Cross-Domain Video Anomaly Detection without Target Domain Adaptation

Aich, Abhishek; Peng, Kuan-Chuan; Roy-Chowdhury, Amit K.

TR2023-001 January 04, 2023

Abstract

Most cross-domain unsupervised Video Anomaly Detection (VAD) works assume that at least few task-relevant target domain training data are available for adaptation from the source to the target domain. However, this requires laborious model- tuning by the end-user who may prefer to have a system that works "out-of-the-box." To address such practical scenarios, we identify a novel target domain (inference-time) VAD task where no target domain training data are available. To this end, we propose a new 'Zero-shot Cross-domain Video Anomaly Detection (zxVAD)' framework that includes a future-frame prediction generative model setup. Different from prior future- frame prediction models, our model uses a novel Normalcy Classifier module to learn the features of normal event videos by learning how such features are different "relatively" to features in pseudo-abnormal examples. A novel Untrained Convolu- tional Neural Network based Anomaly Synthesis module crafts these pseudo-abnormal examples by adding foreign objects in normal video frames with no extra training cost. With our novel relative normalcy feature learning strategy, zxVAD generalizes and learns to distinguish between normal and abnormal frames in a new target domain without adaptation during inference. Through evaluations on common datasets, we show that zxVAD outperformsthestate-of-the-art(SOTA),regardless of whether task-relevant (i.e., VAD) source training data are avail- able or not. Lastly, zxVAD also beats the SOTA methods in inference-time efficiency metrics including the model size, total parameters, GPU energy consumption, and GMACs.

IEEE Winter Conference on Applications of Computer Vision (WACV) 2023

Cross-Domain Video Anomaly Detection without Target Domain Adaptation

Abhishek Aich[⋆], Kuan-Chuan Peng[†], Amit K. Roy-Chowdhury[⋆]
[⋆]University of California, Riverside, USA, [†]Mitsubishi Electric Research Laboratories, USA
{aaich001@, amitrc@ece.}ucr.edu, kpeng@merl.com

Abstract

Most cross-domain unsupervised Video Anomaly Detection (VAD) works assume that at least few task-relevant target domain training data are available for adaptation from the source to the target domain. However, this requires laborious model-tuning by the end-user who may prefer to have a system that works "out-of-the-box." To address such practical scenarios, we identify a novel target domain (inference-time) VAD task where no target domain training data are available. To this end, we propose a new 'Zero-shot Cross-domain Video Anomaly Detection (zxVAD)' framework that includes a future-frame prediction generative model setup. Different from prior future-frame prediction models, our model uses a novel Normalcy Classifier module to learn the features of normal event videos by learning how such features are different "relatively" to features in pseudo-abnormal examples. A novel Untrained Convolutional Neural Network based Anomaly Synthesis module crafts these pseudo-abnormal examples by adding foreign objects in normal video frames with no extra training cost. With our novel relative normalcy feature learning strategy, zxVAD generalizes and learns to distinguish between normal and abnormal frames in a new target domain without adaptation during inference. Through evaluations on common datasets, we show that zxVAD outperforms the state-of-the-art (SOTA), regardless of whether task-relevant (i.e., VAD) source training data are available or not. Lastly, zxVAD also beats the SOTA methods in inference-time efficiency metrics including the model size, total parameters, GPU energy consumption, and GMACs.

1. Introduction

Unsupervised Video Anomaly Detection (VAD) methods [3–33] have been widely used in security and surveillance applications [34–36] over the supervised or weakly-supervised VAD methods [37–45]. This is mainly because unsupervised VAD methods do not need training videos containing abnormal events which are rare and laborious to annotate [35, 36]. Hence, with only normal events in training videos, the unsupervised VAD methods mark the activities unexplained by the trained model as anomalies during testing. Recently, unsupervised VAD works under cross-domain settings have been introduced [1, 2, 46, 47]. Given the video data containing only normal events from a source domain, the goal is to perform VAD in a different target domain. However, these cross-domain VAD (xVAD) works [1, 2, 46, 47] are methods which need access to either the source and target domain VAD training data [1, 2] or strong supervision from pre-trained object detectors (*e.g.*, YOLOv3 [48] in [47]). Collecting such data in the target domain and adapting or tuning the model may not be feasible by the end-user who may want a system that works "out-of-the-box" [49, 50]. Moreover, granting access to such video data may be time-consuming to third-party corporations due to intellectual property and security concerns [51, 52]. This renders the current xVAD works ineffective as they assume access to at least some target domain training data.

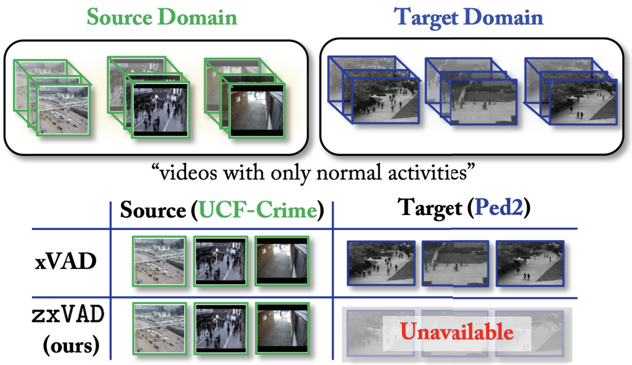


Figure 1: **Problem overview.** Current unsupervised cross-domain VAD works (xVAD) entail adapting to the target domain, assuming access to at least a few training examples [1, 2]. We relax this assumption of having such access to training data from the target domain and tackle a more stringent, yet practical, case using our proposed zero-shot xVAD or zxVAD framework.

Problem Statement. Based on the aforesaid issues, we formally identify the following new unsupervised xVAD problem of detecting anomalies in the target domain with strictly *no access to target domain training data and no prior knowledge of its anomaly types*. More specifically, our goal is to detect anomalies in the *target* domain's testing set, without having any training data on the target side. Fig. 1 contrasts this problem setup with prior xVAD problem definitions.

Proposed framework. We tackle this new problem using a novel xVAD framework, namely 'Zero-shot Cross-domain

Video Anomaly Detection' (zxVAD). The term *zero-shot* implies *no training videos available* from the target domain for adaptation to perform anomaly detection. zxVAD has a generator [53] in a future-frame prediction setup [3] similar to xVAD approaches [1, 2]. However different from these methods, zxVAD's generator training is assisted by a novel *Normalcy Classifier* (NC) module and an *Untrained Convolutional Neural Network (CNN) Anomaly Synthesis* ($\mathcal{O}$) module. Prior unsupervised xVAD works learn features from *only* videos with normal events. This leads to overfitting to the source domain distribution and the poor generalizing ability for target domain VAD [2]. In contrast, zxVAD's generator uses NC and $\mathcal{O}$ modules to learn features of normal activities in input videos, by focusing on how such features are *relatively* different from features of abnormal frames. This "relative" learning strategy enhances the generator's ability in identifying anomalies in target domain without any adaptation at test time.

Normalcy, by definition, is always contextually dependent [42, 54] (*e.g*., running in playgrounds *vs* highways). Hence, to generalize across new target domains (without its training data) where we have no prior knowledge about anomaly types, we propose to learn normal event features that consider the contextual or relative difference between "normal" and "abnormal" patterns. More concretely, rather than learning only the normalcy features (*i.e*. features of normal video frames), our model learns the *relative* normalcy features (*i.e*. difference between features of normal *and* abnormal video frames) using our proposed NC module. These pseudo-abnormal frames are created through our proposed $\mathcal{O}$ module which is capable of localizing objects from both Task-Relevant or VAD data and Task-Irrelevant (TI) or non-VAD data (*i.e*., data irrelevant to the VAD task). The $\mathcal{O}$ module crafts pseudo-abnormal frames by localizing objects in input TI or VAD video frames and pasting them (with random location and size) on normal VAD video frames. Furthermore, a major advantage of introducing TI data to our problem setup is that they can be treated as *video distributions for learning patterns of normal activities* and also assist in creating diverse anomalies. Hence, along with the strategy of learning the relative normalcy difference, zxVAD aims to mitigate the generalizing issue *via* learning this relative normalcy with respect to abnormal frames having different kinds of foreign objects (either from VAD or TI frames). This allows zxVAD to avoid being limited to specific anomaly types in the source domain, making it fundamentally different from supervised specific anomaly learning.

