This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Video Anomaly Detection via Spatio-Temporal Pseudo-Anomaly Generation : A Unified Approach

Ayush K. Rai[1], Tarun Krishna[*,1], Feiyan Hu[*,1], Alexandru Drimbarean[2]
Kevin McGuinness[†,1], Alan F. Smeaton[†,1], Noel E. O'Connor[†,1]
[1]Insight SFI Centre for Data Analytics, Dublin City University [2]Tobii Corporation, Galway
ayush.rai3@mail.dcu.ie

Abstract

Video Anomaly Detection (VAD) is an open-set recognition task, which is usually formulated as a one-class classification (OCC) problem, where training data is comprised of videos with normal instances while test data contains both normal and anomalous instances. Recent works have investigated the creation of pseudo-anomalies (PAs) using only the normal data and making strong assumptions about real-world anomalies with regards to abnormality of objects and speed of motion to inject prior information about anomalies in an autoencoder (AE) based reconstruction model during training. This work proposes a novel method for generating generic spatio-temporal PAs by inpainting a masked out region of an image using a pre-trained Latent Diffusion Model and further perturbing the optical flow using mixup to emulate spatio-temporal distortions in the data. In addition, we present a simple unified framework to detect real-world anomalies under the OCC setting by learning three types of anomaly indicators, namely reconstruction quality, temporal irregularity and semantic inconsistency. Extensive experiments on four VAD benchmark datasets namely Ped2, Avenue, ShanghaiTech and UBnormal demonstrate the effectiveness of our work against other existing state-of-the-art PAs generation and reconstruction based methods under the OCC setting. Our analysis also examines the transferability and generalisation of PAs across these datasets, offering valuable insights by identifying real-world anomalies through PAs. Our results can be reproduced on github.

1. Introduction

Video Anomaly Detection [4, 5, 21–23, 31, 35, 44, 46, 59, 60, 72, 81, 96, 100] refers to the task of discovering the unexpected occurrence of events that are distinct and follow a deviation from known normal patterns. The rarity of anomalies in the real-world and the unbounded nature (open-set recognition [20]) of their diversities and complexities have led to unbalanced training datasets for VAD making it an extremely challenging task. Therefore VAD is commonly addressed as a OCC problem where only normal data is available to train a model [4, 5, 21, 23, 26, 46, 50, 51, 59, 106].

Reconstruction based approaches exploiting an AE are usually adopted to tackle the OCC task [4, 5, 23, 59]. The intuition behind this is that during training, the AE would learn to encode normal instances in its feature space with the assumption that during the test phase a high reconstruction error would correspond to an anomaly and a low reconstruction error would indicate normal behaviour. Contrary to this, [4, 23, 96] observed that when trained in this setting, the AE learns to reconstruct anomalies with high accuracy resulting in a low reconstruction error in the testing phase. Hence, the capability of the AE to distinguish normal and anomalous instances is greatly diminished (Figure 1a in [4]).

[23, 59] introduced a memory-based AE to restrict the reconstruction capability of the AE by recording prototypical normal patterns during training in the latent space therefore shrinking the capability of the AE to reconstruct anomalous data. However, such methods are highly sensitive to memory size. A small-sized memory may hinder reconstruction of normal data as memorising normal patterns can be interpreted as severely limiting the reconstruction boundary of the AE, resulting in failure to reconstruct even the normal events during the testing phase (Figure 1b in [4]).

Astrid *et al*. [4] proposed the generation of two types of PAs (patch based and skip-frame based) to synthetically simulate pseudo-anomalous data from normal data and further introduced a novel training objective for the AE to force the reconstruction of only normal data even if the input samples are anomalous. Patch based PAs are generated by inserting a patch of a specific size and orientation from an intruder dataset (e.g. CIFAR-100) using the SmoothMixS [38] data augmentation method while in order to create skip-frame based PAs, a sequence of frames is sampled with irregular strides to create anomalous movements in the sequence. The

[*] Equal contribution
[†] Equal supervision

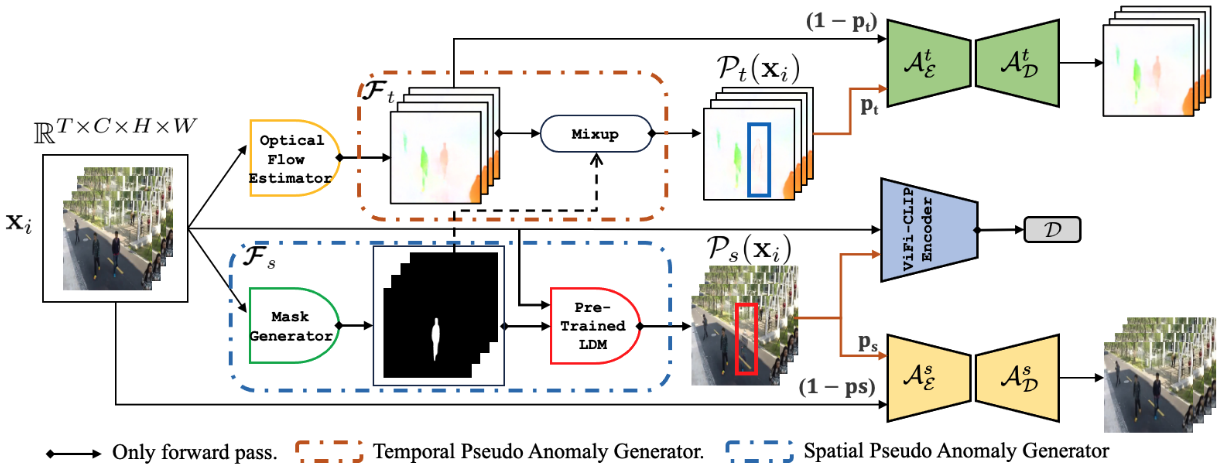
$\mathbb{R}^{T\times C\times H\times W}$
$\mathbf{x}_i$
Optical Flow Estimator
$\mathcal{F}_t$
Mixup
$\mathcal{P}_t(\mathbf{x}_i)$
$(1-\mathbf{p}_t)$
$\mathbf{p}_t$
$\mathcal{A}^t_\mathcal{E}$
$\mathcal{A}^t_\mathcal{D}$
ViFi-CLIP Encoder
$\mathcal{D}$
$\mathcal{F}_s$
Mask Generator
Pre-Trained LDM
$\mathcal{P}_s(\mathbf{x}_i)$
$\mathbf{p}_s$
$(1-\mathbf{ps})$
$\mathcal{A}^s_\mathcal{E}$
$\mathcal{A}^s_\mathcal{D}$
Only forward pass.
Temporal Pseudo Anomaly Generator.
Spatial Pseudo Anomaly Generator

