Learning Not to Reconstruct Anomalies

Marcella Astrid[1,2]
marcella.astrid@ust.ac.kr
Muhammad Zaigham Zaheer[1,2]
mzz@ust.ac.kr
Jae-Yeong Lee[1,2]
jylee@etri.re.kr
Seung-Ik Lee[1,2]
the_silee@etri.re.kr

[1] University of Science and Technology
Daejeon, South Korea
[2] Electronics and Telecommunications
Research Institute
Daejeon, South Korea

Abstract

Video anomaly detection is often seen as one-class classification (OCC) problem due to the limited availability of anomaly examples. Typically, to tackle this problem, an autoencoder (AE) is trained to reconstruct the input with training set consisting only of normal data. At test time, the AE is then expected to well reconstruct the normal data while poorly reconstructing the anomalous data. However, several studies have shown that, even with only normal data training, AEs can often start reconstructing anomalies as well which depletes the anomaly detection performance. To mitigate this problem, we propose a novel methodology to train AEs with the objective of reconstructing only normal data, regardless of the input (i.e., normal or abnormal). Since no real anomalies are available in the OCC settings, the training is assisted by pseudo anomalies that are generated by manipulating normal data to simulate the out-of-normal-data distribution. We additionally propose two ways to generate pseudo anomalies: patch and skip frame based. Extensive experiments on three challenging video anomaly datasets demonstrate the effectiveness of our method in improving conventional AEs, achieving state-of-the-art performance.

1 Introduction

Anomalous event detection in video sequences has recently attracted significant attention [1, 3, 22, 24, 26, 30, 42, 55, 59]. The task is extremely challenging because, in real-life situations such as in surveillance videos, anomalous events do not occur frequently. Moreover, there is no restriction on the types of anomaly events that may occur, making it cumbersome to collect sufficient anomaly examples. Therefore, anomaly detection is often seen as one-class classification (OCC) problem in which only normal data is used to train a novelty detection model [3, 9, 21, 23, 24, 40, 54].

One way to tackle the OCC problem is by using a deep autoencoder (AE) trained to reconstruct normal data [9, 10, 29, 30, 35, 58]. This way, the model is encouraged to encode normalcy information within its latent space. At test time, the trained AE is then expected to only reconstruct normal cases while failing to reconstruct the anomalous cases. However,

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

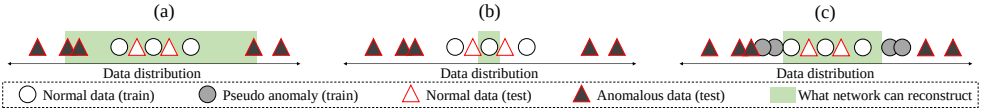

Figure 1: An illustration of reconstruction capability of the three variants of AE: (a) conventional AE trained only on normal data, (b) methods that limit reconstruction of anomalous data without pseudo anomalies, and (c) our approach which encourages the network to reconstruct only normal data with the assistance of pseudo anomalies.

as reported in literature [9, 33, 54, 60] as well as observed in our experiments (baseline performances in Fig. 6), an AE can also often reconstruct anomalous examples. It is a likely outcome as the reconstruction boundary of the AE trained only on normal data would be unconstrained as long as the boundary includes the normal data in the training set. Therefore, the reconstructions between normal and anomalous data may not be discriminative enough to successfully identify the anomalies. The phenomenon is illustrated in Fig. 1(a).

To alleviate this problem, several researchers [9, 35] proposed employing a memory mechanism over the latent space between the encoder and the decoder of an AE to limit the reconstruction capability in the case of anomalous input. The idea is to memorize normal representations learned from the training data. This way, the network is restricted to use the memorized normalcy representations for reconstruction, thus reducing its capability to regenerate anomalous data. However, such a network is highly dependent on the memory size and a small-sized memory may also limit its normal data reconstruction capability. For instance, Fig. 6 of [9] and Fig. 6 (MemAE) show that although anomalous regions have more distortions compared to the baseline, some of the normal portions are distorted as well, which may result in a limited discrimination between normal and anomalous data. The phenomenon can be attributed to the lack of a vivid reconstruction boundary when limiting the reconstruction using only normal data during training, thus resulting in a limited reconstruction capability for the normal data at test time, as illustrated in Fig. 1(b).