Our NC module is designed to distinguish between a pseudo-abnormal and the predicted normal future-frame through novel loss functions. The highlighting attribute of these functions is to consider different properties of normal and abnormal frames through our NC's logit predictions and derived attention maps. Our $\mathcal{O}$ module is uniquely capable of using VAD or TI data with an untrained randomly initialized CNN to create anomalies at no extra training cost. To sum up, we make the following key **contributions**:

Table 1: **Characteristic comparison.** Better than prior unsupervised VAD works (*e.g*., $\mathcal{C}_0$: [54–60], $\mathcal{C}_1$: [3–25], $\mathcal{C}_2$: [46, 47, 61], $\mathcal{C}_3$ (our baselines): [1, 2]), zxVAD needs no prior knowledge (*e.g*. object extraction from VAD videos), can perform cross-domain VAD with no VAD training data, and uses an untrained CNN to create anomalies.

Unsupervised VAD Method Conditions	**Unsupervised VAD Categories**				
	$\mathcal{C}_0$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	zxVAD **(ours)**
no prior **knowledge** required?	✗	✓	✗	✓	✓
show efficacy in **cross-domain** VAD?	✗	✗	✓	✓	✓
works with **no source domain VAD training data**?	✗	✗	✗	✗	✓
create pseudo-anomalies with an **untrained** network?	✗	✗	✗	✗	✓

1. We formally introduce a **new problem setup** in xVAD where the model is trained on the source domain to detect anomalies (at test time) in *a different target domain without any adaptation via target-domain training data* or using any supervision from pre-trained models (*e.g*. YOLOv3).
2. A **novel** xVAD **method** namely zxVAD is proposed, where the model learns the *relative difference between normal and abnormal frames* in source domain and generalize VAD to target without needing target domain training data or any external support from pre-trained models.
3. This "*relative*" difference learning is achieved via **a novel Normalcy Classifier** that uses a **new pseudo-anomaly synthesis module based on an untrained CNN** where anomalies are created with no extra training cost.
4. Notably, for the first time in VAD literature (to the best of our knowledge), we also show that zxVAD **outperforms the SOTA xVAD works** in the proposed problem setup **when trained only with TI data**, in four common benchmarks.
5. zxVAD **beats the SOTA xVAD works** in the proposed problem setup both in AUC on most benchmarks, and in inference-time efficiency metrics (*e.g*., model size, model parameters, GPU energy consumption, and GMACs).

2. Related Works

Unsupervised VAD works. Early unsupervised VAD works formulated the anomaly detection using handcrafted features to characterize the normal event or regular pattern distribution [4–12]. However, these methods were outperformed by the CNN approaches [3, 13–25] (both categorized as $\mathcal{C}_1$ in Tab. 1). Some of these CNN based unsupervised VAD works use generators [53] to model the normal frame distributions [3, 18, 20–22, 58], and further introduce memory modeling networks to record various normal event patterns in videos [1, 2, 18, 21, 58]. Another category of works ($\mathcal{C}_0$ in Tab. 1) [54–60] proposed computationally heavy approaches that used strong priors like object extraction (using pre-trained object-detectors [54, 57]) for VAD, in order to focus only on specific objects to detect anomalies. Compared to aforesaid VAD works in $\mathcal{C}_0$ and $\mathcal{C}_1$, zxVAD *(a)* is designed to tackle unsupervised *cross-domain* VAD problem, *(b)* is a future-frame prediction method with a memory module, and *(c)* needs no strong prior knowledge from object extraction. Finally, few works [47, 54, 60, 62–64] have shown different VAD strategies where pseudo-anomalies are used. For example, [54, 64] uses a generator to create fake anomaly data. [62, 63] propose two different temporal pseudo anomaly synthesizers to craft anomalies from normal videos. Needing no aforesaid extra training efforts, `zxVAD` uses a novel strategy to create anomalies using an *untrained randomly initialized* CNN (details in Sec. 3) at no extra training cost.

Cross-domain setting. The cross-domain scenario in unsupervised VAD has been introduced in [1, 2, 46, 61, 65]. These works operate under the regime of few-shot target domain scene adaptation. For example, [1, 2] ($\mathcal{C}_3$ in Tab. 1) use meta-learning approaches [66] and adapt to the target domain with few scenes for anomaly detection. In contrast, `zxVAD` is specifically designed for cross-domain VAD *without any target domain adaptation*. [46, 47, 61] ($\mathcal{C}_2$ in Tab. 1) provide prior knowledge based methods where videos are subject to object-extraction using pre-trained object detectors [48, 67]. However, `zxVAD` needs no strong priors like object extraction using pre-trained detectors. `zxVAD` is also capable of solely using TI data and outperforming the SOTA in proposed cross-domain VAD setup. Finally, `zxVAD` uses a simple training strategy (details in Sec. 3) rather than using a meta-learning approach to avoid non-trivial computational and memory burdens, as well as vanishing gradients issues [46, 68].

3. Proposed `zxVAD` Framework

Method Overview. We are provided with source domain VAD normal videos to learn features that should ideally transfer across different target domains without needing target domain adaptation. To achieve this *no* adaptation-based cross-domain VAD property, we introduce a novel `zxVAD` framework (illustrated in Fig. 2) based on a future-frame prediction setup that can be trained end-to-end. It consists of an untrained CNN based pseudo-anomaly Synthesis module (Sec. 3.1) where an untrained randomly initialized CNN helps in creating pseudo-anomalies without any extra training burden. These pseudo-abnormal frames along with predicted future-frame are utilized in our novel Normalcy Classifier module (Sec. 3.2) to regularize the backbone generator to learn relative normalcy features. This learning strategy makes `zxVAD` capable of more generalizable VAD performance across different target domains than existing xVAD methods.

Notations. We denote a sample video from VAD datasets as $[\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{L_v}] \in \mathbb{R}^{L_v \times C \times H \times W}$, and TI datasets as $[\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{L_u}] \in \mathbb{R}^{L_u \times C \times H \times W}$, where each video contains L_v and L_u number of frames, and each frame is of height H, width W, and C channels. Our future-frame prediction framework `zxVAD` contains a memory-augmented generator [18] $\mathcal{G}(\cdot)$ with weights $\boldsymbol{\theta}_{\mathcal{G}}$ and memory module $\mathcal{M}$, and a discriminator $\mathcal{D}(\cdot)$ with weights $\boldsymbol{\theta}_{\mathcal{D}}$. As shown in [18, 69, 70], the memory module $\mathcal{M} \in \mathbb{R}^{K \times Q}$ is a matrix with $\boldsymbol{m}_i \in \mathbb{R}^Q, \forall i \in [K]$ vectors (or memory items) that learns to register the prototypical normal features during training. $\mathcal{M}$ takes the output vector $\boldsymbol{z} \in \mathbb{R}^Q$ from $\mathcal{G}(\cdot)$'s encoder and outputs $\widehat{\boldsymbol{z}} = \boldsymbol{w}\mathcal{M} \in \mathbb{R}^Q$ that is forwarded to $\mathcal{G}(\cdot)$'s decoder. Here, $\boldsymbol{w} \in \mathbb{R}^{1 \times K}$ is termed as a soft addressing vector [18]. Each element w_i of $\boldsymbol{w}$ is computed using softmax operation on the cosine similarity between $\boldsymbol{z}$ and $\boldsymbol{m}_i$ [18, 71]. Our proposed anomaly synthesis module is denoted as $\boldsymbol{\mathcal{O}}$ and contains a CNN denoted as $\mathcal{R}(\cdot)$ with weights $\boldsymbol{\theta}_{\mathcal{R}}$. Further, our proposed normalcy classifier module contains a CNN classifier denoted as $\mathcal{N}(\cdot)$ with weights $\boldsymbol{\theta}_{\mathcal{N}}$. We denote the expectation operator, l_p-norm operator, and element-wise multiplication by $\mathbb{E}[\cdot]$, $\|\cdot\|_p$, and $\odot$, respectively.