Figure 1. The overall architecture of our approach consists of spatio-temporal PAs generators. Spatial PAs generator (eq. 2) : $\mathcal{F}_s(\text{stack}(\mathbf{x}, \mathbf{x}\odot \mathbf{m}, \mathbf{m}); \theta)$and temporal PAs (eq. 3) : $\mathcal{F}_t(\phi(\mathbf{x_t}, \mathbf{x_{(t+1)}}))$. The spatial and temporal PAs are sampled with probability p_s and p_t respectively. Our VAD framework unifies estimation of reconstruction quality (eq. 4), temporal irregularity (eq. 5) and semantic inconsistency.

intuition behind this training procedure is based on limiting the reconstruction boundary of the AE near the boundaries of the normal data resulting in more distinctive features between normal and anomalous data (Figure 1c in [4]). A notable limitation of the approach proposed in Astrid *et al.* [4] is its heavy reliance on a predefined set of assumptions and inductive biases. These assumptions encompass various aspects, including the specific intruding dataset selected for patch insertion, the patch's size and orientation, and the idea that altering the movement speed by skipping frames could introduce temporal irregularities into the normal data.

With such assumptions, there is no guarantee that the test anomalies which comprise of an unbounded set of possible anomalous scenarios would comply with pseudo-anomalous samples. This creates a need for more generic solutions for creating PAs from the normal data. Since VAD is an open-set recognition problem and anomalies present an inexhaustible set of possibilities, every pseudo-anomaly synthesiser carries strong or weak inductive biases and thus it is inherently challenging to emulate real-world anomalies through PAs. Furthermore, there are other challenges, such as the fact that certain normal behaviours are rare but possible and therefore not well represented in the normal data. This presents an interesting research question: *"Is it possible to synthetically generate generic PAs by introducing spatio-temporal distortions into normal data in order to detect real-world anomalies effectively?, and importantly, can such PAs transfer across multiple VAD datasets?"*

Our work is motivated by [4] and extends it by addressing its drawbacks and proposing a more generic PAs generator. We focus on generating PAs by injecting two different types of anomaly indicators, the first being distortion added through image inpainting performed by a pre-trained latent diffusion model (LDM) [66], the second being the addition of temporal irregularity through perturbation of the optical flow [95] using mixup [103]. Our simple VAD pipeline focuses on reconstructing the spatio-temporal PAs and also measures the semantic inconsistency between normal samples and PAs using semantically rich ViFi-CLIP [64] features. This *unifies estimation of reconstruction quality, temporal irregularity and semantic inconsistency* under one framework. We conduct an extensive study on understanding the generalisation and transferability of such PAs over real-world anomalies. Overall, our main contributions are:

- We propose a novel and generic spatio-temporal pseudo-anomaly generator for VAD encompassing inpainting of a masked out region in frames using an LDM and applying mixup augmentation to distort the optical flow.
- We introduce a simple unified VAD framework that measures and aggregates three different indicators of anomalous behaviour namely reconstruction quality, temporal irregularity and semantic inconsistency in an OCC setting.
- Extensive experiments on *Ped2, Avenue, ShanghaiTech* and *UBnormal* show that our method though not objectively state-of-the-art (SOTA) but achieves comparable performance as other existing SOTA PAs generation and reconstruction based methods under the OCC setting (Table 1, 2) without any end-to-end finetuning or any post-processing. This validates that our method is a generic video anomaly detector and our spatio-temporal PAs generation process is transferable across multiple datasets.

2. Related Work

2.1. Restricting Reconstruction Capacity of an AE

A standard approach to address VAD is to adopt an OCC strategy by training an AE model to reconstruct the input data [4, 23, 26, 50, 51, 59, 106]. During training, only normal inputs are used for learning the AE with the assumption that reconstruction of anomalies during testing would yield a higher reconstruction error. However, in practice it has been shown that the AE can also reconstruct anomalous data [4, 23, 96]. [23, 59] mitigated this issue by augmenting the AE with memory-based techniques in the latent space to restrict the reconstruction capability of an AE. However the performance of such methods are directly impacted by the choice of the memory size, which may over-constrain the reconstruction power of the AE resulting in poor reconstruction of even the normal events during testing.

To alleviate this issue, [4, 5] utilised data-heuristic based PAs built on strong assumptions to limit the reconstruction capacity of the AE. Patch-based PAs were generated by inserting a patch from an intruding dataset (CIFAR-100) into the normal data by using techniques such as SmoothMixS [38]. For modeling motion-specific anomalous events, PAs were generated by skipping frames with different strides to induce temporal irregularity. The training configuration was set up to minimise the reconstruction loss of the AE with respect to the normal data only. PAs can be interpreted as a type of data-augmentation [6, 37], where instead of creating more data of the same distribution, pseudo-anomalous data is created that belongs to a near-distribution i.e. between the normal and anomaly distributions. [76, 104] adopted adversarial training to generate augmented inputs, which were also effective as an adversarial example for the model.

Our method falls into the category of restricting the reconstruction capability of an AE. Inspired by the method introduced in [4], we propose a novel technique for simulation of generic spatio-temporal PAs without making bold assumptions about dataset specific anomalies.

2.2. Generative Modeling

Generative models have been used to generate out of distribution (OOD) data for various applications in semi-supervised learning (Bad GAN [9], Margin GAN [14]), anomaly detection (Fence GAN [56]), OOD detection (BDSG [11, 18]), medical anomaly detection [83] and novelty detection [55]. However, such methods mostly work with low dimensional data and are not suitable for generating OOD data for VAD. OGNet [96, 99] and G2D [60] exploit a GAN-based generator and discriminator for VAD. During the first phase of training, a pre-trained state of the generator is used to create PAs while in the second phase, binary classification is performed to distinguish between normal and PAs samples.

Several VAD works have exploited DMs though their specific methodologies and goals vary. [77, 78, 90] focus on reconstruction and prediction of spatio-temporal and compact motion features extracted from 3D-ResNet/3D-ResNext based encoders using an end-to-end trainable DM. We design our model from the perspective of generating generic spatio-temporal PAs where a generative model (pre-trained LDM) is availed to generate spatial PAs while the mixup method is exploited to create temporal PAs from optical flow.