In this work, we also propose to limit an AE in reconstructing anomalies; however, in such a way that the normal reconstructions are not affected. Particularly, we introduce a novel training mechanism of an AE with the objective of reconstructing only normal data even if the input is anomalous. Since there are no real anomalies in the training data under the OCC setting, we propose the idea of generating and utilizing pseudo anomalies to assist the training. To this end, two types of pseudo anomaly generation methods are explored, i.e., patch and skip frame based, to simulate out of normal data distribution from normal data. By encouraging to reconstruct only normal data for any kind of input (i.e., normal or pseudo anomalous), AEs are specifically trained to limit their reconstruction boundaries around the normal data hence not affecting the normal reconstructions while distorting anomalies, as illustrated in Fig. 1(c). This results in an improved discrimination between normal and anomalous data, which is evident from the superior performance of our approach both qualitatively and quantitatively.

The contributions of this work are threefold: 1) We propose a pseudo anomaly based novel method of encouraging only normal data reconstructions to train AEs in the OCC setting. 2) We propose two types of pseudo anomalies, patch and skip frame based, to simulate anomalies. 3) In the experiments, we present extensive evaluations and analysis of the proposed training approach using each of the pseudo anomaly types on three challenging video anomaly detection datasets including Ped2 [22], Avenue [26], and ShanghaiTech [30].

2 Related Works

Reconstruction Based Methods: A common way to tackle the one-class classification (OCC) problem is by utilizing autoencoders (AEs) which learn normal data representations by reconstructing the inputs [9, 10, 29, 30, 35, 58]. However, since AEs can also well-reconstruct anomalous data [9, 33, 54, 60], several researchers proposed memory based networks to limit reconstruction capability of AEs [9, 35]. The idea is to use only the learned memory vectors for reconstruction, which helps in achieving higher reconstruction loss for anomalous inputs. However, such a configuration may also restrain the normal data reconstruction capability due to its limited memory size. In contrast, our approach encourages AEs to produce unconstrained reconstructions for normal inputs while limiting the reconstructions for anomalous inputs, thus producing more discriminative anomaly scores.

Non-Reconstruction Methods: Several researchers adopt different schemes for OCC based anomaly detection: focusing only on objects by utilizing object detectors in the frameworks [6, 7, 8, 11, 12, 43, 52]; predicting future frames from the past few consecutive frames with the intuition that it is difficult to predict unseen anomalous data [8, 24, 27, 28, 35]; or incorporating adversarial components [14, 19, 20, 24, 39, 45]. Our approach is different as we do not utilize any additional component and solely rely on the reconstruction based AEs.

Pseudo Anomalies: There have been a few recent attempts towards pseudo anomaly generation for one-class classifiers. Georgescu *et al*. [8] utilize time magnification and separate datasets as pseudo abnormal objects to train an object-centric architecture by flipping the gradient for pseudo abnormal objects. However, this approach can be only applicable to anomalies related to objects and requires a pretrained object detectors. OGNet [54] and G2D [36] propose using an under-trained and adversarially learned generator for generating fake anomaly data to train a binary classifier. Furthermore, OGNet, passes a fusion of two images to a fully-trained generator to produce another type of fake anomalous example. These approaches require a two-phase training, one for adversarial training of generator and the other phase for training binary classifier. Differently, our approach is not restricted to any predefined object classes, carries out the training in an end-to-end manner, and does not require any pretrained networks.

Data Augmentation: Pseudo anomaly generation used in our method can also be viewed as a form of data augmentation technique, widely popular among image classification models, which manipulates training data to increase variety [2, 17, 18, 53, 56]. Typically, the class labels for the augmented data are derived from the already exiting classes in the dataset. In contrast, generating pseudo anomalies can be seen as augmenting data into a new class, i.e., anomaly, which is not a part of the already existing classes.

Non-OCC Methods: In order to enhance the discrimination between normal and anomalous data, some researchers [33, 51] propose to deviate from the fundamental definition of OCC by using real anomaly examples during training. We also acknowledge a recent introduction of several weakly supervised methods using video-level binary annotations for training [42, 55, 59]. However, our approach is not directly comparable to these approaches as we do not train using any real anomaly examples. Instead, our method can be categorized as OCC because we utilize only normal training data to synthesize pseudo anomaly examples.

