

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

Nicolae-Cătălin Ristea^{1,2,◊}, Florinel-Alin Croitoru^{1,◊}, Radu Tudor Ionescu^{1,3,*}, Marius Popescu^{1,3}, Fahad Shahbaz Khan^{4,5}, Mubarak Shah⁶

¹University of Bucharest, Romania, ²NUST Politehnica Bucharest, Romania,

³SecurifAI, Romania, ⁴MBZ University of Artificial Intelligence, UAE,

⁵Linköping University, Sweden, ⁶University of Central Florida, US

Abstract

We propose an efficient abnormal event detection model based on a lightweight masked auto-encoder (AE) applied at the video frame level. The novelty of the proposed model is threefold. First, we introduce an approach to weight tokens based on motion gradients, thus shifting the focus from the static background scene to the foreground objects. Second, we integrate a teacher decoder and a student decoder into our architecture, leveraging the discrepancy between the outputs given by the two decoders to improve anomaly detection. Third, we generate synthetic abnormal events to augment the training videos, and task the masked AE model to jointly reconstruct the original frames (without anomalies) and the corresponding pixel-level anomaly maps. Our design leads to an efficient and effective model, as demonstrated by the extensive experiments carried out on four benchmarks: Avenue, ShanghaiTech, UBnormal and UCSD Ped2. The empirical results show that our model achieves an excellent trade-off between speed and accuracy, obtaining competitive AUC scores, while processing 1655 FPS. Hence, our model is between 8 and 70 times faster than competing methods. We also conduct an ablation study to justify our design. Our code is freely available at: <https://github.com/ristea/aed-mae>.

1. Introduction

In recent years, research on abnormal event detection in video gained significant traction [1, 10, 17, 18, 26–28, 36, 38, 43, 44, 49, 52, 57, 58, 61, 62, 65, 69, 76, 78, 80, 83, 87, 90, 95, 97–100], due to its utter importance in video surveillance. Despite the growing interest, video anomaly detection remains a complex task, owing its complexity to the fact that abnormal situations are context-dependent and do not occur very often. This makes it very difficult to collect a representative set of abnormal events for training state-of-the-art deep learning models in a fully supervised manner. To showcase the rarity and reliance on context of anomalies, we refer to the vehicle ramming attacks carried out by terrorists against pedestrians. As soon as a car is steered

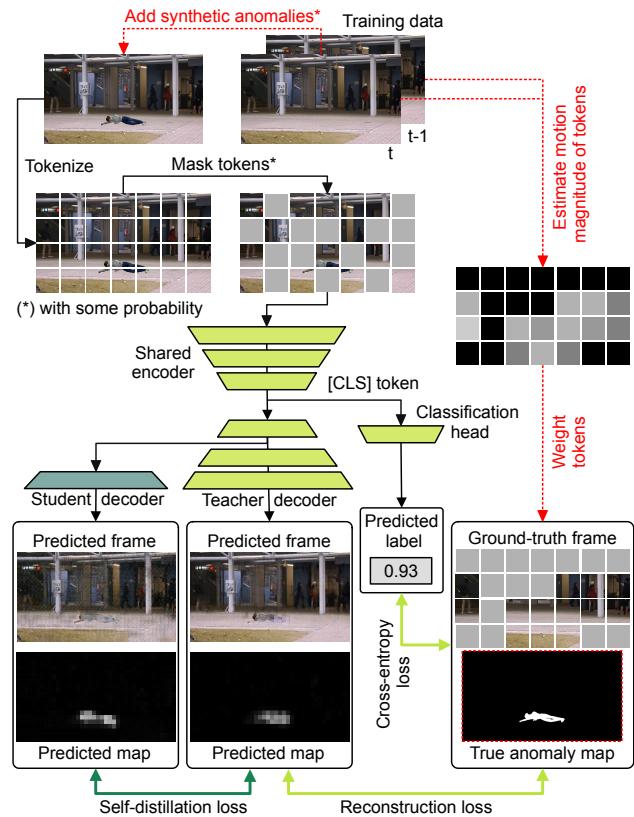


Figure 1. Our masked auto-encoder for abnormal event detection based on self-distillation. At training time, some video frames are augmented with synthetic anomalies. The teacher decoder learns to reconstruct original frames (without anomalies) and predict anomaly maps. The student decoder learns to reproduce the teacher’s output. Motion gradients are aggregated at the token level and used as weights for the reconstruction loss. Red dashed lines represent steps executed only during training.

on the sidewalk, it becomes an abnormal event. Hence, the place where the car is driven (street versus sidewalk) determines the normal or abnormal label of the action, *i.e.* the label depends on context. Furthermore, there are less than 200 vehicle ramming attacks registered to date¹, confirming

¹https://en.wikipedia.org/wiki/Vehicle-ramming_attack

*corresp. author: raducu.ionescu@gmail.com; [◊]equal contribution

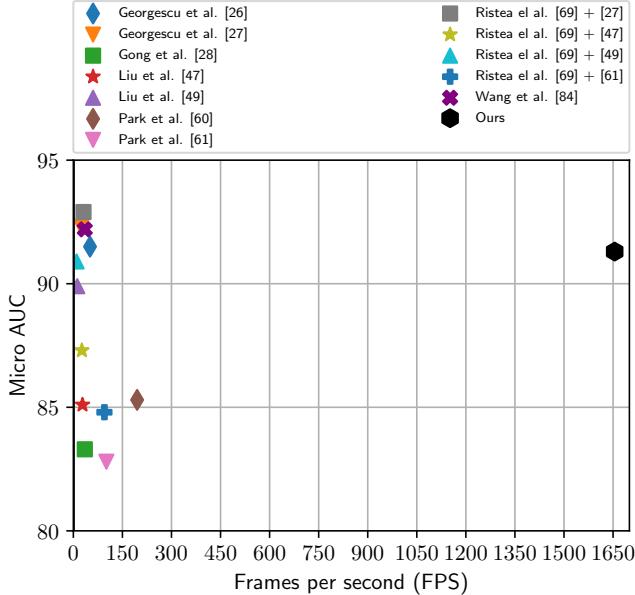


Figure 2. Performance versus speed trade-offs for our self-distilled masked AE and several state-of-the-art methods [26–28, 47, 49, 60, 61, 69, 84] (with open-sourced code), on the Avenue data set. The running times of all methods are measured on a computer with one Nvidia GeForce GTX 3090 GPU with 24 GB of VRAM. Best viewed in color.

the scarcity of such events (even less are caught on video).

Since training anomaly detectors under a fully supervised setting is not possible, most studies dealing with abnormal event detection took a distinct path, proposing variations of outlier detection methods [3, 13, 14, 17, 21, 22, 29, 37, 41, 43, 45, 47, 51, 53, 55, 56, 61–68, 71, 80, 90, 92, 102–105]. Such methods treat abnormal event detection as an outlier detection task, where a normality model trained on normal events is applied on both normal and abnormal events during inference, labeling events deviating from the learned model as abnormal. Different from the mainstream path based on outlier detection, we propose an approach to augment each training video scene with synthetic anomalies, by randomly superimposing temporal action segments from the synthetic UBnormal data set [1] on our real-world data sets. We thus introduce synthetic anomalies at training time, enabling our model to learn in an open-set supervised manner. Additionally, we force our model to reconstruct the original training frames (without anomalies) to limit its ability to reconstruct anomalies, hence generating higher errors when anomalies occur.