2.3. Other VAD Methods

Non-Reconstruction Based Methods: Various non-reconstruction based methods have also been proposed which derive their anomaly scores from various different indicators of anomaly in addition to reconstruction loss. The work presented in [44] utilised a future frame prediction task for VAD and estimated optical flow and gradient loss as supplementary cues for anomalous behaviour. [21, 31] performed object detection as a pre-processing step under the assumption that anomalous events are always object-centric. Several other works added optical flow components [35, 41] to detect anomalous motion patterns and a binary classifier [60, 96] to estimate anomaly scores. In our work, we also use a segmentation mask and optical flow to generate corresponding spatial and temporal PAs during the training phase. However during inference we do not carry out any object detection and perform anomaly detection solely based on reconstruction of whole images and optical flow.

Non-OCC methods: [21] introduced a self-supervised method where different pretext tasks such as arrow of time, middle-box prediction, irregular motion discrimination and knowledge distillation were jointly optimised for VAD. [81] adopted a self-supervised single pre-text task of solving decoupled temporal and spatial jigsaw puzzles Several works have also addressed the VAD problem as a weakly supervised problem through multiple instance learning [72, 85, 102, 108]. Unsupervised VAD methods involve the cooperation of two networks through an iterative process for pseudo-label generation [43, 58, 97, 98, 100, 101]. Zero-shot VAD was introduced in [3] where a model was trained on the source domain to detect anomalies in a target domain without any domain adaptation. USTN-DSC [92] a proposed video event restoration framework for VAD while EVAL [68] presented a technique for video anomaly localisation allowing for human interpretable explanations.

3. Method

3.1. Preliminaries

Latent Diffusion Models (LDMs): Diffusion Probabilistic Models (DMs) [29, 70, 71] are a class of probabilistic generative models that are designed for learning a data distribution $p_{\text{data}}(\mathbf{x})$. DMs iteratively denoise a normally distributed variable by learning the reverse process of a fixed Markov Chain of length T through a denoising score matching objective [71] given by:

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}},\tau\sim p_\tau,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\mathbf{y}-\mathbf{f}_\theta(\mathbf{x}_\tau;\mathbf{c},\tau)||_2^2], \qquad (1)$$

where $\mathbf{x}\sim p_{data}$, the diffused input can be constructed by $\mathbf{x}_\tau=\alpha_\tau\mathbf{x}+\sigma_\tau\epsilon$, $\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ and is fed into a denoiser

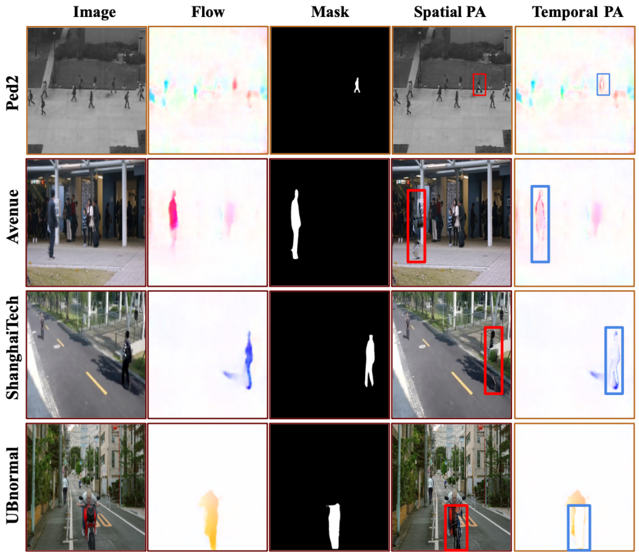

Figure 2. Visualisation of spatial and temporal PAs, using segmentation masks. This approach also works with random masks.

model $\mathbf{f}_\theta$, $(\sigma_\tau, \alpha_\tau)$ denotes the noise schedule parameterised by diffusion-time τ, p_τ is a uniform distribution over τ, $\mathbf{c}$ denotes conditioning information and the target vector $\mathbf{y}$ is either the random noise ϵ or $\mathbf{v} = \alpha_\tau \epsilon - \sigma_\tau \mathbf{x}$. The forward diffusion process corresponds to gradual addition of the gaussian noise to $\mathbf{x}$ such that the logarithmic signal-to-noise ratio $\lambda_\tau = \log(\alpha_\tau^2/\sigma_\tau^2)$ monotonically decreases.

LDMs [66] were proposed to make standard DMs efficient by training a VQGAN [19] based model to project input images i.e. $\mathbf{x} \sim p_{data}$ into a spatially lower dimensional latent space of reduced complexity and then reconstructing the actual input with high accuracy. In particular, a regularised AE [66] is used to reconstruct the input $\mathbf{x}$ such that the reconstruction is given by : $\hat{\mathbf{x}} = \mathbf{f_{de}} \circ \mathbf{f_{en}}(\mathbf{x})$[1] $\approx \mathbf{x}$, where $\mathbf{f_{en}}$ and $\mathbf{f_{de}}$ denotes encoder and decoder respectively. Furthermore an adversarial objective is added using a patch-based discriminator [33] to ensure photorealistic reconstruction. DM is then trained in the latent space by replacing $\mathbf{x}$ with its latent representation $\mathbf{z} = \mathbf{f_{en}}(\mathbf{x})$ in eq. (1). This leads to reduction in number of learnable parameters and memory.

3.2. Generating Spatial-PAs

Real world anomalies are highly context specific without having a ubiquitous definition. Ramachandra *et al.* [63] loosely define them as, the “occurrence of unusual appearance and motion attributes or the occurrence of usual appearance and motion attributes at an unusual locations or times”. Examples of such cases include: an abandoned object in a crowded area or suspicious behaviour of an individual. We address this notion of occurrence of unusual appearance attributes through generation of spatial PAs.

Since LDMs achieve state-of-the-art performance on the image inpainting task, they can be exploited as a spatial PAs generator. In particular, we hypothesise that an off-the-shelf pre-trained LDM model [66] without any finetuning on VAD datasets can inpaint the image with enough spatial distortion that can serve as spatially pseudo-anomalous samples for training a VAD model. We follow the mask generation strategy proposed in LAMA [74] to generate both randomly shaped and object segmentation masks $\mathbf{m}$. We concatenate image $\mathbf{x}$, masked image $\mathbf{x} \odot \mathbf{m}$ [2] and mask $\mathbf{m}$ over the channel dimension and give this 7 channel input to UNet [67]. We denote the normal data samples as $\mathbf{x}$ unless otherwise explicitly stated. The spatial PAs $\mathcal{P}_s(\mathbf{x})$ is given by:

$$\mathcal{P}_s(\mathbf{x}) = \mathcal{F}_s(\text{stack}(\mathbf{x}, \mathbf{x} \odot \mathbf{m}, \mathbf{m}); \theta), \tag{2}$$

where $\mathcal{F}_s$ is the inpainting model that uses latent diffusion with pre-trained model parameters θ. Some examples of the spatial PAs are shown in Figure 2. We avoid regress tuning of LDM hyperparameters due to limited available compute.