3 Methodology

The overall configuration of our approach (Fig. 2) consists of a reconstruction based autoencoder (AE) along with its training objective and a pseudo anomaly generator. Each component is discussed next:

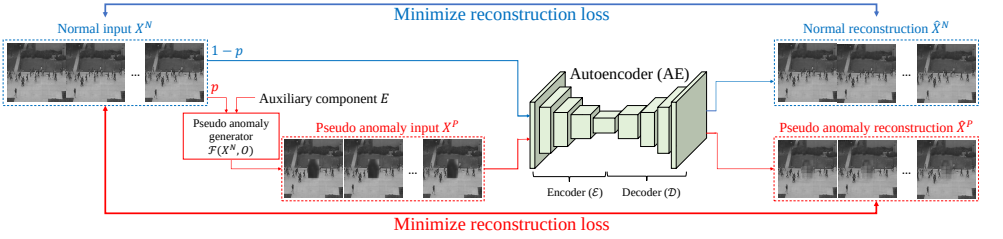

Figure 2: We train an AE for OCC using normal as well as pseudo anomaly sequences. Pseudo anomalies are generated using normal data and extra components to simulate anomalies. With a probability p, pseudo anomalies are introduced in the training examples for the AE. The AE is trained to not to reconstruct anomalies by encouraging to reconstruct the corresponding normal training examples from which the pseudo anomaly is generated.

3.1 Conventional Autoencoders in OCC Setting

For an input X, an AE can be defined as:

$$\hat{X} = \mathcal{D}(\mathcal{E}(X)), \tag{1}$$

where $\mathcal{E}$ and $\mathcal{D}$ are encoder and decoder networks, respectively. The encoder generates a latent code of a typically smaller dimension compared to the input. This code is then transformed into the reconstruction $\hat{X}$ of the input by the decoder. In order to capture rich information from video data, AEs are often designed to take multiple frames as input [9, 10, 35, 58]. Following this convention, we also set our AE model to take X as input of size $T \times C \times H \times W$, where T, C, H, and W are the number of frames, number of channels, height, and width of the frames in the input sequence, respectively.

Typically, to tackle OCC problem, AEs are utilized to learn the normal representations by minimizing the reconstruction loss between the normal input X^N and its reconstruction $\hat{X}^N$ as follows:

$$L^N = \frac{1}{T \times C \times H \times W} \left\| \hat{X}^N - X^N \right\|_F^2, \tag{2}$$

where $\|.\|_F$ is Frobenius norm. With this training setting, an AE is ideally expected to reconstruct only normal data while unable to reconstruct anomalous data. However, as widely reported across the literature, AE can often "generalize" too well and start reconstructing anomalous examples as well [9, 33, 54, 60]. We try to encourage AE to produce only normal data reconstructions for both normal and anomalous inputs, hence the network is unable to reconstruct anomalies at test time.

3.2 Learning Not to Reconstruct Anomalies

We propose a training mechanism with the objective to encourage an AE towards reconstructing only normal data regardless of the input. This means even if the data contain an abnormality, the network will learn to produce a normal reconstruction. In OCC setting, as we do not have access to real anomalous examples during training, we utilize pseudo anomalies X^P, which are generated by altering normal data X^N as:

$$X^P = \mathcal{F}(X^N, E), \tag{3}$$

where $\mathcal{F}(\cdot)$ is a pseudo anomaly generator function and E is an auxiliary component defined based on the intended pseudo anomaly type. Both $\mathcal{F}(\cdot)$ and E are discussed in Section 3.3.

To train the network, we utilize X^P as input with a probability p and X^N with a probability $1-p$, where the hyperparameter p controls the ratio of pseudo anomalies. Given normal and pseudo anomalous data as part of the training, the loss of the network is then defined as:

$$L = \begin{cases} L^N & \text{if } X = X^N, \\ L^P & \text{if } X = X^P, \end{cases} \tag{4}$$

where L^N is the reconstruction loss for normal data as defined in Eq. (2) and L^P is the reconstruction loss for pseudo anomaly data given as:

$$L^P = \frac{1}{T \times C \times H \times W} \left\| \hat{X}^P - X^N \right\|_F^2. \tag{5}$$

It may be noted that, both L^N and L^P encourage the network to reconstruct normal data. Specifically, by minimizing L^P with respect to X^N, the network attempts to remove the perturbations introduced by pseudo anomaly generator.