Backbone description. Given N source domain training videos (with only normal events), we aim to learn a future-frame prediction generator that takes in T input frames and predicts a future frame $\widehat{\boldsymbol{v}}_{T+1}$, *i.e.*, $\mathcal{G}\big([\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_T]\big) = \widehat{\boldsymbol{v}}_{T+1}$. $\mathcal{G}(\cdot)$ is adversarially trained against $\mathcal{D}(\cdot)$ in the Least-Square GAN [53, 72] setup where $\mathcal{D}(\cdot)$ aims to distinguish between $\widehat{\boldsymbol{v}}_{T+1}$ and the ground truth frame $\boldsymbol{v}_{T+1}$. Similar to [2, 18], we introduce a memory module $\mathcal{M}$. In `zxVAD`, $\mathcal{G}(\cdot)$ is further regularized using $\mathcal{N}(\cdot)$ with our proposed four novel objectives (explained in Sec. 3.2) which uses pseudo-anomaly examples generated using an untrained CNN based strategy. Following prior works [1, 73], we optimize $\mathcal{G}(\cdot)$ with the mean square error loss $\mathcal{L}_{\text{MSE}} = \|\widehat{\boldsymbol{v}}_{T+1} - \boldsymbol{v}_{T+1}\|_2^2$, structure similarity loss $\mathcal{L}_{\text{SSM}} = 1 - \text{SSIM}\big(\widehat{\boldsymbol{v}}_{T+1}, \boldsymbol{v}_{T+1}\big)$, where SSIM represents the structural similarity index measure [74] between $\widehat{\boldsymbol{v}}_{T+1}$ and $\boldsymbol{v}_{T+1}$, and Gradient loss $\mathcal{L}_{\text{GD}}$ [3, 73]. To optimize $\mathcal{M}$ and encourage modeling normal videos using sparse but most relevant memory slots, we follow [18] and apply a hard-shrinkage on $\mathcal{M}$'s memory addressing vectors w_i using continuous ReLU activation function with a shrinkage factor λ set as 0.0005. Next, we normalize each element $\widehat{w}_i \leftarrow {}^{\widehat{w}_i}/_{\|\widehat{\boldsymbol{w}}\|_1} \forall i$ and get $\widehat{\boldsymbol{z}} = \widehat{\boldsymbol{w}}\mathcal{M}$. We also apply a sparsity regularizer on $\widehat{\boldsymbol{w}}$ by minimizing its entropy as $\mathcal{L}_{\text{MEM}} = \sum_{i=1}^{N} -\widehat{w}_i \log\big(\widehat{w}_i\big)$ [18]. We combine these losses as

$$\mathcal{L}_{\text{BB}} = \mathcal{L}_{\text{REC}} + \alpha_{\text{MEM}} \mathcal{L}_{\text{MEM}}, \tag{1}$$

where the reconstruction loss is $\mathcal{L}_{\text{REC}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SSM}} + \mathcal{L}_{\text{GD}}$. We set the loss weight $\alpha_{\text{MEM}} = 0.0025$ following [18]. Totally, the weights $\boldsymbol{\theta}_{\mathcal{G}}$, $\boldsymbol{\theta}_{\mathcal{D}}$ and $\boldsymbol{\theta}_{\mathcal{N}}$ are updated during training, while $\boldsymbol{\theta}_{\mathcal{R}}$ is randomly initialized before training and remains fixed. Better than prior works which do not consider the relative difference between normal and abnormal events, `zxVAD` introduces a novel strategy to regularize this backbone generator by learning normal features with respect to pseudo-abnormal features. As our normalcy classifier module utilizes pseudo-anomalies to learn the *relative* normalcy features, we first present our pseudo-anomaly creation strategy.

3.1. Pseudo-Anomaly Synthesis via Untrained CNN

Prior works [54] have focused on creating anomalies using pre-trained object detectors (*i.e.*, YOLOv3 [48] in [54]) that result in issues like additional training overheads. Different from such methods, we present a training-free strategy to extract objects from video frames. These objects can be obtained

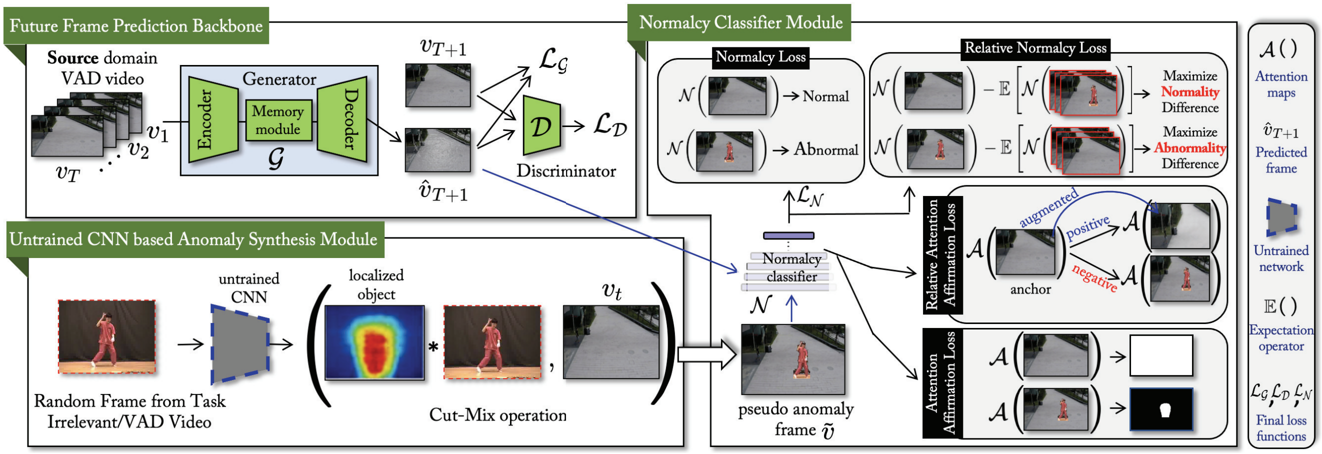

Future Frame Prediction Backbone
Source domain VAD video
v_1
v_2
v_T
Generator
Encoder
Memory module
Decoder
$\mathcal{G}$
v_{T+1}
$\hat{v}_{T+1}$
$\mathcal{L}_\mathcal{G}$
$\mathcal{D}$
$\mathcal{L}_\mathcal{D}$
Discriminator
Normalcy Classifier Module
Normalcy Loss
Normal
Abnormal
Relative Normalcy Loss
Maximize Normality Difference
Maximize Abnormality Difference
$\mathcal{L}_\mathcal{N}$
Normalcy classifier
$\mathcal{N}$
Relative Attention Affirmation Loss
augmented
positive
negative
anchor
Attention Affirmation Loss
pseudo anomaly frame $\tilde{v}$
Untrained CNN based Anomaly Synthesis Module
untrained CNN
localized object
Random Frame from Task Irrelevant/VAD Video
Cut-Mix operation
v_t
$\mathcal{A}()$ Attention maps
$\hat{v}_{T+1}$ Predicted frame
Untrained network
$\mathbb{E}()$ Expectation operator
$\mathcal{L}_\mathcal{G}, \mathcal{L}_\mathcal{D}, \mathcal{L}_\mathcal{N}$ Final loss functions

Figure 2: **Framework overview.** Our `zxVAD` framework contains a Future-Frame Prediction backbone (*top-left*) guided by our Normalcy Classifier module (*right*). To enforce the prediction backbone to learn generalizable features from source domain normal videos and avoid overfitting, we encourage the generative model to learn normalcy features relative to pseudo-abnormal frames using four novel loss functions. These abnormal frames are created using an untrained randomly-initialized CNN through our novel anomaly synthesis module $\mathcal{O}$ (*bottom-left*).

on both VAD and TI video frames (*i.e.*, $\boldsymbol{v}_t$ and $\boldsymbol{u}_t$). For brevity, we refer to the input frame as $\boldsymbol{x}$. Given an input frame $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$, we denote the output of a CNN $\mathcal{R}(\cdot)$ (before the classification layer) as tensor $\boldsymbol{G} \in \mathbb{R}^{d\times h\times w}$. For example, if $\mathcal{R}(\cdot)$ is ResNet152 [75], $\boldsymbol{G}$ is the output of '*conv*5_*x*' with size $2048\times8\times8$ if input size is $3\times256\times256$. We employ SCDA [76] to perform channel-wise summation on $\boldsymbol{G}$ to obtain an attention map $\boldsymbol{A} \in \mathbb{R}^{h\times w}$. We then obtain a binary mask $\boldsymbol{M}$ from $\boldsymbol{A}$ as follows. We set $\boldsymbol{M}_{(i,j)} = 1$ if $\boldsymbol{A}_{(i,j)} > \varsigma$, or 0 otherwise. Here, (i,j) represents position in $h\times w$ locations. We empirically set $\varsigma = 0.1$. $\boldsymbol{M}_{(i,j)} = 1$ indicates the foreground objects. Finally, $\boldsymbol{M}$ is resized from $h\times w$ to $H\times W$. As noted in [77], the idea behind this surprising property that randomly initialized CNN can localize objects is: because the background in the input frame $\boldsymbol{x}$ is relatively texture-less in comparison to the foreground objects in the scene, these background regions have higher chances to be deactivated by nonlinear activation functions like ReLU [78]. The object is finally localized as $\boldsymbol{M}_{\boldsymbol{x}} = \boldsymbol{M} \odot \boldsymbol{x}$. To create pseudo-abnormal frame $\tilde{\boldsymbol{v}}$, we combine $\boldsymbol{M}_{\boldsymbol{x}}$ and one of the input frames to $\mathcal{G}(\cdot)$, *i.e.*, $\boldsymbol{v}_t \in \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_T\}$ by pasting $\boldsymbol{M}_{\boldsymbol{x}}$ on $\boldsymbol{v}_t$ at random location r_z with random size $r_x \times r_y$. We discuss the method to choose the location r_z and size $r_x \times r_y$ in Supplementary Material. Note that most of the video frames used for creating pseudo-anomalies happen to contain at least one foreground object for the untrained CNN to extract. Even if there are no such objects, our untrained CNN will still focus on some patches (on the input frame) and treat them as anomaly on normal event VAD frame.