3.3. Generating Temporal-PAs

We address the notion of unusual motion occurrences (such as person falling to ground) through the generation of temporal PAs. Various video diffusion models [27, 30, 79] have been proposed, which can be exploited to induce temporal irregularity in the video. However due to limited computational resources, we introduce a simple but effective strategy for the generation of temporal PAs by applying a vicinal risk minimisation technique mixup [103] to the optical flow of the normal videos. More specifically, given a *normal* video $\mathbf{v}$, its frame $\mathbf{x_t}$, and its corresponding segmentation mask $\mathbf{m_t}$ and another consecutive frame $\mathbf{x_{(t+1)}}$, we compute the optical flow $\phi(\mathbf{x_t}, \mathbf{x_{(t+1)}})$ using the TVL1 alogrithm [95]. For simplification, we use ϕ as an alias to represent $\phi(\mathbf{x_t}, \mathbf{x_{(t+1)}})$. Let us consider a rectangular patch $\mathbf{p}'$ in ϕ corresponding to the mask $\mathbf{m_t}$ in the frame $\mathbf{x_t}$ with dimensions μ_h and μ_w. In order to perturb the optical flow ϕ, we take another rectangular patch $\mathbf{p_r}'$ at a random location in ϕ with the same dimensions as $\mathbf{p}'$ and apply mixup to yield $\hat{\mathbf{p}}$, which is a convex combination of $\mathbf{p}'$ and $\mathbf{p_r}'$ given by : $\hat{\mathbf{p}} = \lambda \mathbf{p}' + (1-\lambda)\mathbf{p_r}'$, where λ is sampled from a beta distribution with $\alpha = 0.4$ as in [103]. We denote the temporal PAs as $\mathcal{P}_t(\mathbf{x})$ given by:

$$\mathcal{P}_t(\mathbf{x}) = \mathcal{F}_t(\phi(\mathbf{x_t}, \mathbf{x_{(t+1)}})), \tag{3}$$

where $\mathcal{F}_t$ is the temporal PAs generator. Some examples of temporal PAs are depicted in Figure 2. It is important to note that our PAs generation method does not explicitly require segmentation masks, it can also generate PAs using random masks. Since segmentation masks carry semantic meaning,

[1] $\circ$: denotes function composition

[2] $\odot$: denotes point-wise multiplication

using them enables generation of more semantically informative PAs as further validated by our experiments.

3.4. Reconstruction Model

During training regardless of the input ($\mathcal{I}$) i.e normal ($\mathbf{x}/\phi$) or PAs ($\mathcal{P}_s(\mathbf{x})/\mathcal{P}_t(\mathbf{x})$), the network is forced to reconstruct only the normal input using a 3D-CNN (Convolutional Neural Network) based AE model adapted from the convolution-deconvolution network proposed by [23] (Table 2 in supplementary material (supp.)).

We train two different AEs with the aim of limiting their reconstruction capacity by exposing them to spatial and temporal PAs. We represent the spatial (temporal) AE by $\mathcal{A}^s(\mathcal{A}^t)$ with $\mathcal{A}^s_{\text{e}}(\mathcal{A}^t_{\text{e}})$ and $\mathcal{A}^s_{\text{de}}(\mathcal{A}^t_{\text{de}})$ denoting its encoder and decoder respectively. The reconstruction output of $\mathcal{A}^s$ is given by : $\hat{\mathbf{x}} = \mathcal{A}^s_{\text{de}} \circ \mathcal{A}^s_{\text{e}}(\mathbf{x})$ while the reconstruction output of $\mathcal{A}^t$ is computed by : $\hat{\phi} = \mathcal{A}^t_{\text{de}} \circ \mathcal{A}^t_{\text{e}}(\phi)$. In order to train $\mathcal{A}^s$ and $\mathcal{A}^t$, PAs ($\mathcal{P}_s(\mathbf{x})$ or $\mathcal{P}_t(\mathbf{x})$) are given as respective inputs with a probability p_s (or p_t) while the normal data is provided as input with probability of $(1 - p_s)$ (or $(1 - p_t)$). p_s (or p_t) is a hyperparameter to control the ratio of PAs to normal samples. Overall, the loss for $\mathcal{A}^s$ and $\mathcal{A}^t$ is calculated as:

$$\mathcal{L}_{\mathcal{A}^{(s)}} = \frac{1}{\Pi} \begin{cases} ||\hat{\mathbf{x}} - \mathbf{x}||_2^2 & \text{if } \mathcal{I} = \mathbf{x} \\ ||\hat{\mathcal{P}}_s(\mathbf{x}) - \mathbf{x}||_2^2 & \text{if } \mathcal{I} = \mathcal{P}_s(\mathbf{x}), \end{cases} \tag{4}$$

$$\mathcal{L}_{\mathcal{A}^{(t)}} = \frac{1}{\Pi} \begin{cases} ||\hat{\phi} - \phi||_2^2 & \text{if } \mathcal{I} = \phi \\ ||\hat{\mathcal{P}}_t(\mathbf{x}) - \phi||_2^2 & \text{if } \mathcal{I} = \mathcal{P}_t(\mathbf{x}), \end{cases} \tag{5}$$

where $1/\Pi$ is normalisation factor, $\Pi = \mathcal{T} \times C \times H \times W$ and $||.||_2$ is the $\mathcal{L}_2$ norm, where $\mathcal{T}, C, H$ and W are the number of frames, channels, height, and width of frames in the input sequence ($\mathcal{I}$), respectively. *The design of $\mathcal{A}^s(\mathcal{A}^t)$ is purposefully chosen to be simplistic (3D-CNN) instead of complex models (vision transformers [17], 3D ResNets/ResNexts [25, 86]) to explore the degree to which the results can be enhanced by incorporating simple methods.*

3.5. Estimating Semantic Inconsistency

While measuring the spatial reconstruction quality and temporal irregularity between normal and anomalous data is essential for real-world VAD, it is also crucial to learn and estimate the semantic inconsistency (degree of misalignment of semantic visual patterns and cues) between normal and anomalous samples (e.g. abnormal object in the crowded scene). In practice, to emulate this idea in our approach, we extract frame-level semantically rich features from the ViFi-CLIP [64] model (pre-trained on Kinetics-400 [36]) and perform binary classification between normal data samples $\mathbf{x}$ and spatial pseudo-anomalies $\mathcal{P}_s(\mathbf{x})$ using a discriminator $\mathcal{D}$, (Table 1 in supp.), which can be viewed as an auxiliary component to AEs. Intuitively, it is highly likely that latent space representation of PAs will be semantically inconsistent to the normal scenarios.