A large body of work on video anomaly detection has focused on employing auto-encoders (AEs) to address the task [5, 22, 27–29, 36, 49, 81, 85], relying on the poor reconstruction capabilities of these models on out-of-distribution data. Since training is carried out only on normal examples, it is expected for AEs to exhibit high reconstruction errors when anomalies occur. However, several researchers

observed that AEs generalize too well [5, 36], being able to reconstruct anomalies with very high precision. Thus, to better leverage the reconstruction error of AEs in anomaly detection, researchers explored a few alternatives, from the use of dummy [36] or pseudo-anomalies [5, 27] to the integration of memory modules [28, 49, 61]. With the same purpose in mind, we propose to employ masked auto-encoders [30] in anomaly detection, introducing new ways to regulate their generalization capacity. Indeed, we go beyond employing the standard masked AE framework, and propose three novel changes to enhance the anomaly detection performance of our model. First, we propose to weight tokens based on the magnitude of motion gradients, raising the importance of tokens with higher motion in the reconstruction loss. This makes our model focus on reconstructing tokens with high motion, and avoid reconstructing the background scene, which is typically static for surveillance cameras. Second, we attach a classification head to discriminate between normal and pseudo-abnormal instances in the latent encoding space. Third, we integrate a teacher decoder and a student decoder into our masked AE architecture, where the student decoder learns to distill knowledge from the already optimized teacher. To reduce our processing time, we use a shared encoder for the teacher and student models, leading to a process known as self-distillation [101]. During the self-distillation process, the shared encoder is frozen. We leverage the discrepancy between the outputs given by the teacher and student decoders along with the reconstruction error of the teacher to boost anomaly detection performance. Our entire framework, integrating these components into a meticulous design, is shown in Figure 1.

State-of-the-art deep anomaly detectors [10, 26, 27, 84] typically rely on a costly object detection method to increase precision, limiting the processing bandwidth to one video stream per GPU, at around 20–30 FPS. However, for real-world video surveillance, *e.g.* monitoring an entire city with hundreds or thousands of cameras, the processing costs of object-centric video anomaly detectors are simply too high, given their power consumption and that one GPU can cost around \$2,000. To this end, we turn our attention to developing a lightweight model (6 transformer blocks, 3M parameters), capable of processing around 66 video streams at 25 FPS, significantly reducing the processing costs. Different from competing models performing anomaly detection at the object [10, 18, 26, 27, 36, 84, 95] or spatio-temporal cube [15, 21, 35, 37, 41, 48, 51, 53, 55, 60, 71, 75, 103] levels, we present a model that takes whole video frames as input, which is significantly more efficient (see Figure 2).

We carry out comprehensive experiments on four benchmarks: Avenue [51], ShanghaiTech [53], UBnormal [1] and UCSD Ped2 [55]. The empirical results show that our method is 8 to 70 times faster than competing methods [26–28, 47, 49, 60, 61, 69, 84], while achieving comparable ac-

curacy levels. Aside from the main results, we conduct an ablation study showing that our novel design choices are supported by empirical evidence.

In summary, our contribution is threefold:

- We propose a lightweight masked auto-encoder for anomaly detection in video, which learns to reconstruct tokens with higher motion magnitude.
- We introduce a self-distillation training pipeline, leveraging the discrepancy between teacher and student decoders to obtain a significant accuracy boost for our highly efficient model (due to the shared encoder).
- To further boost the performance of our model, we introduce a data augmentation approach based on superimposing synthetic anomalies on normal training videos, which enables the masked AE model to learn with open-set supervision.

2. Related Work

Anomaly detection in video is typically formulated as a one-class learning problem, where only normal data is available at training time. During test time, both normal and abnormal examples are present [59, 65]. There are several categories of anomaly detection approaches, including dictionary learning methods [13, 14, 21, 51, 68, 89], probabilistic models [2, 3, 25, 31, 41, 55, 56, 73, 91], change detection frameworks [15, 35, 48, 58], distance-based models [36, 37, 62, 63, 67, 71, 72, 75, 76, 79, 81] and reconstruction-based approaches [23, 28, 29, 46, 47, 53, 57, 61, 66, 69, 80]. Considering that reconstruction-based methods often reach state-of-the-art performance in anomaly detection [27, 69], a large body of works used the reconstruction-based paradigm in the past few years. To this end, we adopt this paradigm in our study.

With respect to the level at which the anomaly detection is carried out, methods can be categorized into spatio-temporal cube-level methods [37, 47, 48, 51, 55, 56, 66, 67, 71, 75, 76, 96, 103], frame-level methods [47, 66, 67], and object-level methods [10, 18, 19, 26, 27, 36, 49, 84, 95].

Frame-level and cube-level methods. Before the deep learning era, preliminary abnormal event detection models commonly relied on taking short video sequences and dividing them into spatio-temporal cuboids [15, 21, 41, 51, 55, 75, 103]. The cubes are then considered as independent examples, being passed as input to a machine learning model. This mainstream practice continued during the deep learning period [29, 35, 37, 43, 44, 53, 60, 62, 63, 71, 76, 100], when deep networks have been used to extract features [35, 37, 48, 53] or learn [28, 29, 43, 44, 53, 60, 62–64, 71, 89, 100] from these spatio-temporal cubes.

At the same time, some studies considered using entire video frames as input [47, 66, 67]. For example, Liu *et al.* [47] proposed an effective algorithm, which learns to reconstruct the next frame of a short video sequence. A

more complex approach is proposed by Ravanbakhsh *et al.* [67], who employ optical-flow reconstruction to predict the anomalous regions from an input image. In a different study, Ravanbakhsh *et al.* [66] proposed to detect anomalies at the frame level via generative adversarial networks.

Frame-level and cube-level methods have a common characteristic, namely their relatively high processing speed due to the reasonably fast preprocessing steps, as opposed to object-centric methods. Still, frame-level methods hold a stronger advantage in terms of time, since cube-level methods need to process each cube as an independent example. Indeed, it is more efficient to process a mini-batch of frames rather than several mini-batches of spatio-temporal cubes. However, cube-level methods often outperform frame-level methods. To this end, we propose a masked AE that takes whole frames as input, yet learns interactions between video patches, thus integrating the best of both worlds.

To boost the performance of frame-level or cube-level methods, researchers explored the inclusion of various components, such as memory modules [28, 61] or masked convolutional blocks [69]. Although integrating additional modules into the framework leads to accuracy gains, the procedure often comes with efficiency drawbacks. In contrast, our goal is to achieve a superior trade-off between performance and speed, with a higher focus on efficiency. As such, we design a lightweight masked AE based on convolutional vision transformer (CvT) blocks [88], and propose several upgrades resulting in a minimal time overhead. For example, we employ knowledge distillation to leverage the discrepancy between the teacher and the student models. However, to keep the processing time to the bare minimum, we resort to self-distillation [101] and use a shared encoder for the teacher and student networks.

