pretrained encoder ϕ_s on Kinetics 100 [31] dataset. We tried support vector machine, fully-connected layer, and cosine distance head as the classification head [43]. The overall performance of these three heads are $0.65\pm0.10, 0.66\pm0.10$, and 0.63 ± 0.09. Therefore, we choose a fully-connected layer as the classification head on the target domain.

3) Xian et al. [62]: We implement Xian *et al.*'s 3DFSV by using ϕ_s as the encoder learned from the representation learning stage, and use cosine distance as the classification head on the few-shot learning stage. For the testing stage, we will sample the middlemost video clip feature as the video feature. We also test the performance of use maximum/mean as the feature, as the "Pretrain+MAX/MEAN" row shown in Table VIII, but there are only trivial changes on the performance.

F. Domain Adaptation Module (DAM)

We apply the Decoupling the Scene and the Motion (DSM) [59] on the normal samples as the DAM. The implementation is adapted from the official implementation [4]. The hyperparameters we use is mentioned in the Section 4.1 of the main paper.

G. Meta Context Perception Module (MCPM)

We construct the Semantic-Temporal Graph Convolution Network (STGCN) by a 1D convolution layer activated by ReLU followed by a GCNeXt [63] Block. The implementation of GCNeXt is from the official implementation of Xu *et al.*[5]. We will sample video clips with fixed lengths from the beginning of the video with a stride. The minimum and maximum frame lengths are set to ensure learning a meaningful context under memory limitation. If the video length is less than the minimum frame length, we will sample from the beginning again until we sample enough frames. The hyperparameters we use are mentioned in Section 4.1 of the main paper.

In this section, we introduce each anomaly type in Detection of Traffic Anomaly (DoTA) [66] and UCF-Crime [51] datasets as well as the abbreviation in Table. 1 of the main paper.

We construct $N + 1$ different meta-testing sets to test the robustness of our method, where N is the number of anomaly types. N of the meta-testing sets contains only one specific anomaly type, annotated as the anomaly name, and one contains all 10 different types of anomaly, annotated as "All Types". Considering the difference between different

anomaly types, "All Types" is a more complicated task than other meta-testing sets. The introduction of each anomaly type in DoTA is shown as follows:

- AH: Moving ahead or waiting. Collision with another vehicle moving ahead or waiting, 592 videos in total.
- ST: Start-stop or stationary. Collision with another vehicle that starts stops or is stationary, 66 videos in total.
- LA: Lateral. Collision with another vehicle moving laterally in the same direction, 643 videos in total.
- TC: Turning. Collision with another vehicle that turns into or crosses a road, 1330 videos in total.
- OC: Oncoming. Collision with another oncoming vehicle, 528 videos in total.
- VP: Pedestrian. A collision between vehicle and pedestrian, 52 videos in total.
- OO-left, OO-right: Leaving to left and leaving to right. Out-of-control and leaving the roadway to the left or right, 266 videos in total for left, and 203 videos in total for right.
- VO: Collision with an obstacle in the roadway, 64 videos in total.
- UK: Unknown type, 56 videos in total.
- All Types: Samples containing all types, 3800 videos in total.

The introduction of each anomaly type in UCF-Crime is shown as follows:

- Arson: Burning of property (such as a building), 61 videos in total.
- Burglary: Breaking and entering a dwelling, 142 videos in total.
- Explosion: Bursting forth with sudden violence or noise, 61 videos in total.
- Road Accidents: 164 videos in total.
- Shooting: Gun violence directed toward people, 71 videos in total.
- Shoplifting: Stealing displayed goods from a store, 71 videos in total.
- Abuse: 91 videos in total.
- Arrest: 66 videos in total.
- Assault: 63 videos in total.
- Fighting: 72 videos in total.
- Robbery: 152 videos in total.
- Stealing: 104 videos in total.
- Vandalism: 74 videos in total.
- All Types: Samples containing all types, 1342 videos in total.

The full result for UCF-Crime is shown in Table VII

This section describes the ablation studies in more detail by removing or replacing some components of our pipeline. We provide type-level results and more combinations to support the analysis in Section 4.3.2 of the main paper.

The result is shown in Table VIII and Table IX. Pretrain refers to the supervised training on the source domain, the dataset name following the DAM indicates where the normal samples are from for DAM, and MAX/MEAN means directly using the maximum/mean of video clip features as the video feature to replace the MCPM.