3.3 Generating Pseudo Anomalies

In this section, we discuss two distinct methods to generate pseudo anomalies from normal data. Specifically, we formally define the pseudo anomaly generator function $\mathcal{F}(\cdot)$ and its auxiliary component E (Eq. (3)) for each pseudo anomaly category.

3.3.1 Patch Based Pseudo Anomalies

In real world scenarios, an OCC based anomaly detection system may encounter cases like unusual objects. Derived from this motivation, we propose to overlay all normal input frames X^N with an anomalous patch A (i.e., $E = A$), which is essentially taken from an arbitrary image I^A from some other dataset, e.g., CIFAR-100 [16], referred as intruding dataset in this paper. Since such overlaid frames are not actually a part of the normal data, these are anomalous. By default, each patch is placed using SmoothMixS [18], which smooths out the boundary, to prevent the network from latching on to the edges of the patch. Nonetheless, the performance of our method is not strictly dependent on any particular patching technique, as discussed in Section 4.2.2.

To generate the i-th frame of pseudo anomaly X_i^P in the input sequence, we first transform I^A taken from the intruding dataset to $C \times H \times W$, same as the input frame size. Then, using SmoothMixS mask, a patch A of size (σ^w, σ^h) is extracted from I^A with (μ_i^x, μ_i^y) defining its center coordinates. A is then overlaid on the i-th frame X_i^N in the normal sequence. The center coordinates (μ_i^x, μ_i^y) are randomly selected within the image dimensions whereas the patch size (σ^w, σ^h) is randomly selected from 10 pixels to αW for the patch width ($\sigma^w \in [10, \alpha W]$) and from 10 pixels to αH for the patch height ($\sigma^h \in [10, \alpha H]$). α is a hyperparameter to adjust the maximum size of the patch. Detailed visualization of the process can be seen in Fig. 1 of the Supplementary. Furthermore, this technique can be generalized to utilizing a video dataset as an intruder dataset by using T frames, i.e., $I^A = (I_1^A, I_2^A, ..., I_T^A)$, with each time step corresponding to the respective time step of the input.

Moreover, to incorporate movement in pseudo anomalies, the position of a given patch in a particular frame within the sequence is changed based on the previous position as:

$$\mu_i^x = \mu_{i-1}^x + \Delta\mu_i^x, \quad \mu_i^y = \mu_{i-1}^y + \Delta\mu_i^y, \tag{6}$$

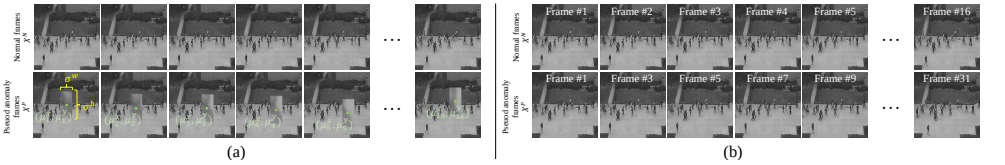

Figure 3: Examples of normal data and the corresponding pseudo anomalies generated using (a) patch and (b) skip frame based methods ($s = 2$).

where $\Delta\mu_i^x$ and $\Delta\mu_i^y$ are each randomly selected within the range of $-\beta$ to β for every i and β is the hyperparameter to adjust the maximum movement of the patch in terms of pixels. An example of patch based pseudo anomaly along with its corresponding normal data can be seen in Fig. 3(a).