Object-level methods. To reduce the number of false positive detections often observed for other methods, some recent studies [10, 26, 27, 36, 49, 84, 95] proposed to look for anomalous objects rather than anomalous frames or cubes. Object-centric methods use the prior information from an object detector, enabling the anomaly detector to focus only on objects. This kind of framework boosts the accuracy by significant margins, currently reaching state-of-the-art performance [10, 84]. However, a considerable drawback is that the inference speed of the whole framework is directly conditioned by the object detector’s speed, which is often much lower than that of the anomaly detection network [26, 36]. Hence, the processing time is significantly limited. In contrast, we perform anomaly detection at the frame level, obtaining an inference speed that is between 32 to 70 times faster than object-centric models [10, 26, 27, 36, 49, 84, 95].

Masked auto-encoders in anomaly detection. He *et al.* [30] proposed masked auto-encoders as a pretraining method to obtain strong backbones for downstream tasks.

Since then, the method has been adopted in various fields, e.g. video processing [24] or multimodal learning [7], with remarkable results. We elaborate the connection to seemingly related masked AEs in the supplementary [11, 94]. The masking framework has also been used for anomaly detection in medical [34] and industrial [39] images. To the best of our knowledge, we are the first to propose a masked transformer-based auto-encoder for video anomaly detection. Moreover, we go beyond applying standard masked AEs, proposing several modifications leading to superior performance levels: emphasizing tokens with higher motion, augmenting training videos with synthetic anomalies, and employing self-distillation.

Knowledge distillation in anomaly detection. Knowledge distillation [6, 32] was originally designed to compress one or multiple large models (teachers) into a lighter neural network (student). Recently adopted in anomaly detection [8, 12, 16, 26, 74, 86], knowledge distillation was deemed useful due to the possibility of leveraging the representation discrepancy between the teacher and the student networks, which is larger in the case of anomalies. For example, Bergmann *et al.* [8] trained an ensemble of student networks on normal data to reproduce the output of a deep feature extractor, which is pretrained on ImageNet [70]. The authors use the difference between the teacher label and the mean over student labels to detect abnormal pixels. Salehi *et al.* [74] employed a more thorough distillation process, called hint learning, in which the multi-level features of a teacher pretrained on ImageNet are distilled into a clone.

Most studies based on knowledge distillation applied the framework to image anomaly detection [8, 12, 16, 74]. With few exceptions [26, 86], knowledge distillation in video anomaly detection remains largely unexplored. Wang et al. [86] employed the teacher-student training paradigm to learn from unlabeled video samples in a self-supervised manner. Georgescu *et al.* [26] integrated knowledge distillation as a proxy task into a multi-task learning framework for video anomaly detection.

Distinct from the aforementioned studies, to our knowledge, we are the first to introduce a variant of self-distillation in anomaly detection. Self-distillation [101] attaches multiple classification heads at various depths to boost the classification performance of a neural classifier. In contrast, we integrate self-distillation into a masked AE, employing two decoders of different depths. Due to the shared encoder, we are able to leverage the reconstruction discrepancy between the teacher and the student with a minimal computational overhead.

3. Method

Overview. We introduce a lightweight teacher-student transformer-based masked AE, which employs a two-stage training pipeline. In the first stage, we optimize a teacher

masked AE via a reconstruction loss that employs a novel weighting mechanism based on motion gradients. In the second stage, we optimize the last (and only) decoder block of a student masked AE, which shares most of the backbone (kept frozen) with its teacher, to preserve efficiency. Next, we describe how to create training videos with synthetic anomalies and train the masked AEs to jointly predict the anomaly maps and overlook (not reconstruct) the anomalies from training frames. Lastly, we introduce a classification head to distinguish between frames with and without synthetic anomalies, which further boosts the performance of our method, with a marginal computational overhead.

Architecture. Our masked AE pursues the architectural principles proposed in [30]. Hence, the entire architecture is formed of visual transformer blocks. In contrast to He *et al.* [30], we replace the ViT [20] blocks with CvT blocks [88], aiming for higher efficiency. Our processing starts by dividing the input images into non-overlapping tokens and removing a certain number of tokens. The encoder embeds the remaining tokens via convolutional projection layers, and the result is processed by transformer blocks. The decoder operates on a complete set of tokens, those removed being replaced with mask tokens. Its architecture is symmetric to that of the encoder. For efficiency reasons, we only use three blocks for the encoder and three blocks for the decoder. Each block is equipped with four attention heads. To achieve further speed gains, we replace all dense layers inside the CvT blocks with pointwise convolutions. We consider the architecture described so far as a teacher network. A student decoder branches out from the teacher after the first transformer block of the main decoder, adding only one extra transformer block (as shown in Figure 1).

Motion gradient weighting. Masked AEs [30] have been originally applied on natural images. In this context, reconstructing randomly masked tokens is a viable solution, since images have high foreground and background variations. However, abnormal event detection data sets [2, 51, 53, 55] contain videos from fixed cameras with static backgrounds [65]. Learning to reconstruct the static background via masked AEs is both trivial and useless. Hence, naively training masked AEs to reconstruct randomly masked tokens in video anomaly detection is suboptimal. To this end, we propose to take into account the magnitude of the motion gradients when computing the reconstruction loss.

Let $\mathbf{x}_t \in \mathbb{R}^{h \times w \times c}$ be the video frame at index t . Let n be the number of non-overlapping visual tokens (patches) of size $d \times d \times c$ from each frame \mathbf{x}_t , where c is the number of input channels, and d is a hyperparameter that directly determines n . Let $\left\{\mathbf{p}_i^{(t)}\right\}_{i=1}^n \in \mathbb{R}^{d \times d \times c}$ denote the set of tokens in frame \mathbf{x}_t , and $\left\{\hat{\mathbf{p}}_i^{(t)}\right\}_{i=1}^n \in \mathbb{R}^{d \times d \times c}$ the corresponding set of reconstructed tokens.

Following Ionescu *et al.* [36], we estimate the motion

gradient map \mathbf{g}_t of frame \mathbf{x}_t by computing the absolute difference between consecutive frames, which are previously filtered with a 3×3 median filter. Next, we divide the gradient magnitude map \mathbf{g}_t into non-overlapping patches, obtaining the set of gradient patches $\{\mathbf{r}_i^{(t)}\}_{i=1}^n \in \mathbb{R}^{d \times d \times c}$. Inside each gradient patch, we compute the maximum gradient magnitude per channel. Then, we compute the channel-wise mean over the maximum gradient magnitudes, as follows:

$$m_i^{(t)} = \frac{1}{c} \sum_{l=1}^c \max_{j,k} \{\mathbf{r}_{ijkl}^{(t)}\}, \forall j, k \in \{1, \dots, d\}. \quad (1)$$

Finally, we compute the token-wise weights for the reconstruction loss as follows:

$$w_i^{(t)} = \frac{m_i^{(t)}}{\sum_{j=1}^n m_j^{(t)}}, \forall i \in \{1, \dots, n\}. \quad (2)$$

Introducing the resulting weights $w_i^{(t)}$ into the conventional token-level reconstruction loss leads to an objective that pushes the masked AE to focus on reconstructing the patches with high motion magnitude. Formally, our weighted mean squared error loss is given by:

$$\mathcal{L}_{\text{wMSE}}(\mathbf{x}_t, \theta_T) = \frac{1}{n} \sum_{i=1}^n w_i^{(t)} \cdot \|\mathbf{p}_i^{(t)} - \hat{\mathbf{p}}_i^{(t)}\|_2^2, \quad (3)$$

where θ_T are the weights of our teacher masked AE. Although our reconstruction loss focuses on tokens with high motion, the masked tokens are still chosen randomly.