3.2. Learning Normality w.r.t. Abnormality

Our Normalcy Classifier Module is a classifier $\mathcal{N}(\cdot)$ that is optimized by the following four loss functions. These loss functions are complementary to each other as follows: *normalcy* loss and *attention affirmation* loss focus on the difference between normal and abnormal frames, whereas *relative normalcy* loss and *relative attention affirmation* loss focus on how relatively different are normal frames from abnormal frames (and vice-versa). For clarity, we drop the subscript of the predicted frame $\hat{\boldsymbol{v}}_{T+1}$ and mark it as $\hat{\boldsymbol{v}}$. The data distribution of normal and pseudo-abnormal frames are denoted as ρ and κ, respectively.

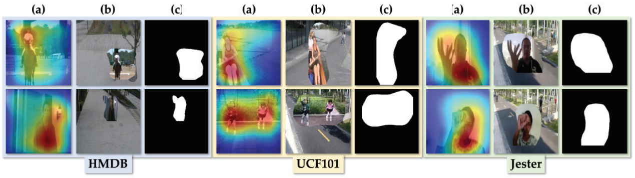


Figure 3: **Pseudo-abnormal examples.** We show pseudo-abnormal frames (marked as **(b)**) created using our pseudo-anomaly synthesis strategy. The untrained randomly initialized CNN is ResNet50 [75] which localizes objects in the TI frames (marked as **(a)**). We also show examples of ground-truth masks $\tilde{\boldsymbol{M}}$ used in loss $\mathcal{L}_{\text{RAA}}$ (marked as **(c)**). See more examples in Supplementary Material.

Normalcy loss $\mathcal{L}_{\text{N}}$. Given the predicted future-frame $\hat{\boldsymbol{v}}$ and pseudo-abnormal frame $\tilde{\boldsymbol{v}}$, $\mathcal{L}_{\text{N}}$ optimizes $\mathcal{N}(\cdot)$ to increase the probability that $\hat{\boldsymbol{v}}$ is 'normal' (label set as 1) and $\tilde{\boldsymbol{v}}$ is 'abnormal' (label set as 0), using following loss function.

$$\mathcal{L}_{\text{N}} = {}^1\!/_2\mathbb{E}_{\hat{\boldsymbol{v}}\sim\rho}\Big[\big(\mathcal{N}(\hat{\boldsymbol{v}})-1\big)^2\Big] + {}^1\!/_2\mathbb{E}_{\tilde{\boldsymbol{v}}\sim\kappa}\Big[\big(\mathcal{N}(\tilde{\boldsymbol{v}})\big)^2\Big] \tag{2}$$

Relative normalcy loss $\mathcal{L}_{\text{RN}}$. Abnormal events can be viewed as deviation with respect to normal events. We argue that the key missing attribute of (2) is that the probability of normal data being *normal* ($\mathcal{N}(\hat{\boldsymbol{v}})$) should increase as the probability of abnormal data being *normal* ($\mathcal{N}(\tilde{\boldsymbol{v}})$) decreases and vice-versa. Rather than just maximizing $\mathbb{P}[\hat{\boldsymbol{v}}$ is normal$]$, we also ask $\mathcal{N}(\cdot)$ to maximize $\mathbb{P}[\hat{\boldsymbol{v}}$ is more normal than $\tilde{\boldsymbol{v}}]$ ($\mathbb{P}[\cdot]$ denotes probability operator). We define this novel relative normalcy loss below:

$$\begin{aligned}\mathcal{L}_{\text{RN}} = {}^1\!/_2\mathbb{E}_{\hat{\boldsymbol{v}}\sim\rho}\Big[\big(\mathcal{N}(\hat{\boldsymbol{v}}) - \mathbb{E}_{\tilde{\boldsymbol{v}}\sim\kappa}[\mathcal{N}(\tilde{\boldsymbol{v}})] - 1\big)^2\Big] +& \\ {}^1\!/_2\mathbb{E}_{\tilde{\boldsymbol{v}}\sim\kappa}\Big[\big(\mathcal{N}(\tilde{\boldsymbol{v}}) - \mathbb{E}_{\hat{\boldsymbol{v}}\sim\rho}[\mathcal{N}(\hat{\boldsymbol{v}})] + 1\big)^2\Big]&\end{aligned} \tag{3}$$

Attention affirmation loss $\mathcal{L}_{\text{AA}}$. The decision of $\mathcal{N}(\cdot)$ on the normal frame $\hat{\boldsymbol{v}}$ and abnormal frame $\tilde{\boldsymbol{v}}$ should be based on the following information: *(1)* $\mathcal{N}(\cdot)$ should consider the whole scene in $\widehat{\boldsymbol{v}}$ to classify it as '*normal*,' and *(2)* $\mathcal{N}(\cdot)$ should consider the foreign object (introduced by our module $\boldsymbol{\mathcal{O}}$ in $\tilde{\boldsymbol{v}}$) to classify it as '*abnormal*.' Our strategy in Sec. 3.1 allows us to obtain the exact location of foreign objects in $\tilde{\boldsymbol{v}}$. Hence, we leverage this knowledge and create ground-truth masks of $\tilde{\boldsymbol{v}}$. We first initialize a tensor $\tilde{\boldsymbol{M}}$ with zeroes. Next, we update this tensor by pasting $\boldsymbol{M}$ after resizing to $r_x \times r_y$ at location r_z (obtained from $\boldsymbol{\mathcal{O}}$ in Sec. 3.1). We show examples of $\tilde{\boldsymbol{M}}$ in Fig. 3. We extract feature maps from the last convolutional layer of $\mathcal{N}(\cdot)$ and apply SCDA [76] to obtain attention maps $\mathcal{A}(\widehat{\boldsymbol{v}})$ and $\mathcal{A}(\tilde{\boldsymbol{v}})$ for normal and abnormal frames, respectively. $\mathcal{A}(\cdot)$ denotes the operation to extract attention maps from $\mathcal{N}(\cdot)$. We enforce this constraint via the attention affirmation loss $\mathcal{L}_{\text{AA}}$ as ($\mathbb{1}$ is a tensor of the same size as $\mathcal{A}(\widehat{\boldsymbol{v}})$ filled with ones):

$$\mathcal{L}_{\text{AA}} = {}^1\!/_2\big(\mathbb{1} - \mathcal{A}(\widehat{\boldsymbol{v}})\big)^2 + {}^1\!/_2\big(\tilde{\boldsymbol{M}} - \mathcal{A}(\tilde{\boldsymbol{v}})\big)^2, \tag{4}$$

Relative attention affirmation loss $\mathcal{L}_{\text{RAA}}$. Similar to the concept of $\mathcal{L}_{\text{RN}}$, we argue that $\mathcal{L}_{\text{AA}}$ does not consider the relative difference of attention maps from normal frames with respect to attention maps from abnormal frames. Hence, we propose a relative attention affirmation loss $\mathcal{L}_{\text{RAA}}$ that aims to learn this difference. We create two attention map pairs: *(Pair-1)* $\mathcal{A}(\widehat{\boldsymbol{v}})$ and $\mathcal{A}(g(\widehat{\boldsymbol{v}}))$, and *(Pair-2)* $\mathcal{A}(\widehat{\boldsymbol{v}})$ and $\mathcal{A}(\tilde{\boldsymbol{v}})$. The function $g(\cdot)$ denotes a series of transformations (*Color Jitter*, *Random Affine*, and *Random Perspective*) applied to $\widehat{\boldsymbol{v}}$ using the package Kornia [79] (related parameters are provided in Supplementary Material). The relative difference between the attention on 'augmented normal' frame should be smaller than that of the 'pseudo-abnormal' frame with respect to the 'normal' frame. We enforce this difference with a margin m that simultaneously enhances the intra-class compactness between normal and augmented-normal frames and inter-class discrepancy between normal and pseudo-abnormal frames. We design $\mathcal{L}_{\text{RAA}}$ using the ArcFace loss [80] enforcing this margin as follows.