4. Experimental Setup

4.1. Implementation Details

a). Training Spatial ($\mathcal{A}^s$) and Temporal ($\mathcal{A}^t$) AE's: We closely follow the training procedure described in [4] to train $\mathcal{A}^s$ and $\mathcal{A}^t$. The architecture of $\mathcal{A}^s$ and $\mathcal{A}^t$ is adapted from [23], however instead of relying on single channel image as input we use all 3 channels. $\mathcal{A}^s$ and $\mathcal{A}^t$ were trained on respective datasets from scratch with the objective defined in eq. 4 and eq. 5 respectively on 2 NVIDIA GeForce 2080 Ti GPUs with effective batch size ($\mathcal{B}$) of 24 distributed across the GPUs (12 each). The input to $\mathcal{A}^s$ and $\mathcal{A}^t$ is of size $(\mathcal{B} \times \mathcal{T} \times 3 \times 256 \times 256)$, where $\mathcal{T} = 16$. The spatial and temporal PAs were sampled by probability $p_s = 0.4$ and $p_t = 0.5$ respectively. $\mathcal{A}^s$ is trained with Adam optimiser for 25 epochs with a learning rate of 1e-4. During training, the reconstruction loss is calculated across all 16 frames of the sequence. The training of the $\mathcal{A}^t$ follows a similar procedure, however the input to the model is the optical flow representing normal events i.e ϕ and temporal PAs $\mathcal{P}_t(\mathbf{x})$.
b). Training the Discriminator ($\mathcal{D}$): During the training phase, the input to $\mathcal{D}$ has a batch size of 16 and feature dimension of 512. The model was trained using a SGD optimiser with a learning rate of 0.02, momentum of 0.9 and weight decay of 10^{-3} for 20 epochs. The groundtruth for normal and PAs samples are given labels 0 and 1 respectively. See section 2 (supp.) for ViFi-CLIP [64] feature extraction and additional details. Figure 1 depicts the complete pipeline.

4.2. Inference

During inference (Figure 2 in supp.), our goal is to *temporally localise the anomaly* by measuring all three types of anomaly indicators of all frames in the test video in the given dataset i.e reconstruction quality, temporal irregularity and semantic inconsistency. Therefore, our anomaly score holistically combines these aspects to gain deeper insights into real-world anomalies in videos.

In order to measure the reconstruction quality, we follow the recent works of [12, 44, 59], which utilise normalised Peak Signal to Noise Ratio P_t (PSNR) between the test input frame at time t and its reconstruction from $\mathcal{A}^s$ to calculate the anomaly score $\omega_1^{(t)}$. The input to $\mathcal{A}^s$ is given in a non-overlapping sliding window fashion with dimensions $1 \times 16 \times 3 \times 256 \times 256$, where batch size is 1 and 16 (window size) represents number of frames. At test time, only the 9^{th} frame of a sequence is considered for anomaly score calculation as in [4]. For measuring temporal irregularity, a similar strategy is followed as for frames but instead of measuring the PSNR, the normalised $\mathcal{L}_2$ loss (denoted by

Table 1. Micro AUC score comparison between our approach and state-of-the-art methods on test split of Ped2 [42], Avenue (Ave) [47] and ShanghaiTech (Sh) [53]. Best and second best performances are highlighted as **bold** and underlined, in each category and dataset.

	Methods	Ped2 [42]	Ave [47]	Sh [53]
Miscellaneous	OLED [34]	99.02%	-	-
	AbnormalGAN [65]	93.50%	-	-
	Smeureanu *et al.* [69]	-	84.60%	-
	AMDN [88, 89]	90.80%	-	-
	STAN [39]	96.50%	87.20%	-
	MC2ST [45]	87.50%	84.40%	-
	Ionescu *et al.* [32]	-	88.90%	-
	BMAN [40]	96.60%	**90.00%**	76.20%
	AMC [57]	96.20%	86.90%	-
	Vu *et al.* [80]	**99.21%**	71.54%	-
	DeepOC [84]	-	86.60%	-
	TAM-Net [35]	98.10%	78.30%	-
	LSA [1]	95.40%	-	72.50%
	Ramachandra *et al.* [62]	94.00%	87.20%	-
	Tang *et al.* [75]	96.30%	85.10%	73.00%
	Wang *et al.* [82]	-	87.00%	**79.30%**
	OGNet [96]	98.10%	-	-
	Conv-VRNN [48]	96.06%	85.78%	-
	Chang *et al.* [8]	96.50%	86.00%	73.30%
	USTN-DSC [92]	98.10%	89.90%	73.8%
	EVAL [68]	-	86.06%	76.63%
Object-centric	MT-FRCN [28]	92.20%	-	-
	Ionescu *et al.* [31] [3]	94.30%	87.40%	78.70%
	Doshi and Yilmaz [15, 16]	97.80%	86.40%	71.62%
	Sun *et al.* [73]	-	89.60%	74.70%
	VEC [94]	97.30%	90.20%	74.80%
	Georgescu *et al.* [22]	**98.70%**	**92.30%**	**82.70%**

	Methods	Ped2 [42]	Ave [47]	Sh [53]
Non deep learning	MDT [54]	82.90%	-	-
	Lu *et al.* [47]	-	**80.90%**	-
	AMDN [89]	90.80%	-	-
	Del Giorno *et al.* [10]	-	78.30%	-
	LSHF [105]	**91.00%**	-	-
	Xu *et al.* [87]	88.20%	-	-
	Ramachandra and Jones [61]	88.30%	72.00%	-
Prediction	Frame-Pred [44]	95.40%	85.10%	72.80%
	Dong *et al.* [13]	95.60%	84.90%	73.70%
	Lu *et al.* [49]	96.20%	85.80%	**77.90%**
	MNAD-Pred [59]	97.00%	88.50%	70.50%
	AnoPCN [93]	96.80%	86.20%	73.60%
	AMMC-Net [7]	96.90%	86.60%	73.70%
	DLAN-AC [91]	**97.60%**	**89.90%**	74.70%
Reconstruction	AE-Conv2D [26]	90.00%	70.20%	60.85%
	AE-Conv3D [107]	91.20%	71.10%	-
	AE-ConvLSTM [52]	88.10%	77.00%	-
	TSC [53]	91.03%	80.56%	67.94%
	StackRNN [53]	92.21%	81.71%	68.00%
	MemAE [24]	94.10%	83.30%	71.20%
	MNAD-Recon [59]	90.20%	82.80%	69.80%
	Baseline (without PAs)	92.49%	81.47%	71.28%
	STEAL Net [5]	**98.40%**	**87.10%**	73.70%
	LNTRA Astrid *et al.* [4] - Patch based	94.77%	84.91%	72.46%
	LNTRA Astrid *et al.* [4] - Skip-frame based	96.50%	84.67%	**75.97%**
	Ours w/o $\mathcal{D}$	93.52%	86.51%	71.76%
	Ours w/ $\mathcal{D}$	93.53%	86.61%	71.65%

Table 2. Micro AUC score comparison between our approach and existing state-of-the-art methods on val split of UBnormal [2].