3.3.2 Skip Frame Based Pseudo Anomalies

In addition to the anomalous objects as described in Section 3.3.1, in real-world scenarios, anomalies may sometimes look normal in their appearance while depicting anomalous movements. To create a system that is inclusive of this scenario, similar in essence with motion magnification of [8], we propose skipping frames to generate anomalous movements from the normal training sequences. Given a video consisting of K images $(I_1, I_2, ..., I_K)$, a normal temporally-consistent input frame sequence X^N of length T is taken from the video starting from a random n-th index as follows:

$$X^N = (I_n, I_{n+1}, ..., I_{n+(T-1)}) = (I_{n+t})_{0 \le t < T}. \tag{7}$$

To generate the corresponding pseudo anomaly X^P, we take the first frame of X^N, then replace its following frames with skipped frames $E = (I_{n+s}, I_{n+2s}, ..., I_{n+(T-1)s})$ as:

$$X^P = (I_n, E) = (I_n, I_{n+s}, ..., I_{n+(T-1)s}) = (I_{n+ts})_{0 \le t < T, s > 1}, \tag{8}$$

where s is a hyperparameter controlling the number of skipped frames. An example of a pseudo anomaly generated with skip frame method and its corresponding normal sequence can be seen in Fig. 3(b).

3.4 Inference

At test time, we process the input sequences and their anomaly scores. Concurrent with other recent anomaly detection methods [5, 24, 35], we utilize Peak Signal to Noise Ratio (PSNR) P_t between an input frame and its reconstruction to compute the anomaly score as follows:

$$P_t = 10 \log_{10} \frac{M_{\hat{I}_t}^2}{\frac{1}{R} \left\| \hat{I}_t - I_t \right\|_F^2}, \tag{9}$$

where t is the frame index, I_t is the t-th frame input, $\hat{I}_t$ is the reconstruction of I_t, R is the total number of pixels in $\hat{I}_t$, and $M_{\hat{I}_t}$ is the maximum possible pixel value of $\hat{I}_t$. Finally, following [5, 24, 35], the anomaly score S_t is obtained using min-max normalization of P_t as:

$$S_t = 1 - \frac{P_t - \min_t(P_t)}{\max_t(P_t) - \min_t(P_t)}, \tag{10}$$

where a higher S_t value represents higher reconstruction error compared to the other frames in the test video and vice versa.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our approach on three publicly available video anomaly detection datasets, i.e., Ped2 [22], Avenue [26], and ShanghaiTech [30]. We utilize the standard division of the datasets in which training splits consist of only normal videos. Whereas, every video in each of the test sets contains one or more anomalous portions. Further details about the datasets are provided in the Supplementary.

Evaluation Criteria. We evaluate our approach using the widely popular frame-level area under the ROC curve (AUC) metric [54]. The ROC curve is obtained by varying the anomaly score thresholds to plot false and true positive rates across the whole test set, i.e., one ROC curve for a dataset. Higher AUC values represent more accurate results.

Parameters and Implementation Details. For the AE architecture, we use a 3D convolution-deconvolution network similarly proposed by Gong *et al*. [9]. The AE takes an input sequence X (Eq. (1)) of size $16 \times 1 \times 256 \times 256$ and produces its reconstruction of the same size. During training, the reconstruction loss is calculated across all of the 16 frames. At test time, only the 9th frame of a sequence is considered for anomaly score calculation (Eq. (9) - (10)). Further details on the architecture are provided in the Supplementary. By default, for all datasets, we utilize $p = 0.2$, $s = \{2,3,4,5\}$, $\alpha = 0.5$, CIFAR-100 [16] as the intruder dataset, and SmoothMixS [18] as the patching technique. Moreover, β is set to 10 for Ped2 and 25 for the other datasets. $s = \{2,3,4,5\}$ means s is randomly selected as 2, 3, 4 or 5 each time we generate pseudo anomaly. To observe the robustness of our method, in Section 4.2.2, we also perform evaluations by varying hyperparameters. Training is carried out separately for the model trained without pseudo anomalies (referred as baseline; see Section 3.1), the model trained using patch based pseudo anomalies, and the model trained using skip frame based pseudo anomalies. The code is provided at *https://github.com/aseuteurideu/LearningNotToReconstructAnomalies*.

4.2 Quantitative Results

4.2.1 Comparisons with the Baseline and SOTA Methods

Table 1 shows the AUC comparisons of our overall model with the existing state-of-the-art (SOTA) approaches on Ped2 [22], Avenue [26], and ShanghaiTech [30] datasets. For a fair comparison, we classify various SOTA methods into five categories: 1) Non-deep learning approaches, 2) Object-centric methods which utilize object detectors to focus only on the detected objects, 3) Prediction based methods that predict a future frame to detect anomalies, 4) Reconstruction based approaches that use reconstruction of the input to detect anomalies, and 5) Miscellaneous methods which are either the tasks not belonging to the aforementioned categories or employing a combination of these. Our method falls in the category of reconstruction based methods.