$$\mathcal{L}_{\text{RAA}} = \frac{-1}{N}\sum_{i=0}^{N-1}\log\left(\frac{e^{s(\cos(\omega_{y_i}+m))}}{e^{s(\cos(\omega_{y_i}+m))}+\sum_{j=0,j\neq y_i}^{1} e^{s\cos(\omega_j)}}\right), \tag{5}$$

where label y_i is set as 1 for normal frame $\widehat{\boldsymbol{v}}$ and augmented frame $g(\widehat{\boldsymbol{v}})$, and 0 for pseudo-abnormal frame $\tilde{\boldsymbol{v}}$. We transform $\mathcal{A}(\boldsymbol{x})$ with $\psi_{y_i} = \|\boldsymbol{W}_{y_i}\|\|\texttt{vec}(\mathcal{A}(\boldsymbol{x}))\|\cos(\omega_{y_i})$ (with $\omega_{y_i} \in [0,\pi]$ as the angle between $\boldsymbol{W}_{y_i}$ and $\texttt{vec}(\mathcal{A}(\boldsymbol{x}))$). Here, $\texttt{vec}(\cdot)$ is a vectorizing operation. $\|\boldsymbol{W}_{y_i}\|$ and $\|\texttt{vec}(\mathcal{A}(\boldsymbol{x}))\|$ are normalized to 1 which leads to $\psi_{y_i} = \cos(\omega_{y_i})$. With ArcFace loss, $\boldsymbol{W}_{y_i}$ behaves as a centre for each class (*i.e.* normal and abnormal) [80] which creates a distance margin penalty of m. We set scaling factor s = 64 and margin m = 28.6 degrees following [81]. $\mathcal{L}_{\text{RAA}}$ can be implemented as any triplet metric learning loss [81]. However, we choose the ArcFace loss as it has been shown to perform well in recent non-VAD works [82–84].

Final learning objectives. To summarize, `zxVAD` is trained end-to-end with $\mathcal{G}(\cdot)$ learning loss $\mathcal{L}_{\mathcal{G}}$, $\mathcal{D}(\cdot)$ learning loss $\mathcal{L}_{\mathcal{D}}$, and $\mathcal{N}(\cdot)$ learning loss $\mathcal{L}_{\mathcal{N}}$ as follows:

$$\begin{gathered}
\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\text{BB}} + \alpha_{\mathcal{D}}\mathbb{E}_{\widehat{\boldsymbol{v}}\sim\rho}\Big[{}^1\!/_2\big(\mathcal{D}(\widehat{\boldsymbol{v}})-1\big)^2\Big] + \\
\alpha_{\mathcal{N}}\mathbb{E}_{\widehat{\boldsymbol{v}}\sim\rho}\Big[{}^1\!/_2\big(\mathcal{N}(\widehat{\boldsymbol{v}})-1\big)^2\Big], \\
\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{\widehat{\boldsymbol{v}}\sim\rho}\Big[{}^1\!/_2\big(\mathcal{D}(\widehat{\boldsymbol{v}})\big)^2\Big] + \mathbb{E}_{\widehat{\boldsymbol{v}}\sim\rho}\Big[{}^1\!/_2\big(\mathcal{D}(\boldsymbol{v})-1\big)^2\Big], \\
\mathcal{L}_{\mathcal{N}} = \alpha_{\text{n}}\mathcal{L}_{\text{N}} + \alpha_{\text{rn}}\mathcal{L}_{\text{RN}} + \alpha_{\text{aa}}\mathcal{L}_{\text{AA}} + \alpha_{\text{raa}}\mathcal{L}_{\text{RAA}}
\end{gathered} \tag{6}$$

We set $\alpha_{\mathcal{D}} = 0.05$ following [3]. The rest of loss weights $\alpha_{\mathcal{N}} = 0.5, \alpha_{\text{n}} = 1, \alpha_{\text{rn}} = 0.01, \alpha_{\text{aa}} = 1$, and $\alpha_{\text{raa}} = 1$ are set empirically.

Discussion on why `zxVAD` works. `zxVAD` trains $\mathcal{G}(\cdot)$ in predicting the future-frame $\widehat{\boldsymbol{v}}$ of input normal video by considering the difference with respect to pseudo-abnormal frames (*via* our normalcy classifier module). Not considered in prior works, this strategy specifically helps $\mathcal{G}(\cdot)$ to learn contextual relative difference between normal and abnormal frames to alleviate overfitting to source domain normal video features. The overfitting issue is further mitigated when the abnormal examples created from our pseudo-anomaly module contains various kinds of objects as "foreign entities" in VAD normal frames. This allows $\mathcal{G}(\cdot)$ to learn the relative normalcy difference from extremely diverse kinds of pseudo-anomalies, making it capable of detecting different anomaly types (in inference-time) in multiple target domains without any prior knowledge.

To make the further discussion concise, we show the statistics and acronyms of the VAD and TI datasets in Tab. 2.

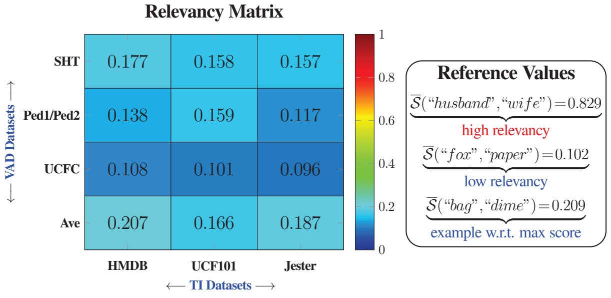
3.3. Introduction to Task-Irrelevant (TI) Datasets

In this section, we discuss the utilities of task-irrelevant or non-VAD videos for unsupervised VAD. Task-relevant or VAD datasets provided by VAD research community are known to be limited in scale as shown in Tab. 2 and [1, 42, 88]. (*e.g.* Ave [9], Ped1, Ped2 [10] datasets have <100 training videos). Further, it is difficult to collect different kinds of scenarios of normal activities with such limited scale. Hence, we propose to introduce the utility of Task-Irrelevant (TI) datasets to the task of VAD.

We define a dataset as '*Task-Irrelevant*' which is freely available from different other video downstream or non-VAD tasks (*e.g.*, video classification, action recognition, *etc.*). Examples of such datasets are UCF101 [86] and HMDB [85] (see Tab. 2). Such datasets were originally introduced for non-VAD works, specifically curated for large-scale deep learning-based tasks. For example, Jester was originally introduced for video classification of 25 hand gesture classes [87]. To show the performance using diverse types of datasets in our `zxVAD` task, we choose Jester, UCF101, and HMDB to be our TI datasets. Please see Supplementary Material for dataset examples. Next, we discuss how the task-relevancy of these datasets is measured with respect to the VAD task, followed by two simple strategies to use these datasets in the proposed problem scenario. Note that `zxVAD` needs nothing from the TI-VAD relevancy measure to operate. The purpose is to only validate TI data's irrelevancy to the VAD task.

Table 2: **Dataset statistics.** We highlight the difference in amount of training data between VAD and TI datasets. *: the train/test disjoint camera (dc) split is provided by [1]. As stated in [1], UCFC dataset does not contain ground truth frame-level labels and hence is not considered for evaluation.

	Task-Relevant/VAD Datasets						**Task-Irrelevant/non-VAD Datasets**		
Property \ Dataset	Shanghai-Tech [16]	Shanghai-Tech* [1]	UCF-Crime [37]	Ped1 [10]	Ped2 [10]	CUHK-Avenue [9]	HMDB51 [85]	UCF101 [86]	20BN-JESTER [87]
Acronym	SHT	SHT_{dc}	UCFC	Ped1	Ped2	Ave	HMDB	UCF101	Jester
# of training / testing videos	330 / 107	147 / 33	950 / -	34 / 36	16 / 12	16 / 21	6,766 / -	13,320 / -	50,420 / -
# of abnormal instances	47	33	–	40	12	21	N/A	N/A	N/A



(0 : ‘*lower*’, 1 : ‘*higher*’ semantic similarity)

Figure 4: **Relevancy measure between VAD and TI labels.** Using the relevancy score matrix between the TI and VAD labels, we find that TI datasets have low semantic similarity with the VAD datasets. The maximum score occurs between HMDB and Ave.