---

[3]https://github.com/liubingyuu/FTEM_BSR_CDFSL

[4]https://github.com/FingerRec/DSM-decoupling-scene-motion

[5]https://github.com/frostinassiky/gtad

TABLE VII: 2-way 5-shot result of different anomaly types in UCF-Crime.

Methods	Arson Road Accidents	Burglary Shooting	Explosion Shoplifting	Abuse Robbery	Arrest Stealing	Assault Vandalism	Fighting	All Types
Deep SVDD [49]	0.59 0.57	0.55 0.52	0.59 0.63	0.54 0.51	0.65 0.47	0.59 0.55	0.47	0.55
Liu *et al.* [37]	0.70±0.09 0.61±0.11	0.59±0.10 0.56±0.10	0.65±0.10 0.65±0.10	0.66±0.10 0.59±0.09	0.59±0.10 0.67±0.10	0.66±0.10 0.64±0.10	0.66±0.10	0.57±0.09
Xian *et al.* [62]	0.80±0.08 0.68±0.09	0.62±0.08 0.60±0.09	0.73±0.09 0.70±0.09	0.75±0.09 0.60±0.10	0.65±0.10 0.71±0.10	0.72±0.09 **0.71±0.09**	0.72±0.10	0.61±0.11
Ours	**0.86±0.07** **0.75±0.09**	**0.69±0.09** **0.72±0.09**	**0.78±0.09** **0.77±0.09**	**0.80±0.08** **0.70±0.10**	**0.67±0.10** **0.76±0.08**	**0.73±0.09** 0.69±0.10	**0.76±0.09**	**0.67±0.09**

TABLE VIII: 2-way 5-shot result of different anomaly types in DoTA.

Modules	AH VP	ST OO-LEFT	LA OO-RIGHT	TC VO	OC UK	All Types
DAM(BDD100K)	0.60±0.10 0.51±0.09	0.55±0.09 0.71±0.09	0.57±0.11 0.68±0.10	0.61±0.10 0.55±0.10	0.61±0.10 0.53±0.09	0.60±0.10
DAM(DoTA)	0.53±0.10 0.49±0.08	0.47±0.08 0.56±0.09	0.52±0.08 0.58±0.10	0.53±0.10 0.49±0.08	0.53±0.09 0.48±0.09	0.53±0.09
DAM(BDD100K)+MCPM	0.61±0.09 0.52±0.09	0.57±0.09 0.71±0.09	0.56±0.10 0.70±0.10	0.61±0.10 0.62±0.09	0.63±0.10 0.49±0.08	0.61±0.09
DAM(DoTA)+MCPM	0.53±0.09 0.49±0.08	0.49±0.09 0.59±0.10	0.52±0.10 0.60±0.10	0.54±0.09 0.53±0.09	0.55±0.09 0.46±0.07	0.54±0.11
Pretrain	0.66±0.10 0.65±0.09	0.53±0.08 0.72±0.09	0.61±0.10 0.70±0.10	0.65±0.10 0.60±0.09	0.63±0.10 0.59±0.10	0.65±0.09
Pretrain+DAM(BDD100K)	0.80±0.03 0.72±0.05	0.70±0.06 0.85±0.01	0.73±0.04 0.83±0.03	0.76±0.03 0.66±0.06	0.75±0.04 0.70±0.07	0.75±0.04
Pretrain+DAM(DoTA)	0.81±0.04 0.73±0.06	0.69±0.06 0.81±0.04	0.73±0.07 0.78±0.06	0.77±0.05 0.67±0.05	0.75±0.05 0.66±0.06	0.75±0.05
Pretrain+MAX	0.67±0.09 0.64±0.10	0.53±0.08 0.71±0.10	0.62±0.10 0.68±0.10	0.67±0.10 0.61±0.10	0.66±0.10 0.58±0.08	0.64±0.10
Pretrain+MEAN	0.66±0.09 0.64±0.10	0.53±0.09 0.71±0.10	0.61±0.10 0.70±0.10	0.67±0.10 0.62±0.09	0.66±0.10 0.59±0.10	0.64±0.10
Pretrain+MCPM	0.67±0.10 0.71±0.09	0.55±0.09 0.70±0.10	0.62±0.10 0.70±0.10	0.68±0.10 0.64±0.11	0.67±0.10 0.58±0.08	0.65±0.10
Pretrain+DAM(DoTA)+MAX	0.82±0.09 0.72±0.09	0.71±0.09 0.80±0.08	0.77±0.09 0.76±0.09	0.80±0.09 0.77±0.10	0.79±0.09 0.66±0.11	0.78±0.09
Pretrain+DAM(DoTA)+MEAN	0.81±0.09 0.74±0.10	0.70±0.09 0.81±0.08	0.75±0.10 0.78±0.08	0.78±0.09 0.74±0.09	0.77±0.09 0.66±0.10	0.76±0.10
Pretrain+DAM(BDD100K)+MCPM	0.83±0.08 0.77±0.09	0.71±0.10 **0.84±0.07**	0.74±0.10 **0.83±0.07**	0.79±0.08 0.73±0.09	0.79±0.07 0.72±0.10	0.77±0.09
Pretrain+DAM(DoTA)+MCPM	**0.84±0.07** **0.79±0.07**	**0.75±0.09** **0.84±0.07**	**0.80±0.08** 0.80±0.09	**0.82±0.08** **0.82±0.09**	**0.83±0.07** **0.69±0.10**	**0.81±0.08**