Reconstruction Methods	UBnormal [2]
Baseline (without PAs)	54.06 %
LNTRA Astrid *et al.* [4] - Patch based	57.09 %
LNTRA Astrid *et al.* [4] - Skip-frame based	55.48 %
Ours w/o $\mathcal{D}$	57.53 %
Ours w/ $\mathcal{D}$	**57.98 %**

$\omega_2^{(t)}$) is computed between the input test ϕ at time t and its reconstruction from $\mathcal{A}^t$. For measuring semantic inconsistency, the sequence of input frames is fed into $\mathcal{D}$ in a sliding window fashion (window size = 16). We compute the output probability of a frame at time t to be anomalous from its ViFi-CLIP feature representation and denote it by $\omega_3^{(t)}$. The aggregate anomaly score is given by the weighted average:

$$\omega_{agg}^{(t)} = \begin{cases} \eta_1\omega_1^{(t)} + \eta_2\omega_2^{(t)} + \eta_3\omega_3^{(t)}, & \text{w/ } \mathcal{D} \\ \eta_1\omega_1^{(t)} + \eta_2\omega_2^{(t)}, & \text{w/o } \mathcal{D}; (\eta_3 = 0) \end{cases} \tag{6}$$

where η_1, η_2, η_3 are tuned for every dataset. (Refer to section 3 in supp. material for further details)

4.3. Results

We performed extensive and exhaustive quantitative and qualitative assessments on four datasets namely Ped2 [42], Avenue [47], ShanghaiTech [53] and UBnormal [2].

Baselines: We compare our results with memory based AE [24, 59] and other reconstruction based method trained with pseudo-anomalous samples created using other simulation techniques [4, 5]. The network trained without any PAs is represented as the standard *baseline*. The model design of the AE is fixed across all the experimental settings. Object-level information is only considered for perturbing the normal data during training while at inference we evaluate results strictly based on reconstruction and classification outputs i.e. without any object detection. Hence our method is not directly comparable to object-centric methods.

1. Quantitative Assessment: In Table 1, we report micro AUC comparisons of overall scores of our model and existing SOTA methods on test sets of Ped2, Avenue and Shanghai datasets. We follow the same practice as in [4] of dividing the SOTA methods into 5 categories.

Our method is closest to reconstruction based methods though we also avail the discriminator $\mathcal{D}$ as the auxiliary component to learn the distance between normal data distribution and PAs distribution. For clarity, we provide results *with and without* $\mathcal{D}$ for all the datasets. Compared to memory-based networks, our unified framework trained on synthetically generated spatio-temporal PAs outperforms MemAE [24] and MNAD-Reconstruction [59] on Avenue and Shanghai while on Ped2 surpasses MNAD-Reconstruction and achieves comparable performance as MemAE. We also compare our results with other PAs generator methods such as STEAL Net [5] and LNTRA [4]. We observe that on the Avenue dataset our model outper-

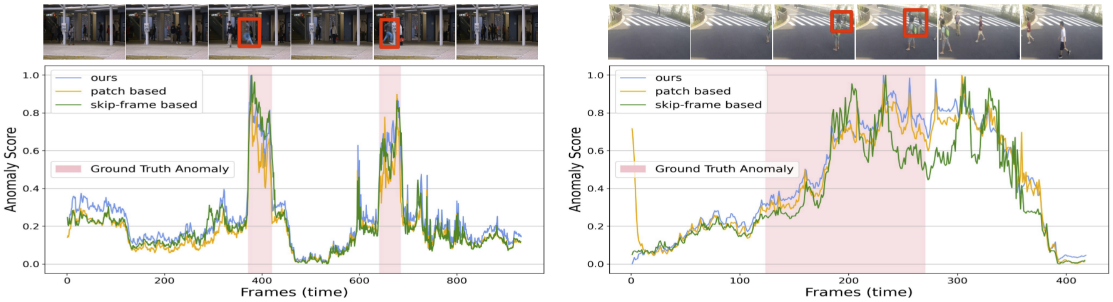

Figure 3. Qualitative Assessment : Visualisation of anomaly score over time for sample videos in Avenue (left) and ShanghaiTech (right). Compared with other PAs generator and reconstruction based methods in LNTRA [4] - patch and skip-frame based.

forms LNTRA (patch, skip-frame based) though marginally lags behind STEAL-Net whereas STEAL-Net and LNTRA achieve better performance than our model on Ped2 and Shanghai dataset. However such methods generate PAs under bold assumptions and inductive biases which may cause them to fail in particular cases. We report such cases in the Ablation study (Figure 4). In Table 3 we show that the transfer performance of our model is comparable with other PAs generation methods (see section 4.4). We do employ optical flow like other methods (Frame-Pred [44]) and observe that our results outperform Frame-Pred on the Avenue, achieve comparable performance on ShanghaiTech and are marginally less on Ped2.

In Table 2, we show a comparison between baseline, [4] and our approach on the validation set of the UBnormal dataset by training only on the normal videos in the train split. This is done to ensure consistency in evaluation under the OCC setting (refer to section 1 in supp for data-split details). The training and evaluation for baseline and LNTRA (patch, skip-frame) based methods on UBnormal was performed using scripts provided by the authors of LNTRA[4]. We observe that our method outperforms baseline and LNTRA achieving micro AUC score of 57.98% and implying that our PAs are generic and applicable for more diverse anomalous scenarios. Both in Table 1, 2 we notice that the effect of adding $\mathcal{D}$ is minimal, which validates the intuition that VAD cannot be directly addressed as a classification problem.

Table 1, 2 show that no single reconstruction-based method excels on all datasets. This is because anomalies are context-dependent. Different methods have inductive biases that work for specific datasets but not others. Our work provides a generic solution towards generating PAs without making bold assumptions about dataset's anomalies.

2. Qualitative Assessment: We conduct qualitative analysis of the anomaly score over time for sample videos in Avenue, Shanghai (Figure 3) and Ped2, UBnormal (Figure 3 in supp). We also compare our model's anomaly score over time with those obtained from LNTRA (skip-frame, patch-based). It can be concluded that on the Avenue and Ped2 datasets, our method detects anomalies fairly well and performance is equivalent with LNTRA models. Though there exist certain failure cases in the Shanghai and UBnormal datasets which occur due to anomalies occurring due to abnormal interaction between two objects i.e. fighting between two individuals in Shanghai and accident with a bike in UBnormal. Though our PAs generator is generic, end-to-end finetuning is further needed to emulate such complex real-world anomalies.