Measuring relevancy of non-VAD datasets. Following [89, 90], we use word2vec [91] (pre-trained on Google News dataset [91]) to measure the task-relevancy of our introduced TI datasets: Jester, UCF101, and HMDB. We first compute an embedding vector of the input labels (in case the label contains more than one word, we average the embedding). Next, we compute the mean absolute cosine similarity $\overline{\mathcal{S}} \in [0,1]$ of the embedding for all possible pairs of labels between the TI datasets and abnormal classes of VAD datasets. This is denoted as $\overline{\mathcal{S}} = {}^1\!/_{\ell_P \ell_Q} \sum_{p=1}^{\ell_P} \sum_{q=1}^{\ell_Q} \left|\text{cos-sim}\left(\boldsymbol{\pi}_p, \boldsymbol{\pi}_q\right)\right|$, where ℓ_P and ℓ_Q are the total number of labels in the TI and VAD dataset, $\boldsymbol{\pi}_p$ and $\boldsymbol{\pi}_q$ are the word2vec representation of the p-th and q-th label of the TI and VAD dataset, respectively. cos-sim$(\cdot)$ denotes the cosine similarity operation on input vectors. A value of $\overline{\mathcal{S}}$ closer to 0 indicates a higher degree of irrelevancy (or lower degree of relevance). In Fig. 4, we show the mean cosine similarity $\overline{\mathcal{S}}$ for all the TI (*i.e.*, HMDB, UCF101, Jester) and VAD (*i.e.*, SHT, Ped1/Ped2, UCFC, Ave) datasets used in this paper. Following reference values: $\overline{\mathcal{S}}(\text{“}object\text{”}, \text{“}scene\text{”}) = 0.829$, $\overline{\mathcal{S}}(\text{“}bag\text{”}, \text{“}dime\text{”}) = 0.209$, and $\overline{\mathcal{S}}(\text{“}fox\text{”}, \text{“}paper\text{”}) = 0.109$, we find that the maximum semantic similarity $\overline{\mathcal{S}} = 0.207$ occurs between Ave and HMDB indicating that all the TI datasets are quite irrelevant to the task of VAD problem.

Methods to use TI datasets. We provide two methods to use TI datasets. *Firstly*, unsupervised VAD methods learn features from normal events during training. These events are particularly marked by continuous activities without any sudden disruption from alien objects. Such kinds of videos are readily available in other video downstream tasks like action recognition, where a sample video only contains frames from a continuous activity. In cases where there is no VAD training data available in the source domain (worst-case scenario), we show later in Sec. 4 that training `zxVAD` solely with TI datasets reaches the SOTA results across 3 different target domain datasets. We hypothesize that TI datasets represent the recording of normal activities as in VAD training data with normal videos. Hence, learning from such TI data helps in modeling features similar to normal videos. *Secondly*, we recommend using TI frames to create anomalies containing diverse types of objects (see Fig. 3). Using our proposed method to create pseudo-anomaly frames using TI data (details in Sec. 3.1), our generator learns features from normal frames *relative* to abnormal frames. Such pseudo-anomalies contain diverse foreign entities extracted from TI video frames allowing our generator to learn relative normalcy difference in a broad manner.

4. Experiments and Results

Implementation details. We implement our framework in PyTorch [92]. The generator $\mathcal{G}(\cdot)$ is an U-Net [93] adapted from [3] with a memory module at its bottleneck similar to [20]. The discriminator $\mathcal{D}(\cdot)$ and normalcy classifier $\mathcal{N}(\cdot)$ are Patch-GAN discriminators [94]. We provide more details of our implementation in Supplementary Material.

Evaluation details. We evaluate `zxVAD` under three training scenarios with respect to types of available source data: (1) *Both VAD and TI data are available*: $\mathcal{G}(\cdot)$ takes VAD videos and $\boldsymbol{\mathcal{O}}$ takes TI frames as input, (2) *Only one of the VAD or TI data are available*: Both $\mathcal{G}(\cdot)$ and $\boldsymbol{\mathcal{O}}$ take VAD or TI videos as input. We did not observe any performance gain empirically when $\mathcal{G}(\cdot)$ takes both VAD and TI videos as input, so we drop this case as it adds a computational burden. We compare `zxVAD` with [1, 2] using the area under ROC curve (AUC), model storage, total parameters, GPU energy consumption, inference time FPS, and GMACs.

Baselines. Since the problem of ‘cross-domain VAD without target domain training data adaptation’ is identified by us, we cannot find other methods which are designed for such a setup. The latest and closest baselines we found are rGAN [1] and MPN [2], which are designed for the xVAD task without needing strong priors from VAD frame object extraction. Since both methods report their performance under the proposed problem setup, we use them as our baselines. Even though we outperform strong prior based xVAD methods [46, 47, 54] without any such computationally expensive operation under our problem setup, we do not consider them as part of our baselines for fair comparisons with respect to [1, 2]. In Tab.

Table 3: **Comparison in Efficiency and Same-dataset testing.** We beat our baselines in most of the same-dataset testing, and outperform them in the listed efficiency metrics. *: GPU energy consumption is measured by testing on Ped2. †: rGAN [1] does not provide its official testing code for inference-time metric evaluation.

	Efficiency Metrics					Same Dataset Testing		
Method	**Parameters** (↓) (millions)	**GMACs** (↓)	**Energy** (↓) (Joules)*	**Storage** (↓) (MegaByte)	**FPS** (↑)	**SHT$_{\text{dc}}$**	**Ped2**	**SHT**
rGAN [1]	19.0	1384.52	–†	79.85	2.1	70.11	96.90	**77.90**
MPN [2]	12.7	55.09	10.65	53.14	166.8	67.47	96.20	73.80
zxVAD	**8.73**	**43.10**	**6.81**	**34.92**	**208.5**	**70.85**	**96.95**	71.60

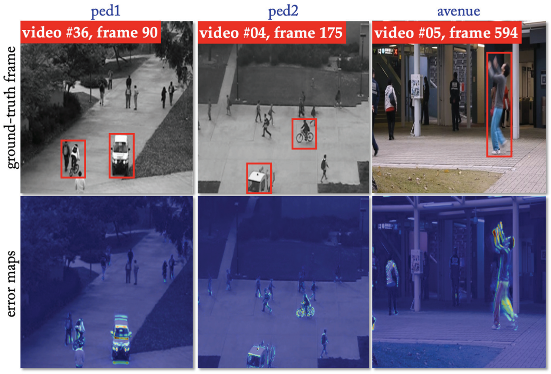


Figure 6: **Difference maps.** We show examples of cross-domain frame prediction comparison on three datasets (source: SHT). The lighter colors in difference map mean larger prediction error indicating anomalies. Red boxes indicate ground truth anomalies. Best viewed in color.

3, and 4, *paper* denotes results as reported and *code* denotes results computed using official code, if available.

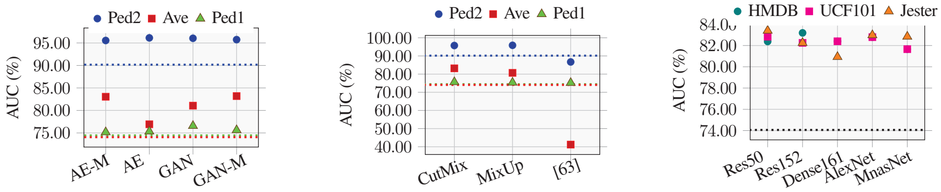
Ablation study. We show the ablation study of our proposed loss functions in zxVAD in Tab. 5(a) on the SHT$_{\text{dc}}$ dataset. Tab. 5(a) shows that each of our proposed loss functions contributes to the AUC, and jointly training with them all achieves the best AUC. In Fig. 5(b), we analyze different combinations of an autoencoder and generative adversarial network with (AE-M,GAN-M) and without our memory module (AE, GAN) as our zxVAD backbone. In Fig. 5(c), we analyze the impact of different mixing strategies (MixUp [96], CutMix [97] within our module $\mathcal{O}$ and compare with recent SOTA strategy called Patch [63] that proposes a pseudo-anomaly method. In Fig. 5(d), we analyze impact of changing $\mathcal{R}(\cdot)$ with (ResNet50, ResNet152 [75], DenseNet161 [98], AlexNet [99], MnasNet [100]) in zxVAD. Fig. 5 shows that regardless of the backbone choice, the pseudo-anomaly strategy, and the architecture of $\mathcal{R}(\cdot)$, zxVAD still outperforms the SOTA baselines in most settings, which supports that zxVAD is flexible w.r.t. these factors.