Self-distillation. Knowledge distillation has already shown its utility in anomaly detection [8, 12, 16, 26, 74]. Intuitively, since the teacher and student models are both trained on normal data, their reconstructions should be very similar for normal test samples. However, their behavior is not guaranteed to be similar on abnormal examples. Therefore, the magnitude of the teacher-student output gap (discrepancy) can serve as a means to quantify the anomaly level of a given sample. Unfortunately, this approach implies using both teacher and student models during inference, virtually splitting our processing speed in half. To slash the additional burden of using another model during inference, we propose to employ a novel variant of self-distillation with a shared encoder and two decoders, a teacher and a student. More precisely, the student branches out from the original architecture after the first transformer block of the teacher decoder, essentially adding only one transformer block.

Our training process is carried out in two stages. In the first phase, the teacher is trained with the loss defined in Eq. (3). In the second phase, we freeze the weights of the shared backbone and train only the student decoder via self-distillation. The self-distillation loss is similar to the one defined in Eq. (3). The main difference is that instead of reconstructing the patches from the real image, the student learns to reconstruct the ones produced by the teacher. Let



Figure 3. Four synthetic anomalies (with red contours) taken from the UBnormal data set [1] and overlaid on training frames from Avenue [51]. Best viewed in color.

$\{\tilde{\mathbf{p}}_i^{(t)}\}_{i=1}^n \in \mathbb{R}^{d \times d \times c}$ denote the patches reconstructed by the student. Then, the self-distillation loss can be expressed as follows:

$$\mathcal{L}_{\text{SD}}(\hat{\mathbf{x}}_t, \theta_S) = \frac{1}{n} \sum_{i=1}^n w_i^{(t)} \cdot \|\hat{\mathbf{p}}_i^{(t)} - \tilde{\mathbf{p}}_i^{(t)}\|_2^2, \quad (4)$$

where $\hat{\mathbf{x}}_t$ is the frame reconstructed by the teacher, and θ_S are the weights of the student decoder. Notice that we keep the motion gradient weights $w_i^{(t)}$ during self-distillation.

Synthetic anomalies. As observed in other studies [5, 36], AEs tend to generalize too well to out-of-distribution data. This behavior is not desired in anomaly detection, since methods based on AEs rely on having high reconstruction errors for abnormal examples and low reconstruction errors for normal ones. To this end, we propose to augment the training videos with abnormal events. Since collecting abnormal training examples from the real-world is not possible, we resort to adding synthetic (virtual) anomalies. We leverage the recently introduced UBnormal data set [1] and its accurate pixel-level annotations to crop out abnormal events and blend them in our training videos, while ensuring the temporal consistency of the added events. The resulting examples, some depicted in Figure 3, are used to augment the training set with extra data.

The synthetic examples help our model in three ways. First, in the reconstruction loss, we consider the original training frames (without superimposed anomalies) as the ground-truth, essentially forcing our model to overlook the anomalies. Formally, in Eq. (3), we use the patches $\{\mathbf{p}_i^{(t)}\}_{i=1}^n$ from the normal version of frame \mathbf{x}_t . Second, we add the anomaly map as an additional channel to our target image. In the anomaly map, we set normal pixels to 0 and abnormal pixels to 1. This change implies that, in Eq. (3) and Eq. (4), all patches will have an additional channel. Third, we use the ground-truth anomaly map to enhance the weights defined in Eq. (2). The added synthetic anomalies do not necessarily yield motion gradients with high magnitude. Hence, it is possible to have low weights in Eq. (3) and Eq. (4) for patches that correspond to anomaly regions. This is not desirable if we want the model to detect anom-

lies. To this end, we propose to add the anomaly maps and the gradients together, before computing the weights as in Eq. (2). Formally, in Eq. (1), we replace $\left\{ \mathbf{r}_i^{(t)} \right\}_{i=1}^n$ with $\left\{ \mathbf{r}_i^{(t)} + \mathbf{a}_i^{(t)} \right\}_{i=1}^n$, where $\left\{ \mathbf{a}_i^{(t)} \right\}_{i=1}^n$ is the set of the patches extracted from the anomaly map.

Classification head. We further harness the synthetic anomalies to train a classification head applied on the final [CLS] token of the shared encoder. The head is trained to discriminate between frames with and without synthetic anomalies. This head is trained using binary cross-entropy:

$$\mathcal{L}_{CE}(\hat{\mathbf{x}}_t, \theta_E) = -y_t \cdot \log(\hat{y}_t) - (1-y_t) \cdot \log(1 - \hat{y}_t), \quad (5)$$

where $y_t \in \{0, 1\}$ is 1 if the frame contains an anomaly and 0 otherwise, \hat{y}_t is the prediction, and θ_E represents the set of weights of the shared encoder.

Inference. During inference, we pass each frame \mathbf{x}_t through both teacher and student models to obtain the reconstructed frames $\hat{\mathbf{x}}_t$ and $\tilde{\mathbf{x}}_t$, respectively. Then, we compute output pixel-level anomaly map as:

$$\mathbf{o}_t = \alpha \cdot \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2 + \beta \cdot \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2 + \gamma \cdot \hat{y}_t, \quad (6)$$

where α , β and γ are hyperparameters that control the importance of the individual anomaly score components. Following [15, 35], we apply spatio-temporal 3D filtering to smooth the anomaly volumes. To obtain the frame-level anomaly scores, we keep the maximum value from each output map \mathbf{o}_t and subsequently apply another temporal Gaussian filter to smooth the values.

4. Experiments

4.1. Experimental Setup

Data sets. We verify the performance of our method on four data sets for video anomaly detection: Avenue [51], ShanghaiTech [53], UBnormal [1] and UCSD Ped2 [55]. ShanghaiTech is the largest data set, with 270K frames for training and about 50K for testing. UBnormal is the second largest, with about 116K training frames and 93K testing frames. Avenue is a popular benchmark containing 15K frames for training and another 15K for testing. UCSD Ped2 holds a total of 4.5K frames, out of which 2.5K are used for training. UBnormal is a benchmark that uses an open set evaluation, where training and test anomalies belong to disjoint category sets. For the other three data sets, the training videos contain only normal events, and the test ones include both normal and abnormal scenarios. To augment the normal training videos, we sample abnormal events from the UBnormal data set [1]. UBnormal is a synthetic (virtual) data set containing anomalies simulated by video game characters, which alleviates the burden of collecting anomalies from the real world. The probability of augmenting a frame from Avenue, ShanghaiTech or UCSD Ped2 is 0.25.