H. Impact of MCPM for different anomaly types

In the Meta Context Perception Module part of Section 4.3.2 in the main paper, we propose that the performance of MCPM is dependent on the length of the event. The event length in UCF-Crime is longer than events in DoTA. Therefore, as a result, shown in Table IX, the impact of MCPM is larger compared with the performance on DoTA. Meanwhile, the performance for each anomaly type is improved, which proves that MCPM learns the meta context, which can be adapted to different anomaly types and domains under the few-shot setting.

I. Comparison between MCPM and MAX/MEAN

In order to prove that the STGCN updated video clip features are better than the original video clip features, we remove the STGCN in MCPM and extract the video feature by directly using the maximum/mean of video clip features. As the result shown in Table VIII and Table IX, the performance of MCPM is better than MAX/MEAN in most types of anomaly without DAM and in each anomaly type with DAM, which proves the effectiveness of MCPM.

J. Impact of DAM for different anomaly types

The goal of DAM is to obtain a better video clip encoder ϕ_t^-. As the result shown in Table VIII and Table IX, the performance will be significantly better with DAM regardless of the setting of other modules. This result suggests the robustness of DAM as expected. Considering different sources of normal samples for DAM, we found that the effectiveness of DAM is related to the quality of normal samples. Such quality can be evaluated by the distance to the target domain, the

TABLE IX: 2-way 5-shot result of different anomaly types in UCF-Crime.

Modules	Arson Road Accidents	Burglary Shooting	Explosion Shoplifting	Abuse Robbery	Arrest Stealing	Assault Vandalism	Fighting	All Types
Pretrain	0.80±0.08 0.68±0.09	0.62±0.09 0.60±0.09	0.73±0.09 0.70±0.09	0.75±0.09 0.60±0.10	0.65±0.09 0.71±0.10	0.72±0.09 0.71±0.09	0.72±0.10	0.62±0.10
Pretrain+MAX	0.83±0.08 0.67±0.11	0.64±0.11 0.64±0.11	0.76±0.08 0.73±0.10	0.75±0.10 0.60±0.11	0.64±0.11 0.76±0.10	0.70±0.12 0.69±0.10	0.75±0.10	0.66±0.11
Pretrain+MEAN	0.82±0.08 0.68±0.10	0.64±0.10 0.63±0.10	0.76±0.09 0.72±0.09	0.73±0.09 0.60±0.10	0.66±0.10 0.75±0.08	0.69±0.10 0.68±0.10	0.74±0.08	0.64±0.10
Pretrain+MCPM	0.82±0.09 0.72±0.09	0.66±0.10 0.65±0.10	0.76±0.09 0.73±0.09	0.76±0.09 0.61±0.10	0.65±0.09 0.75±0.09	0.71±0.09 0.69±0.08	0.76±0.09	0.64±0.10
Pretrain+DAM	0.84±0.07 0.68±0.09	0.64±0.09 0.64±0.09	0.76±0.08 0.72±0.08	0.75±0.09 0.64±0.10	0.62±0.10 0.71±0.10	0.70±0.11 0.67±0.10	0.75±0.09	0.64±0.10
Pretrain+DAM+MAX	0.84±0.08 0.70±0.10	0.69±0.11 0.70±0.10	0.75±0.09 0.75±0.09	0.78±0.10 0.68±0.12	0.65±0.11 **0.78±0.09**	0.72±0.11 **0.71±0.12**	0.76±0.10	0.65±0.12
Pretrain+DAM+MEAN	0.86±0.07 0.71±0.10	0.66±0.10 0.70±0.10	0.77±0.08 0.76±0.09	0.78±0.10 0.69±0.10	0.65±0.10 0.77±0.09	0.71±0.09 0.68±0.11	**0.76±0.09**	0.66±0.10
Pretrain+DAM+MCPM	**0.86±0.07** **0.75±0.09**	**0.69±0.09** **0.72±0.09**	**0.78±0.09** **0.77±0.08**	**0.80±0.08** **0.70±0.10**	**0.67±0.10** 0.76±0.08	**0.73±0.09** 0.69±0.10	**0.76±0.09**	**0.67±0.09**