Same-dataset experiments. We compare zxVAD with [1, 2] on the SHT$_{\text{dc}}$, SHT and Ped2 datasets. Tab. 3 shows that zxVAD outperforms both baselines in AUC in such experiment. For example, zxVAD shows better generalization ability across different camera angles than the baselines in the SHT$_{\text{dc}}$ dataset with the least efficiency metrics like model parameters and GMACs. We also find that using extra TI data (HMDB and UCF101) can improve the AUC further compared to baselines (results in Supplementary Material).

Cross-dataset experiments. We compare zxVAD with [1, 2] under the cross-dataset setting. In the top two sections of Tab. 4, we train zxVAD with either the SHT or UCFC dataset with optional TI data and test it on the Ped1, Ped2, and Ave datasets. Tab. 4 shows that zxVAD outperforms both baselines in AUC under most settings, regardless of whether the extra TI data are used, which supports that zxVAD has better generalization ability across different datasets (with different types of anomalies under different scenes) than the baselines. For example, when our model is trained on the SHT dataset [16], it outperforms existing xVAD methods in the proposed problem setup in detecting anomalies like "chasing" and "brawling" in SHT's test set as well as anomalies like "bicycles" and " cars" in Ped1/Ped2's test set *without* performing any kind of adaptation on Ped1/Ped2's training set. This shows that our method is not specific to anomalies in the source domain, but generalizes well to target domain scenes during inference without adaptation. The bottom section of Tab. 4 shows that even without using any source domain VAD training data at all, zxVAD still outperforms [1, 2] in most settings by training with only TI data, which supports our proposed mechanism of using the TI data under the proposed problem setup. These encouraging results suggest that making use of TI data is a promising research direction for the zxVAD problem. Interestingly, when either $\mathcal{G}$ *or* $\mathcal{O}$ *or* both use TI data, it's not surprising to see slightly lower AUC than if both $\mathcal{G}$ *and* $\mathcal{O}$ use VAD relevant data, *i.e. more* relevant source data lead to *less* source-target domain gap, resulting in *better* AUC. This is confirmed by average AUC (Tab. 4) when source is *only* VAD: 84.26%, *VAD w/ TI*: 83.46%, and *only* TI: 82.30%. We also analyzed the impact of the amount videos needed when solely training with TI data with HMDB and UCF101 in zxVAD setup and found that even as little as ∼1.25% of UCF101 or ∼8% of HMDB is enough to outperform the SOTA (details in Supplementary Material). Following [1], we do not perform a cross-domain evaluation with Ped1/Ped2 as a source as the training dataset is too small to make reasonable conclusions. In Fig. 7, we show that zxVAD outperforms existing strong prior based unsupervised xVAD methods [46, 47, 54] that report cross-domain VAD testing performance when source domain data is SHT. This implies that zxVAD provides a computationally efficient and reduced supervision approach with no need for object extraction from videos (using YOLOv3 [48] in [47, 54] and CenterNet [101] in [46]) both in source and target domain, under the proposed problem setup. Compared to [47] (in Fig. 7) and [63] (in Fig. 5(c)), our untrained CNN based abnormal example generation strategy results in superior VAD for the proposed problem setup. Our "relative normalcy" learning approach optimizes the VAD model to learn features that differentiate normal events from (pseudo)-abnormal events, rather than focusing on learning *only* patterns of normal events as in prior xVAD works. Results in Tab. 4 (when zxVAD uses only TI data) validate this claim as zxVAD still outperforms SOTA on target VAD by learning

Loss Functions					AUC (%) on SHT$_{dc}$
$\mathcal{L}_{BB}$	$\mathcal{L}_{N}$	$\mathcal{L}_{RN}$	$\mathcal{L}_{AA}$	$\mathcal{L}_{RAA}$	
✓					68.32
✓	✓				68.99
✓	✓	✓			69.33
✓	✓	✓	✓		69.61
✓	✓	✓	✓	✓	**70.85**

(a) Loss ablation



(b) Impact of Backbones

(c) Mixing methods

(d) Impact of $\mathcal{R}(\cdot)$

Figure 5: **Component Analysis of** zxVAD. Tab. 5(a) shows the loss ablation using SHT$_{dc}$; Fig. 5(b) compares the cross-domain performance of zxVAD with different future-frame prediction backbones on three datasets (source: SHT); Fig. 5(c) compares the impact of different mixing strategies in module $\mathcal{O}$ with SOTA method [63] that also presents a pseudo-anomaly method on three datasets (source: SHT); Fig. 5(d) compares the impact of network $\mathcal{R}(\cdot)$ in module $\mathcal{O}$ on three TI datasets (source: SHT, target: Ave). *Dotted* lines in Fig. 5(b), 5(c) (three datasets), and 5(d) (one dataset) show SOTA (MPN [2] with Ped1: 74.45%, Ped2: 90.17%, Ave: 74.06%) in respective cross-domain VAD when source is SHT.

Table 4: **Cross-dataset testing.** Comparison with xVAD works that need no background-subtraction. The best and second best AUC are marked in **bold** and underline, respectively. ‡: For MPN [2], the publicized code [95] gives lower AUC than what was reported in their paper.

VAD Training Data (Input to $\mathcal{G}(\cdot)$)	Auxiliary Data (Input to $\mathcal{O}$)	Method	VAD Testing Data		
			Ped1	Ped2	Ave
SHT	N/A	rGAN [1] (paper)	73.10	81.95	71.43
SHT	N/A	MPN [2] (paper)	74.45	90.17	74.06
SHT	N/A	MPN [2] (code)‡	66.05	84.73	74.06
SHT	SHT	zxVAD **(ours)**	**76.14**	95.78	82.28
SHT	HMDB	zxVAD **(ours)**	75.62	95.74	**83.19**
SHT	UCF101	zxVAD **(ours)**	75.41	**95.80**	82.25
SHT	Jester	zxVAD **(ours)**	75.93	95.62	82.49
UCFC	N/A	rGAN [1] (paper)	66.87	62.53	64.32
UCFC	N/A	MPN [2] (paper)	75.52	86.04	**82.26**
UCFC	UCFC	zxVAD **(ours)**	**78.61**	**91.65**	81.11
UCFC	HMDB	zxVAD **(ours)**	78.02	87.66	81.50
UCFC	UCF101	zxVAD **(ours)**	76.27	86.80	81.45
UCFC	Jester	zxVAD **(ours)**	78.39	88.71	81.55
HMDB	HMDB	zxVAD **(ours)**	76.66	**91.53**	**81.92**
UCF101	UCF101	zxVAD **(ours)**	75.67	85.84	81.78
Jester	Jester	zxVAD **(ours)**	**78.12**	91.23	78.06

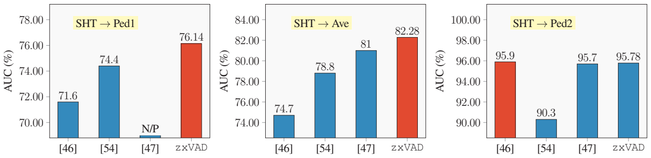


Figure 7: **Cross-dataset testing.** Comparison with VAD works that need background-subtraction. The best AUC is marked in red. 'N/P' in leftmost plot means 'Not Provided.'