4.4. Ablation Studies

1: How transferable are PAs? We also examine how well PAs transfer across various VAD datasets. We use our pre-trained model on UBnormal dataset, which contains a wide range of anomalies and backgrounds, making it suitable for transferability. We tested the model on rest of the datasets without fine-tuning. Our results in Table 3 show that our model outperforms the patch-based method on all other datasets while achieves competitive performance compared to the skip-frame based method. This provides an interesting insight that our PAs are generic and transferable.

Table 3. Transfer Performance : micro-AUC scores.

Method	Ped2	Avenue	Shanghai
Patch [4]	78.80 %	43.94 %	61.57 %
Skip-Frame [4]	85.21 %	83.82 %	70.52 %
Ours w/ $\mathcal{D}$	85.37 %	83.50 %	70.07 %

2: How to interpret PAs? In Figure 4, we compare error heatmaps generated using a model trained with patch and skip-frame based PAs and with our spatial-PAs on all the respective datasets. Since skip-frame and patch based PAs carry strong assumptions, they tend to have problems detecting complicated real-world anomalies in ShanghaiTech such as a baby carriage (anomalous object) whereas our model trained with spatial-PAs yields high error for such cases. Furthermore, our PAs also give strong results on the synthetic dataset UBnormal, where patch and skip-frame based PAs fail to detect complex violent scenes as temporal irregularity induced through skip-frames is not generic. However, even

[4] https://github.com/aseuteurideu/LearningNotToReconstructAnomalies

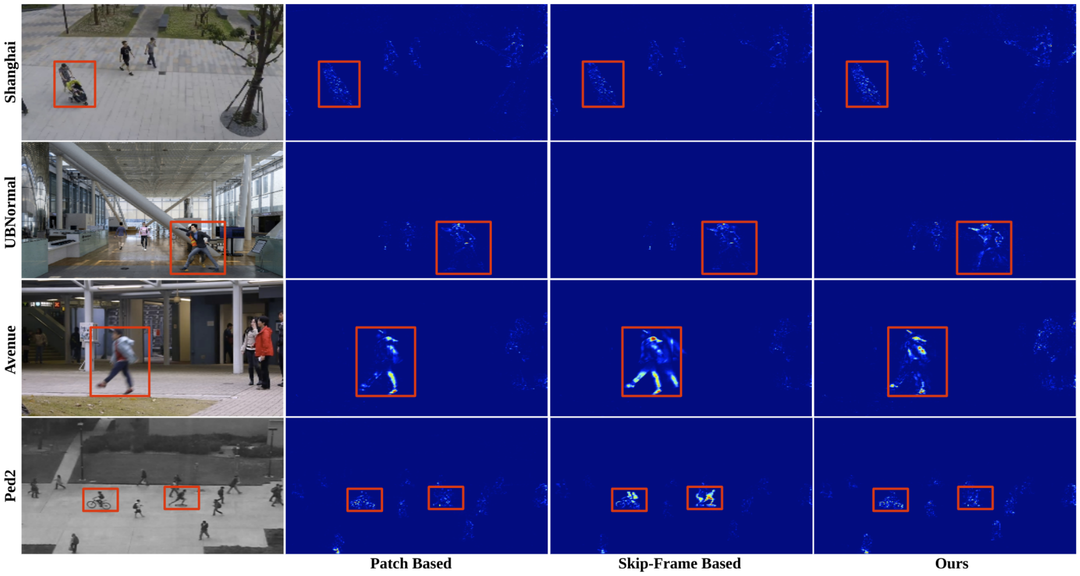

Figure 4. Visualisation of error heatmap for sample videos. Compared with other PAs generator methods in LNTRA [4].

Table 4. Effect of Random and Segmentation masks on micro-AUC scores, using the output of $\mathcal{A}^s$ when trained with $p_s = 0.4$.

Mask Type	Ped2	Avenue
Random Mask	91.18 %	83.13 %
Segmentation Mask	92.71 %	84.51 %

our spatial-PAs, which are not explicitly trained to detect temporal anomalies are able to determine such real-world anomalies. On Avenue and Ped2 datasets, our model yields comparable error to patch based PAs for an anomalous activity however we observe that skip-frame based PAs overly estimates the reconstruction error for the same. Intuitively this indicates that even though skip-frame performs reasonably well on benchmark datasets but it is susceptible to amplification of the error. An explanation for this phenomena could be due to underlying strong assumption of skipping frames based on a specific stride value to model temporal irregularity. These observations validate that our PAs are generalised and enable understanding of which real-world anomalies can be detected using which type of PAs.

3: Random vs Segmentation masks: Table 4 shows the effect of using random and segmentation masks for generating spatial PAs. We observe that using a segmentation mask gives better AUC score on Ped2 and Avenue dataset, which is intuitively justified as segmentation masks contain more semantic information. Despite this, our method is flexible in terms of type of mask chosen.

5. Conclusions and Discussion

In this paper we presented a novel and generic spatio-temporal PAs generator vital for VAD tasks without incorporating strong inductive biases. We achieve this by adding perturbation in the frames of normal videos by inpainting a masked out region using a pre-trained LDM and by distorting optical flow by applying mixup-like augmentation (Figure 2). We also introduced a simple unified VAD framework that learns three types of anomaly indicators i.e. reconstruction quality, temporal irregularity and semantic inconsistency in an OCC setting (Figure 1). Extensive evaluation shows that our framework though not objectively SOTA but achieves comparable performance to other SOTA reconstruction methods and PA generators with predefined assumptions across multiple datasets (Table 1, 2) without any end-to-end finetuning or any post-processing. This indicates the effectiveness, generalisation and transferability of our PAs.

However, there are limitations with this work. First, our model was not trained in an end-to-end fashion and doesn't avail more powerful architectures (vision transformers or 3D-ResNets) due to limited computational resources, which might boost the performance. It will also be interesting to make this setting adaptive by learning a policy network to select which anomaly indicator among poor reconstruction quality, temporal irregularity and semantic inconsistency contributes more towards detection of real-world anomalies. Second, the notion of generating latent space PAs for VAD through LDMs or manifold mixup remains to be investigated.