Evaluation. We evaluate all models following recent related works [1, 27, 69], considering both micro and macro

Type	Method	AUC								FPS	
		Avenue		Shanghai		UBnormal		Ped2			
		Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro		
Object-centric	[10]	91.6	92.5	83.8	90.5	62.1	86.5	-	-	20	
	[18]	86.4	-	71.6	-	-	-	97.8	-	-	
	[26]	91.5	92.8	82.4	90.2	55.4	84.5	97.5	99.8	51	
	[27]	92.3	90.4	82.7	89.3	61.3	85.6	98.7	99.7	24	
	[33]	-	-	85.9	-	71.8	-	-	-	-	
	[36]	87.4	90.4	78.7	84.9	-	-	94.3	97.8	-	
	[40]	80.2	-	73.7	-	50.7	-	-	-	-	
	[49]	89.9	93.5	74.2	83.2	-	-	99.3	-	12	
	[50]	91.8	92.3	83.8	87.8	-	-	-	-	-	
	[54] + [10]	91.6	92.4	83.6	90.6	-	-	-	-	20	
	[54] + [27]	93.2	91.8	83.3	89.3	-	-	-	-	31	
	[54] + [49]	89.5	93.6	75.2	83.8	-	-	-	-	10	
	[69] + [27]	92.9	91.9	83.6	89.5	-	-	-	-	31	
	[69] + [49]	90.9	92.2	75.5	83.7	-	-	-	-	10	
	[84]	92.2	-	84.3	-	-	-	99.0	-	35	
	[95]	89.6	-	74.8	-	-	-	97.3	-	-	
Frame or cube level	[4]	87.1	-	75.9	-	-	-	96.5	-	-	
	[5]	84.7	-	73.7	-	-	-	98.4	-	-	
	[9]	-	-	-	-	68.5	80.3	-	-	37	
	[28]	83.3	-	71.2	-	-	-	94.1	-	35	
	[37]	88.9	-	-	-	-	-	-	-	-	
	[43]	90.0	-	-	-	-	-	96.6	-	-	
	[47]	85.1	81.7	72.8	80.6	-	-	95.4	-	28	
	[48]	84.4	-	-	-	-	-	87.5	-	-	
	[54] + [47]	89.1	84.8	74.6	83.3	-	-	-	-	26	
	[54] + [61]	86.4	86.3	70.6	80.3	-	-	-	-	94	
	[57]	86.9	-	-	-	-	-	96.2	-	-	
	[60]	85.3	-	72.2	-	-	-	96.3	-	195	
	[61]	82.8	86.8	68.3	79.7	-	-	97.0	-	101	
	[62]	72.0	-	-	-	-	-	88.3	-	-	
	[63]	87.2	-	-	-	-	-	93.0	-	-	
	[66]	-	-	-	-	-	-	93.5	-	-	
	[67]	-	-	-	-	-	-	88.4	-	-	
	[69] + [47]	87.3	84.5	74.5	82.9	-	-	-	-	26	
	[69] + [61]	84.8	88.6	69.8	80.2	-	-	-	-	95	
	[76]	84.6	-	-	-	-	-	-	-	-	
	[77]	-	-	-	76.5	50.3	76.8	-	-	56	
	[78]	89.6	-	74.7	-	-	-	-	-	-	
	[80]	85.1	-	73.0	-	-	-	96.3	-	-	
	[82]	-	-	76.1	-	-	-	-	-	-	
	[87]	87.0	-	79.3	-	-	-	-	-	-	
	[89]	-	-	80.4	-	-	-	-	-	-	
	[90]	86.6	-	-	-	-	-	96.9	-	-	
	[93]	90.1	-	78.6	-	62.7	-	-	-	-	
	[96]	90.2	-	-	-	-	-	97.3	-	-	
	[100]	-	-	78.9	-	-	-	-	-	-	
	[103]	-	-	-	-	-	-	91.0	-	-	
Ours		91.3	90.9	79.1	84.7	58.5	81.4	95.4	98.4	1655	

Table 1. Micro and macro AUC scores (in %) of several state-of-the-art frame-level, cube-level and object-level methods versus our self-distilled masked AE on Avenue, ShanghaiTech, UBnormal and UCSD Ped2. The top three scores for each category of methods are shown in red, green, and blue. All reported running times (including those of the baselines) are measured on a machine with an Nvidia GeForce GTX 3090 GPU with 24 GB of VRAM.

AUC metrics. The area under the ROC curve (AUC) expresses the overlap between the ground-truth frame-level annotations and the anomaly scores predicted by a model, at multiple thresholds. At a given threshold, a frame is labeled as abnormal if the predicted anomaly score is above the threshold. For the micro AUC, the test frames from all

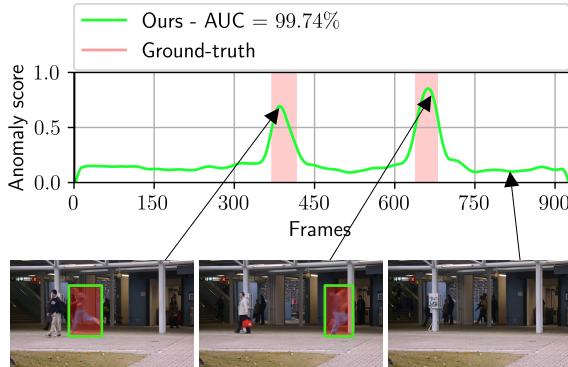


Figure 4. Predictions for test video 04 from Avenue. The abnormal bounding boxes are given by the convex hull of the patches labeled as abnormal. Best viewed in color.

videos are concatenated before computing the AUC over all frames. For the macro AUC, the AUC of each test video is first computed, and the resulting AUC scores are averaged to obtain a single value.

Hyperparameters. The encoder module is formed of three CvT blocks, each with a projection size of 256 and four attention heads. The teacher decoder contains three CvT blocks, while the student decoder contains only one block. All decoder blocks have four attention heads and a projection dimension of 128. Since the data sets have different input resolutions and objects vary in size, we adapt the patch size to each data set. Thus, we set the patch size to 16×16 on Avenue, 8×8 on ShanghaiTech and UBnormal, and 4×4 on UCSD Ped2. Regardless of the data set, the teacher network is trained for 100 epochs, while the student is trained for 40 epochs. We optimize the networks with Adam [42], using a learning rate of 10^{-4} and mini-batches of 100 samples. The hyperparameters in Eq. (6) are set to $\alpha = 0.4$, $\beta = 0.3$ and $\gamma = 0.3$, for all data sets.

4.2. Results

We present our results in Table 1 and discuss them below.

Results on Avenue. Our method obtains a micro AUC score of 91.3% on Avenue, being only 1.9% below the state-of-the-art object-centric method. Remarkably, in the category of frame-level methods, we reach the best micro and macro AUC scores. Taking into account that our method is much faster than all other methods, we consider that its performance is remarkable. In Figure 4, we illustrate the anomaly scores for test video 04 from Avenue. Here, our model is close to perfect, highlighting its ability to capture anomalies, such as people running.

Results on ShanghaiTech. On ShanghaiTech, our method reaches the top macro AUC score and the third-best micro AUC score, when compared with other frame-level frameworks. Object-centric methods generally surpass frame-level methods, but the former methods have much lower processing speeds (see Figure 2).

Motion weights	Self-distillation	Synthetic data	Anomaly maps	Classif. head	Avenue		Shanghai	
					AUC		AUC	
					Micro	Macro	Micro	Macro
					84.0	85.6	69.7	80.1
✓					84.8	86.3	71.3	80.9
✓	✓				88.5	86.0	76.3	83.8
✓	✓	✓			88.5	86.9	77.0	83.0
✓	✓	✓	✓		90.5	89.6	77.3	84.2
✓	✓	✓	✓	✓	91.3	90.9	79.1	84.7

Table 2. Impact of each novel component on the micro and macro AUC scores (in %), on Avenue [51] and ShanghaiTech [53].