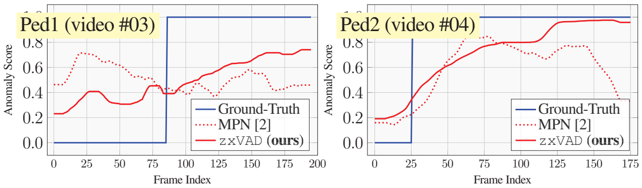


Figure 8: **Anomaly detection curve.** We compare our cross-domain (source: SHT) anomaly detection on Ped1/Ped2 against MPN [2]. Larger value in curves indicates possible anomalies.

such differentiating features from TI videos. [102] is a few-shot VAD method that puts together three off-the-shelf pre-trained models (YOLOv4 [103], AlphaPose [104], Flownet2 [105]) to perform xVAD. Even with such costlier storage, high training overhead, and strong priors from different distributions, zxVAD easily beats [102] by 11.76% (Ave), 13.85% (Ped2) with source as SHT, and 10.12% (Ave), 29.12% (Ped2) with source as UCFC with extremely less parameters and no initial priors. Finally, we provide qualitative evaluation under the cross-domain setting with anomaly curves of two testing videos of Ped1 and Ped2 when trained with SHT in Fig. 8, where zxVAD provides better cross-domain detection ability than MPN [2]. We also visualize difference maps in Fig. 6 (absolute error between ground truth and the predicted frame) that indicate the presence of anomalies by zxVAD in three datasets under cross-domain setting after training with SHT. We show more such qualitative results of zxVAD in Supplementary Material. In addition to the above, zxVAD achieves such results with much better inference-time efficiency than the baselines. Tab. 3 shows that zxVAD outperforms [1, 2] in model size, total parameters, GPU energy consumption (computed by pyJoules [106] following [107, 108]), and GMACs by 34.3%, 31.3%, 36.1%, and 21.76%, respectively.

5. Conclusion

We identify a new unsupervised xVAD problem of detecting anomalies in the target domain where no target domain training data are available. To tackle this problem, we propose a novel framework named 'Zero-shot Cross-domain Video Anomaly Detection' (zxVAD). zxVAD aims to learn features of normal activities in input videos by learning how such features are *relatively* different from features of pseudo-abnormal frames. Finally, zxVAD outperforms the SOTA baselines in most settings under common benchmarks not only in AUC but also in four efficiency metrics, regardless of whether the source domain VAD training data are available or not. Our results demonstrates the potential of task-irrelevant data as a promising direction for addressing the xVAD problem. As part of our future work, we will extend our method to enhance it's ability in directly localizing the anomaly in the videos.

Acknowledgement. AA and KP were supported by Mitsubishi Electric Research Laboratories. AA and ARC were partially supported by ONR grants N00014-19-1-2264 and N00014-18-1-2252.

References

[1] 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.

[2] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15425–15434, 2021.

[3] 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.

[4] Frederick Tung, John S Zelek, and David A Clausi. Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. *Image and Vision Computing*, 29(4):230–240, 2011.

[5] Shandong Wu, Brian E Moore, and Mubarak Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2054–2060. IEEE, 2010.

[6] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan. Semi-supervised adapted hmms for unusual event detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 611–618. IEEE, 2005.

[7] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008.

[8] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2921–2928. IEEE, 2009.

[9] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[10] 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. IEEE, 2010.

[11] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011.

[12] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. IEEE, 2011.

[13] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017.

[14] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE international conference on computer vision*, pages 2895–2903, 2017.

[15] 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.

[16] 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.

[17] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[18] 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.

[19] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020.

[20] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020.

[21] 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.

[22] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[23] Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, and Chengkun Wu. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13987–13998, 2022.

[24] Keval Doshi and Yasin Yilmaz. Multi-task learning for video surveillance with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3889–3899, 2022.

[25] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2249–2259, 2022.

[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14744–14754, 2022.

[27] MyeongAh Cho, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, 129:108703, 2022.

[28] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. A causal inference look at unsupervised video anomaly detection. 2022.

[29] Yuxin Zhang, Jindong Wang, Yiqiang Chen, Han Yu, and Tao Qin. Adaptive memory networks with self-supervised learning for unsupervised anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[30] Lin Wang, Haishu Tan, Fuqiang Zhou, Wangxia Zuo, and Pengfei Sun. Unsupervised anomaly video detection via a double-flow convlstm variational autoencoder. *IEEE Access*, 10:44278–44289, 2022.

[31] Guodong Shen, Yuqi Ouyang, and Victor Sanchez. Video anomaly detection via prediction network with enhanced spatio-temporal memory exchange. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3728–3732. IEEE, 2022.

[32] Yuqi Ouyang, Guodong Shen, and Victor Sanchez. Look at adjacent frames: Video anomaly detection without offline training. *arXiv preprint arXiv:2207.13798*, 2022.

[33] Joo-Yeon Lee, Woo-Jeoung Nam, and Seong-Whan Lee. Multi-contextual predictions with vision transformer for video anomaly detection. *arXiv preprint arXiv:2206.08568*, 2022.

[34] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

[35] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347, 2020.

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

[37] 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.

[38] 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.

[39] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, pages 3023–3030, 2019.

[40] 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. *arXiv preprint arXiv:2101.10030*, 2021.

[41] Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 173–183, 2021.

[42] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 2021.

[43] 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*, pages 12742–12752, 2021.

[44] Jaeyoo Park, Junha Kim, and Bohyung Han. Learning to adapt to unseen abnormal activities under weak supervision. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[45] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *arXiv preprint arXiv:2207.08003*, 2022.

[46] Pankaj Raj Roy, Guillaume-Alexandre Bilodeau, and Lama Seoud. Predicting next local appearance for video anomaly detection. *arXiv preprint arXiv:2106.06059*, 2021.

[47] 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. *arXiv preprint arXiv:2008.12328*, 2020.

[48] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[49] Takafumi Moriya, Tomohiro Tanaka, Takahiro Shinozaki, Shinji Watanabe, and Kevin Duh. Evolution-strategy-based automation of system development for high-performance speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):77–88, 2018.

[50] Jiayi Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah. Auptimizer-an extensible, open-source framework for hyperparameter tuning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 339–348. IEEE, 2019.

[51] Alevtina Krotova, Armin Mertens, and Marc Scheufen. Open data and data sharing: An economic analysis. Technical report, IW-Policy Paper, 2020.

[52] Michael Mattioli. Disclosing big data. *Minn. L. Rev.*, 99:535, 2014.

[53] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[54] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.

[55] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017.

[56] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 184–192, 2020.

[57] 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.

[58] 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*, pages 13588–13597, 2021.

[59] Chun-Yu Chen and Yu Shao. Crowd escape behavior detection and localization based on divergent centers. *IEEE Sensors Journal*, 15(4):2431–2439, 2014.

[60] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020.

[61] 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.

[62] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pages 207–214, 2021.

[63] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. *arXiv preprint arXiv:2110.09742*, 2021.

[64] Masoud Pourreza, Bahram Mohammadi, Mostafa Khaki, Samir Bouindour, Hichem Snoussi, and Mohammad Sabokrou. G2d: Generate to detect anomaly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2003–2012, 2021.

[65] Yutao Hu, Xin Huang, and Xiaoyan Luo. Adaptive anomaly detection network for unseen scene without fine-tuning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 311–323. Springer, 2021.

[66] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[67] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[68] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 113–124, 2019.

[69] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[70] Jack W Rae, Jonathan J Hunt, Tim Harley, Ivo Danihelka, Andrew Senior, Greg Wayne, Alex Graves, and Timothy P Lillicrap. Scaling memory-augmented neural networks with sparse reads and writes. *arXiv preprint arXiv:1610.09027*, 2016.

[71] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016.

[72] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[73] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[74] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[76] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017.

[77] Yun-Hao Cao and Jianxin Wu. A random cnn sees objects: One inductive bias of cnn and its applications. *Proceedings of the 6th AAAI Conference on Artificial Intelligence*, 2022.

[78] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

[79] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

[80] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[81] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Pytorch metric learning, 2020.

[82] Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11416–11425, 2021.

[83] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11772–11781, 2021.

[84] Arman Afrasiyabi, Jean-Francois Lalonde, and Christian Gagne. Mixture-based feature space learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9041–9051, 2021.

[85] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[86] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[87] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[88] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20143–20153, 2022.

[89] Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. Zero-shot deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 764–781, 2018.

[90] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2635–2644, 2015.

[91] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[92] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[93] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[94] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[95] The official codes of the CVPR21 paper: Learning Normal Dynamics in Videos with Meta Prototype Network. `https://github.com/ktr-hubrt/MPN`.

[96] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[97] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[98] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[99] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[100] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

[101] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[102] Keval Doshi and Yasin Yilmaz. A modular and unified framework for detecting and localizing video anomalies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3982–3991, 2022.

[103] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[104] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[105] 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.

[106] pyJoules. `https://pypi.org/project/pyJoules/`.

[107] Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2021.

[108] Krishnateja Killamsetty, Durga Sivasubramanian, Baharan Mirzasoleiman, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: A gradient matching based data subset selection for efficient learning. *Proceedings of the 38th International Conference on Machine Learning*, 2021.