6. Acknowledgement

This research is supported by Science Foundation Ireland under the Grant Number SFI/12/RC/2289_P2, funded by the European Regional Development Fund and Tobii FotoNation. Dedicated to Dr. Kevin McGuinness, with lasting gratitude for his invaluable contributions and enduring influence.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2019. 6

[2] 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 (CVPR)*, pages 20143–20153, 2022. 6

[3] Abhishek Aich, Kuan-Chuan Peng, and Amit K Roy-Chowdhury. Cross-domain video anomaly detection without target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2579–2591, 2023. 3

[4] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. In *BMVC*, 2021. 1, 2, 3, 5, 6, 7, 8

[5] 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*, pages 207–214, 2021. 1, 3, 6

[6] Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger–Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, Xavier Glorot, Xavier Muller, Sylvain Pannetier Lebeuf, Razvan Pascanu, Salah Rifai, François Savard, and Guillaume Sicard. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172, Fort Lauderdale, FL, USA, 2011. PMLR. 3

[7] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 938–946. AAAI Press, 2021. 6

[8] Yunpeng Chang, Tu Zhigang, Xie Wei, and Yuan Junsong. Clustering driven deep autoencoder for video anomaly detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[9] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. *Advances in neural information processing systems*, 30, 2017. 3

[10] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision*, pages 334–349. Springer, 2016. 6

[11] Nikolaos Dionelis, Mehrdad Yaghoobi, and Sotirios A Tsaftaris. Boundary of distribution support generator (bdsg): Sample generation on the boundary. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 803–807. IEEE, 2020. 3

[12] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020. 5

[13] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020. 6

[14] Jinhao Dong and Tong Lin. Margingan: adversarial training in semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3

[15] 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. 6

[16] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 254–255, 2020. 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[18] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 3

[19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[20] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 1

[21] 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. 1, 3

[22] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4505–4523, 2021. 6

[23] 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 (ICCV)*, 2019. 1, 2, 3, 5

[24] 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 International Conference on Computer Vision*, pages 1705–1714, 2019. 6

[25] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[26] 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. 1, 2, 6

[27] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 4

[28] 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. 6

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 4

[31] 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. 1, 3, 6

[32] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using narrowed normality clusters. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1951–1960. IEEE, 2019. 6

[33] 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. 4

[34] John Taylor Jewell, Vahid Reza Khazaie, and Yalda Mohsenzadeh. One-class learned encoder-decoder network with adversarial context masking for novelty detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3591–3601, 2022. 6

[35] Xiangli Ji, Bairong Li, and Yuesheng Zhu. Tam-net: Temporal enhanced appearance-to-motion generative network for video anomaly detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 1, 3, 6

[36] 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. 5

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 3

[38] Jin-Ha Lee, Muhammad Zaigham Zaheer, Marcella Astrid, and Seung-Ik Lee. Smoothmix: A simple yet effective data augmentation to train robust classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1, 3

[39] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatio-temporal adversarial networks for abnormal event detection. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1323–1327. IEEE, 2018. 6

[40] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2019. 6

[41] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2020. 3

[42] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014. 6

[43] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. A causal inference look at unsupervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1620–1629, 2022. 3

[44] 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. 1, 3, 5, 6, 7

[45] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *BMVC*, page 71, 2018. 6

[46] 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. 1

[47] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 6

[48] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 6

[49] 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. 6

[50] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, 2017. 1, 2

[51] 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 (ICCV)*, 2017. 1, 2

[52] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017. 6

[53] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *ICCV, Oct*, 1(2):3, 2017. 6

[54] 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. 6

[55] Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees GM Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it until you make it: Towards accurate near-distribution novelty detection. In *NeurIPS ML Safety Workshop*. 3

[56] Phuc Cuong Ngo, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence gan: Towards better anomaly detection. In *2019 IEEE 31St International Conference on tools with artificial intelligence (ICTAI)*, pages 141–148. IEEE, 2019. 3

[57] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283, 2019. 6

[58] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020. 3

[59] 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. 1, 2, 3, 5, 6

[60] 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 (WACV)*, pages 2003–2012, 2021. 1, 3

[61] 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. 6

[62] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2598–2607, 2020. 6

[63] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2293–2312, 2020. 4

[64] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 2, 5

[65] 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. 6

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4

[67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[68] Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18717–18726, 2023. 3, 6

[69] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. Deep appearance features for abnormal behavior detection in video. In *International Conference on Image Analysis and Processing*, pages 779–789. Springer, 2017. 6

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[72] 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. 1, 3

[73] 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. 6

[74] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov,

Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 4

[75] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129: 123–130, 2020. 6

[76] Zhiqiang Tang, Yunhe Gao, Leonid Karlinsky, Prasanna Sattigeri, Rogerio Feris, and Dimitris Metaxas. Onlineaugment: Online data augmentation with less domain knowledge. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 313–329. Springer, 2020. 3

[77] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In *International Conference on Image Analysis and Processing*, pages 49–62. Springer, 2023. 3

[78] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Exploring diffusion models for unsupervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2540–2544, 2023. 3

[79] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022. 4

[80] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Phung. Robust anomaly detection in videos using multilevel representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5216–5223, 2019. 6

[81] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 494–511. Springer, 2022. 1, 3

[82] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2463–2471, 2020. 6

[83] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 35–45. Springer, 2022. 3

[84] Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems*, 31(7):2609–2622, 2019. 6

[85] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 3

[86] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[87] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. *Neurocomputing*, 143:144–152, 2014. 6

[88] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *BMVC*, 2015. 6

[89] 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. 6

[90] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5527–5537, 2023. 3

[91] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 6

[92] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14592–14601, 2023. 3, 6

[93] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813, 2019. 6

[94] 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. 6

[95] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, pages 214–223. Springer, 2007. 2, 4

[96] 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. 1, 3, 6

[97] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*, pages 358–376. Springer, 2020. 3

[98] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020. 3

[99] Muhammad Zaigham Zaheer, Jin-Ha Lee, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Stabilizing adversarially learned one-class novelty detection using pseudo anomalies. *IEEE Transactions on Image Processing*, 31:5963–5975, 2022. 3

[100] 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. 1, 3

[101] Muhammad Zaigham Zaheer, Jin-Ha Lee, Marcella Astrid, Arif Mahmood, and Seung-Ik Lee. Cleaning label noise with clusters for minimally supervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2020. 3

[102] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023. 3

[103] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2, 4

[104] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. *arXiv preprint arXiv:1912.11188*, 2019. 3

[105] Ying Zhang, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Shun Sakai. Video anomaly detection based on locality sensitive hashing filters. *Pattern Recognition*, 59:302–311, 2016. 6

[106] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1933–1941, New York, NY, USA, 2017. Association for Computing Machinery. 1, 2

[107] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 6

[108] Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards open set video anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 395–412. Springer, 2022. 3