Strategy	Avenue		Shanghai	
	AUC		AUC	
	Micro	Macro	Micro	Macro
Teacher	84.0	85.6	74.8	82.3
Teacher + Student	85.4	85.8	75.1	82.1
Teacher + Teacher-Student Difference	88.5	86.0	76.3	83.8
Teacher + Student + Teacher-Student Difference	86.9	85.8	75.8	83.6

Table 3. Impact of strategies to combine the outputs of the teacher and student models on Avenue [51] and ShanghaiTech [53]. These results do not include the synthetic anomalies and the classification head.

Data set	Measure	Percentage of synthetic data			
		0%	25%	50%	75%
Avenue	Micro AUC	88.5	91.3	90.6	89.9
	Macro AUC	86.0	90.9	89.4	87.7
Shanghai	Micro AUC	76.3	79.1	77.9	77.7
	Macro AUC	83.8	84.7	84.4	84.1

Table 4. Impact of varying the proportion of synthetic anomalies on the Avenue [51] and ShanghaiTech [53] data sets.

Results on UBnormal. In terms of the micro AUC, the best frame-level method on UBnormal is TimeSformer [9], which benefits from large-scale pretraining. Notably, our method obtains a higher macro AUC than TimeSformer. The micro AUC of our method is fairly close to the micro AUC levels of the better object-centric approaches. In terms of speed, our method is significantly faster than all the other methods reporting results on UBnormal.

Results on UCSD Ped2. Our framework obtains a micro AUC of 95.4%, being 3.9% below the state-of-the-art performance of Liu *et al.* [49]. While being slightly below in terms of micro AUC, our macro AUC score is remarkably on par with the object-centric method of Ionescu *et al.* [36], and only 1.4% below that of Georgescu *et al.* [26]. Considering that our FPS is more than 30 times higher compared with these methods [26, 36] on the same GPU, our framework provides a clearly superior accuracy-speed trade-off.

Ablation study. In Table 2, we illustrate the impact of each novel component on our model’s performance. The first model is a vanilla masked AE, which obtains a rather low performance on both Avenue and ShanghaiTech. Each and every component contributes towards boosting the performance of the vanilla model. We observe that self-distillation gives the highest boost in terms of the micro AUC. How-

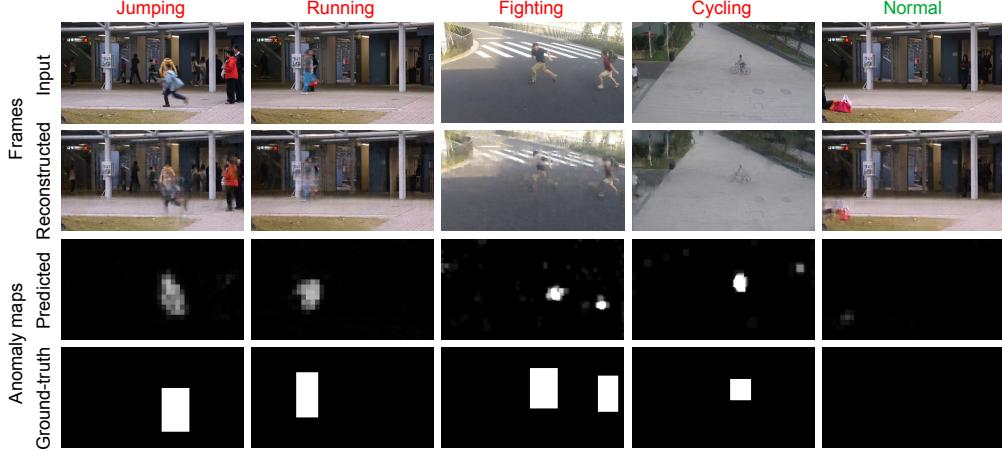


Figure 5. Examples of frames and anomaly maps reconstructed by our teacher. The first four columns correspond to abnormal examples from the Avenue and ShanghaiTech data sets, while the last column shows a normal example. Best viewed in color.

ever, to surpass the 90% milestone on Avenue, it is mandatory to introduce the prediction of anomaly maps in the learning task. An additional boost is given by our classification head.

The self-distillation procedure gives us a few possible strategies to combine the outputs of the teacher and the student. Hence, we investigate this aspect and report the results in Table 3. We observe that the best micro AUC is obtained when we combine the teacher reconstruction error with the teacher-student discrepancy.

The proportion of synthetic examples per mini-batch is another aspect that can influence our model’s performance. In Table 4, we report the micro and macro AUC scores for three possible augmentation levels. We observe that augmentation is always useful, but, for a better outcome, it requires a moderate percentage (25%).

Performance-speed trade-off. In Figure 2, we compare our model with several other methods in terms of the performance-speed trade-off. This comparison undoubtedly shows that our method reaches a far better processing speed, while achieving fairly good performance. To strengthen this observation, we also compare the methods in terms of GFLOPs and number of parameters in Table 5. We underline that the method of Gong *et al.* [28] might seem small in terms of the number of parameters, but it is slowed down by its input, which is formed of a cuboid constructed by stacking 16 consecutive frames. Moreover, the method relies on a memory module, where each memory slot records the features of one pixel in the activation maps. Although their method has twice as many parameters as our own, the large input volume and the sizable memory bank reduce the speed to 35 FPS. With an FPS of 1655, our method proves to be significantly lighter than all its competitors.

Qualitative results. In Figure 5, we present the frames and anomaly maps reconstructed by the teacher in four abnormal scenarios from Avenue and ShanghaiTech. Moreover,

Method	GFLOPs ↓	#Params (M) ↓	FPS ↑
Georgescu <i>et al.</i> [26]	107.9	65	51
Georgescu <i>et al.</i> [27]	121.6	67	24
Liu <i>et al.</i> [49]	179.5	320	12
Park <i>et al.</i> [60]	84	64	195
Gong <i>et al.</i> [28]	55.2	6	35
Ours	0.8	3	1655

Table 5. Comparing methods in terms of floating point operations (GFLOPs), number of parameters, and FPS.

in the fifth column, we illustrate the behavior of the teacher in a normal scenario. In all four abnormal cases, the reconstruction error is visibly higher for anomalous regions. This effect is mostly due to our training procedure based on synthetic data augmentation. The most obvious example is in the fourth column, where the bicycle seen in a pedestrian area is almost entirely removed from the output. Additionally, the predicted anomaly maps are well aligned with the ground-truth ones.

5. Conclusion

In this work, we proposed a lightweight masked auto-encoder (3M parameters, 0.8 GFLOPs) for video anomaly detection, which learns to reconstruct tokens with high motion gradients. Our framework is based on self-distillation, leveraging the discrepancy between teacher and student decoders for anomaly detection. Moreover, we boost the performance of our model by introducing a data augmentation technique based on overlapping synthetic anomalies on normal training data. Our highly efficient framework reached an unprecedented speed of 1655 FPS, with a minimal performance gap with respect to the state-of-the-art object-centric approaches.

Acknowledgments. Work supported by a grant of the Romanian Ministry of Education and Research, CNCS-UEFISCDI, project number PN-III-P2-2.1-PED-2021-0195, within PNCDI III.