resolution of the video, and the amount of videos. DoTA has no distance to the target domain, and 4136 videos are used for DAM, which leads to the best performance. BDD100K [67] is close to the target domain, and 10,000 high-resolution videos are used for DAM; therefore, the result can be comparable to the best performance. For UCF-Crime, most of the videos are captured by surveillance cameras, whose resolution is not high. Meanwhile, only 950 normal videos are used for DAM; therefore, the effectiveness of DAM is not as significant as one on DoTA, although it has no distance to the target domain.

In this section, we provide t-SNE plots for each anomaly type as well as the "All types" in DoTA and UCF-Crime to prove our modules improve the feature space to be more separable under various types of anomaly. The results are shown from Fig. 6 to Fig. 30. Note that the t-SNE plot might be slightly different for different runs. These plots show the evolution of features from ϕ_s to ϕ_t, which proves the effectiveness and robustness of Anomaly Crossing.

These plots support our analysis of ablation studies. The performance is related to the grouping of the features. For most anomaly types, with the help of DAM, the feature extracted by ϕ_t^- will be more separable compared with those extracted by ϕ_s. For some long-time events (*i.e.*, events in UCF-Crime), the performance of DAM is not as significant as others, leading to less separable features. With the help of learning meta context from MCPM, the video feature extracted by ϕ_t will be more separable, proving that MCPM supplements DAM. As a result, ϕ_s will be adapted to ϕ_t by DAM and MCPM as a better representation of the target domain.
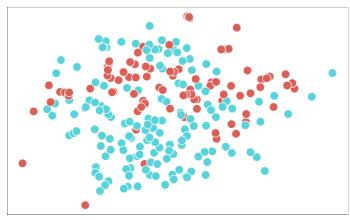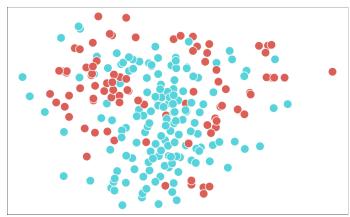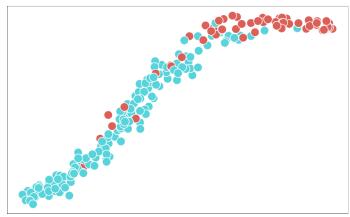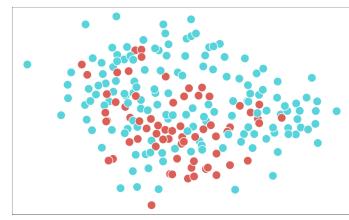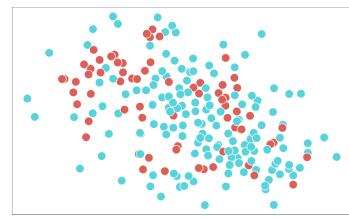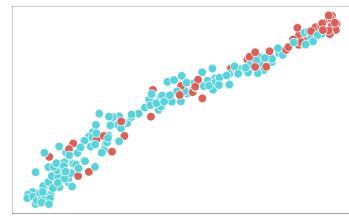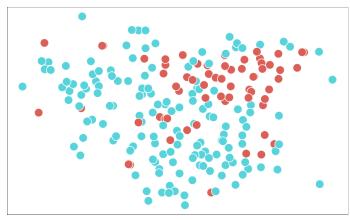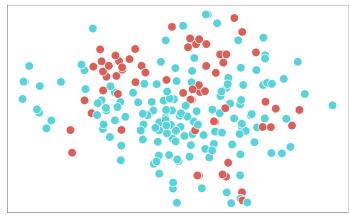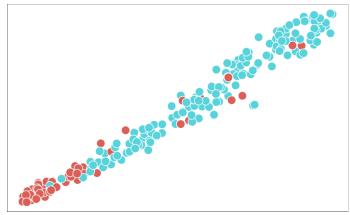

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 6: t-SNE plots for "Lateral" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 7: t-SNE plots for "Leave to left" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 8: t-SNE plots for "Leave to right" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 9: t-SNE plots for "Moving ahead or waiting" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 10: t-SNE plots for "Obstacle" in DoTA
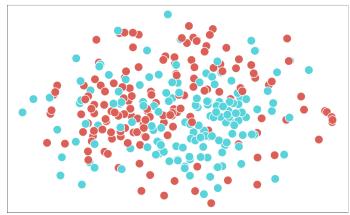
(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 11: t-SNE plots for "Oncoming" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 12: t-SNE plots for "Pedestrian" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 13: t-SNE plots for "Start Stop or Stationary" in DoTA
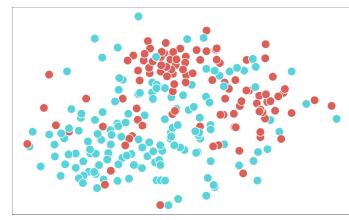
(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 14: t-SNE plots for "Turning" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 15: t-SNE plots for "Unknown" in DoTA

(a) ϕ_s

(b) ϕ_t^-

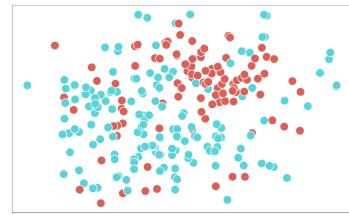(c) ϕ_t

Fig. 16: t-SNE plots for "All Types" in DoTA

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 17: t-SNE plots for "Arson" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 18: t-SNE plots for "Burglary" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 19: t-SNE plots for "Explosion" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

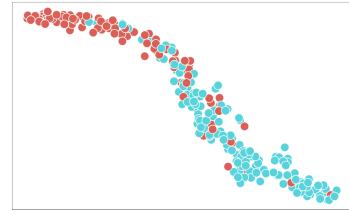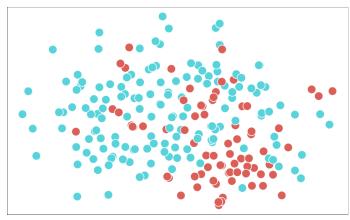Fig. 20: t-SNE plots for "RoadAccidents" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 21: t-SNE plots for "Shooting" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 22: t-SNE plots for "Shoplifting" in UCF Crime

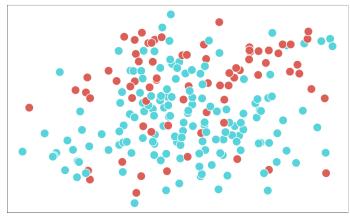(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 23: t-SNE plots for "Abuse" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t
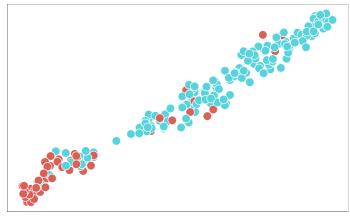
Fig. 24: t-SNE plots for "Arrest" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 25: t-SNE plots for "Assault" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 26: t-SNE plots for "Fighting" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 27: t-SNE plots for "Robbery" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 28: t-SNE plots for "Stealing" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

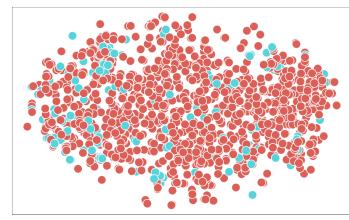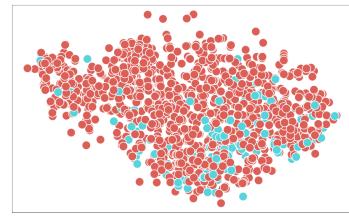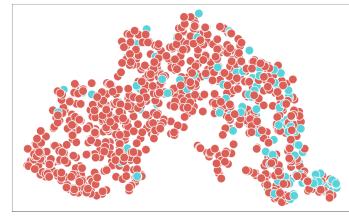Fig. 29: t-SNE plots for "Vandalism" in UCF Crime

(a) ϕ_s

(b) ϕ_t^-

(c) ϕ_t

Fig. 30: t-SNE plots for "All" in UCF Crime