References

- [1] Andra Acsintoaie, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UB-normal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *Proceedings of CVPR*, pages 20143–20153, 2022. [1](#), [2](#), [5](#), [6](#)
- [2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David 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. [3](#), [4](#)
- [3] Borislav Antic and Bjorn Ommer. Video parsing for abnormality detection. In *Proceedings of ICCV*, pages 2415–2422, 2011. [2](#), [3](#)
- [4] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. In *Proceedings of BMVC*, 2021. [6](#)
- [5] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic Temporal Anomaly Guided End-to-End Video Anomaly Detection. In *Proceedings of ICCVW*, pages 207–214, 2021. [2](#), [5](#), [6](#)
- [6] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Proceedings of NIPS*, pages 2654–2662, 2014. [4](#)
- [7] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *Proceedings of ECCV*, pages 348–367. Springer, 2022. [4](#)
- [8] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings. In *Proceedings of CVPR*, pages 4183–4192, 2020. [4](#), [5](#)
- [9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of ICML*, 2021. [6](#), [7](#)
- [10] Antonio Bärbläu, 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. *Computer Vision and Image Understanding*, 229: 103656, 2023. [1](#), [2](#), [3](#), [6](#)
- [11] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. SdAE: Self-distillated Masked Autoencoder. In *Proceedings of ECCV*, pages 108–124, 2022. [4](#)
- [12] Hekai Cheng, Lu Yang, and Zulong Liu. Relation-based knowledge distillation for anomaly detection. In *Proceedings of PRCV*, pages 105–116, 2021. [4](#), [5](#)
- [13] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In *Proceedings of CVPR*, pages 2909–2917, 2015. [2](#), [3](#)
- [14] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *Proceedings of CVPR*, pages 3449–3456, 2011. [2](#), [3](#)
- [15] Allison Del Giorno, J. Andrew Bagnell, and Martial Hebert. A Discriminative Framework for Anomaly Detection in Large Videos. In *Proceedings of ECCV*, pages 334–349, 2016. [2](#), [3](#), [6](#)
- [16] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of CVPR*, pages 9737–9746, 2022. [4](#), [5](#)
- [17] Fei Dong, Yu Zhang, and Xiushan Nie. Dual Discriminator Generative Adversarial Network for Video Anomaly Detection. *IEEE Access*, 8:88170–88176, 2020. [1](#), [2](#)
- [18] Keval Doshi and Yasin Yilmaz. Any-Shot Sequential Anomaly Detection in Surveillance Videos. In *Proceedings of CVPRW*, pages 934–935, 2020. [1](#), [2](#), [3](#), [6](#)
- [19] Keval Doshi and Yasin Yilmaz. Continual Learning for Anomaly Detection in Surveillance Videos. In *Proceedings of CVPRW*, pages 254–255, 2020. [3](#)
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021. [4](#)
- [21] Jayanta K. Dutta and Bonny Banerjee. Online Detection of Abnormal Events Using Incremental Coding Length. In *Proceedings of AAAI*, pages 3755–3761, 2015. [2](#), [3](#)
- [22] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, Martin D. Levine, and Fei Xiao. Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder. *Computer Vision and Image Understanding*, 195:102920, 2020. [2](#)
- [23] Ye Fei, Chaoqin Huang, Cao Jinkun, Maosen Li, Ya Zhang, and Cewu Lu. Attribute Restoration Framework for Anomaly Detection. *IEEE Transactions on Multimedia*, 24: 116–127, 2022. [3](#)
- [24] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners. In *Proceedings of NeurIPS*, 2022. [4](#)
- [25] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. Learning deep event models for crowd anomaly detection. *Neurocomputing*, 219:548–556, 2017. [3](#)
- [26] Mariana-Iuliana Georgescu, Antonio Bärbläu, 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 CVPR*, pages 12742–12752, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [27] Mariana Iuliana Georgescu, Radu 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, 2022. [2](#), [3](#), [6](#), [8](#)
- [28] 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 ICCV*, pages 1705–1714, 2019. 1, 2, 3, 6, 8
- [29] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *Proceedings of CVPR*, pages 733–742, 2016. 2, 3
- [30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of CVPR*, pages 16000–16009, 2022. 2, 3, 4
- [31] Ryota Hinami, Tao Mei, and Shin’ichi Satoh. Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge. In *Proceedings of ICCV*, pages 3639–3647, 2017. 3
- [32] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. In *Proceedings of NIPS Deep Learning and Representation Learning Workshop*, 2014. 4
- [33] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of ICCV*, pages 13545–13554, 2023. 6
- [34] Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-Supervised Masking for Unsupervised Anomaly Detection and Localization. *IEEE Transactions on Multimedia*, 2022. 4
- [35] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of ICCV*, pages 2895–2903, 2017. 2, 3, 6
- [36] 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 CVPR*, pages 7842–7851, 2019. 1, 2, 3, 4, 5, 6, 7
- [37] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using Narrowed Normality Clusters. In *Proceedings of WACV*, pages 1951–1960, 2019. 2, 3, 6
- [38] Xiangli Ji, Bairong Li, and Yuesheng Zhu. TAM-Net: Temporal Enhanced Appearance-to-Motion Generative Network for Video Anomaly Detection. In *Proceedings of IJCNN*, pages 1–8, 2020. 1
- [39] Jielin Jiang, Jiale Zhu, Muhammad Bilal, Yan Cui, Neeraj Kumar, Ruihan Dou, Feng Su, and Xiaolong Xu. Masked Swin Transformer Unet for Industrial Anomaly Detection. *IEEE Transactions on Industrial Informatics*, 19(2):2200–2209, 2022. 4
- [40] Asiegbu Miracle Kanu-Asiegbu, Ram Vasudevan, and Xiaoxiao Du. BiPOCO: Bi-Directional Trajectory Prediction with Pose Constraints for Pedestrian Anomaly Detection. In *Proceedings of SL4AD*, 2022. 6
- [41] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *Proceedings of CVPR*, pages 2921–2928, 2009. 2, 3
- [42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015. 7
- [43] 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. 1, 2, 3, 6
- [44] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-Aware Spatio-Temporal Relation Learning for Video Anomaly Detection. In *Proceedings of ECCV*, pages 333–350, 2022. 1, 3
- [45] 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. 2
- [46] Zhenyu Li, Ning Li, Kaitao Jiang, Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Superpixel Masking and Inpainting for Self-Supervised Anomaly Detection. In *Proceedings of BMVC*, 2020. 3
- [47] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection – A New Baseline. In *Proceedings of CVPR*, pages 6536–6545, 2018. 2, 3, 6
- [48] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier Two-Sample Test for Video Anomaly Detections. In *Proceedings of BMVC*, 2018. 2, 3, 6
- [49] 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 ICCV*, pages 13588–13597, 2021. 1, 2, 3, 6, 7, 8
- [50] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating Anomalies for Video Anomaly Detection With Prompt-Based Feature Mapping. In *Proceedings of CVPR*, pages 24500–24510, 2023. 6
- [51] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal Event Detection at 150 FPS in MATLAB. In *Proceedings of ICCV*, pages 2720–2727, 2013. 2, 3, 4, 5, 6, 7
- [52] Yiwei Lu, Frank Yu, Mahesh Kumar, Krishna Reddy, and Yang Wang. Few-Shot Scene-Adaptive Anomaly Detection. In *Proceedings of ECCV*, pages 125–141, 2020. 1
- [53] Weixin Luo, Wen Liu, and Shenghua Gao. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In *Proceedings of ICCV*, pages 341–349, 2017. 2, 3, 4, 6, 7
- [54] Neelu Madan, Nicolae-Catalin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):525–542, 2024. 6
- [55] Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos. Anomaly Detection in Crowded Scenes. In *Proceedings of CVPR*, pages 1975–1981, 2010. 2, 3, 4, 6
- [56] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proceedings of CVPR*, pages 935–942, 2009. 2, 3
- [57] Trong-Nguyen Nguyen and Jean Meunier. Anomaly Detection in Video Sequence With Appearance-Motion Correspondence. In *Proceedings of ICCV*, pages 1273–1283, 2019. 1, 3, 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 CVPR*, pages 12173–12182, 2020. [1](#), [3](#)
- [59] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021. [3](#)
- [60] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. FastAno: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of WACV*, pages 2249–2259, 2022. [2](#), [3](#), [6](#), [8](#)
- [61] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-guided Normality for Anomaly Detection. In *Proceedings of CVPR*, pages 14372–14381, 2020. [1](#), [2](#), [3](#), [6](#)
- [62] Bharathkumar Ramachandra and Michael Jones. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of WACV*, pages 2569–2578, 2020. [1](#), [3](#), [6](#)
- [63] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a Siamese network to localize anomalies in videos. In *Proceedings of WACV*, pages 2598–2607, 2020. [3](#), [6](#)
- [64] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. Perceptual metric learning for video anomaly detection. *Machine Vision and Applications*, 32:1432–1769, 2021. [3](#)
- [65] 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, 2022. [1](#), [3](#), [4](#)
- [66] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal Event Detection in Videos using Generative Adversarial Nets. In *Proceedings of ICIP*, pages 1577–1581, 2017. [3](#), [6](#)
- [67] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-Play CNN for Crowd Motion Analysis: An Application in Abnormal Event Detection. In *Proceedings of WACV*, pages 1689–1698, 2018. [3](#), [6](#)
- [68] Huamin Ren, Weifeng Liu, Soren Ingvar Olsen, Sergio Escalera, and Thomas B. Moeslund. Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection. In *Proceedings of BMVC*, pages 28.1–28.13, 2015. [2](#), [3](#)
- [69] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection. In *Proceedings of CVPR*, pages 13576–13586, 2022. [1](#), [2](#), [3](#), [6](#)
- [70] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, Karpathy A., A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. [4](#)
- [71] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017. [2](#), [3](#)
- [72] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. Deep-Anomaly: Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes. *Computer Vision and Image Understanding*, 172:88–97, 2018. [3](#)
- [73] Babak Saleh, Ali Farhadi, and Ahmed Elgammal. Object-Centric Anomaly Detection by Attribute-Based Reasoning. In *Proceedings of CVPR*, pages 787–794, 2013. [3](#)
- [74] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution Knowledge Distillation for Anomaly Detection. In *Proceedings of CVPR*, pages 14902–14912, 2021. [4](#), [5](#)
- [75] Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *Proceedings of CVPR*, pages 2112–2119, 2012. [2](#), [3](#)
- [76] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. Deep Appearance Features for Abnormal Behavior Detection in Video. In *Proceedings of ICIAP*, pages 779–789, 2017. [1](#), [3](#), [6](#)
- [77] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of CVPR*, pages 6479–6488, 2018. [6](#)
- [78] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos. In *Proceedings of ACMMM*, pages 184–192, 2020. [1](#), [6](#)
- [79] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64(C):187–201, 2017. [3](#)
- [80] 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. [1](#), [2](#), [3](#), [6](#)
- [81] Hanh T.M. Tran and David Hogg. Anomaly Detection using a Convolutional Winner-Take-All Autoencoder. In *Proceedings of BMVC*, 2017. [2](#), [3](#)
- [82] Anil Osman Tur, Nicola Dall’Asen, Cigdem Beyan, and Elisa Ricci. Exploring diffusion models for unsupervised video anomaly detection. In *Proceedings of ICIP*, pages 2540–2544, 2023. [6](#)
- [83] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Phung. Robust Anomaly Detection in Videos Using Multi-level Representations. In *Proceedings of AAAI*, pages 5216–5223, 2019. [1](#)
- [84] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Proceedings of ECCV*, pages 494–511, 2022. [2](#), [3](#), [6](#)
- [85] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan. Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal

- Autoencoder. In *Proceedings of ICIP*, pages 2276–2280, 2018. 2
- [86] Xusheng Wang, Mingtao Pei, and Zhengang Nie. Self-trained video anomaly detection based on teacher-student model. In *Proceedings of MLSP*, pages 1–6, 2021. 4
- [87] Ziming Wang, Yuxian Zou, and Zeming Zhang. Cluster Attention Contrast for Video Anomaly Detection. In *Proceedings of ACMMM*, pages 2463–2471, 2020. 1, 6
- [88] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of ICCV*, pages 22–31, 2021. 3, 4
- [89] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *Proceedings of ECCV*, pages 729–745, 2022. 3, 6
- [90] 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. 1, 2, 6
- [91] Shandong Wu, Brian E. Moore, and Mubarak Shah. Chaotic Invariants of Lagrangian Particle Trajectories for Anomaly Detection in Crowded Scenes. In *Proceedings of CVPR*, pages 2054–2060, 2010. 3
- [92] 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. 2
- [93] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature Prediction Diffusion Model for Video Anomaly Detection. In *Proceedings of ICCV*, pages 5527–5537, 2023. 6
- [94] Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Self-supervised video representation learning with motion-aware masked autoencoders. *arXiv preprint arXiv:2210.04154*, 2022. 4
- [95] 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 ACMMM*, pages 583–591, 2020. 1, 2, 3, 6
- [96] Jongmin Yu, Younkwon Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3572–3586, 2022. 3, 6
- [97] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-Training 3D Point Cloud Transformers With Masked Point Modeling. In *Proceedings of CVPR*, pages 19313–19322, 2022. 1
- [98] 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 CVPR*, pages 14183–14193, 2020.
- [99] 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 *Proceedings of ECCV*, pages 358–376, 2020.
- [100] Muhammad Zaigham Zaheer, Arif Mahmood, Haris M. Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative Cooperative Learning for Unsupervised Video Anomaly Detection. In *Proceedings of CVPR*, pages 14744–14754, 2022. 1, 3, 6
- [101] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-Distillation: Towards Efficient and Compact Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2022. 2, 3, 4
- [102] Xinfeng Zhang, Su Yang, Jilong Zhang, and Weishan Zhang. Video Anomaly Detection and Localization using Motion-field Shape Description and Homogeneity Testing. *Pattern Recognition*, page 107394, 2020. 2
- [103] 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. 2, 3, 6
- [104] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Online Detection of Unusual Events in Videos via Dynamic Sparse Coding. In *Proceedings of CVPR*, pages 3313–3320, 2011.
- [105] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection. In *Proceedings of CVPR*, pages 1237–1246, 2019. 2