Comparing to the other approaches of the same category, i.e., reconstruction, our model achieves the best performance on all three benchmark datasets. Interestingly, both of our models trained with different kinds of pseudo anomalies achieve better performance than the memory based networks, such as MNAD-Recon [35] and MemAE [9], considering that we use a very similar network architecture with these approaches and a common goal of limiting the AE capability of reconstructing anomalies. It may also be noted that our proposed pseudo anomaly based trained models provide consistent gains over the respective baselines on all three datasets. This clearly demonstrates the superiority of our proposed approach, i.e., training AEs by encouraging only normal data reconstructions assisted by pseudo anomalies.

	Methods	Ped2 [24]	Ave [26]	Sh [30]
Miscellaneous	AbnormalGAN [33]	93.5%	-	-
	Smeureanu *et al.* [40]	-	84.6%	-
	AMDN [43, 50]	90.8%	-	-
	STAN [13]	96.5%	87.2%	-
	MC2ST [25]	87.5%	84.4%	-
	Ionescu *et al.* [18]	-	88.9%	-
	BMAN [20]	96.6%	**90.0%**	76.2%
	AMC [42]	96.2%	86.9%	-
	Vu *et al.* [45]	**99.21%**	71.54%	-
	DeepOC [41]	-	86.6%	-
	TAM-Net [16]	98.1%	78.3%	-
	LSA [1]	95.4%	-	72.5%
	Ramachandra *et al.* [35]	94.0%	87.2%	-
	Tang *et al.* [44]	96.3%	85.1%	73.0%
	Wang *et al.* [46]	-	87.0%	**79.3%**
	OGNet [52]	98.1%	-	-
	Conv-VRNN [7]	96.06%	85.78%	-
	Chang *et al.* [3]	96.5%	86.0%	73.3%
Object-centric	MT-FRCN [31]	92.2%	-	-
	Ionescu *et al.* [18] [1]	94.3%	87.4%	78.7%
	Doshi and Yilmaz [9, 8]	97.8%	86.4%	71.62%
	Sun *et al.* [38]	-	89.6%	74.7%
	VEC [51]	97.3%	89.6%	74.8%
	Georgescu *et al.* [8]	**98.7%**	**92.3%**	**82.7%**

	Methods	Ped2 [24]	Ave [26]	Sh [30]
Non deep learning	MPPCA [15]	69.3%	-	-
	MPPC+SFA [15]	61.3%	-	-
	Mehran *et al.* [22]	55.6%	-	-
	MDT [22]	82.9%	-	-
	Lu *et al.* [26]	-	**80.9%**	-
	AMDN [43]	90.8%	-	-
	Del Giorno *et al.* [6]	-	78.3%	-
	LSHF [39]	**91.0%**	-	-
	Xu *et al.* [48]	88.2%	-	-
	Ramachandra and Jones [36]	88.3%	72.0%	-
Prediction	Frame-Pred [24]	95.4%	85.1%	72.8%
	Dong *et al.* [5]	95.6%	84.9%	73.7%
	Lu *et al.* [26]	96.2%	85.8%	**77.9%**
	MNAD-Pred [30]	**97.0%**	**88.5%**	70.5%
Reconstruction	AE-Conv2D [10]	90.0%	70.2%	60.85%
	AE-Conv3D [26]	91.2%	71.1%	-
	AE-ConvLSTM [26]	88.10%	77.00%	-
	TSC [10]	91.03%	80.56%	67.94%
	StackRNN [10]	92.21%	81.71%	68.00%
	MemAE [4]	94.1%	83.3%	71.2%
	MNAD-Recon [30]	90.2%	82.8%	69.8%
	Baseline	92.49%	81.47%	71.28%
	Ours: Patch based	94.77%	**84.91%**	72.46%
	Ours: Skip frame based	**96.50%**	84.67%	**75.97%**

[1] Micro AUC reported in [8]

Table 1: AUC performance comparison of our approach with several existing SOTA methods on Ped2, Avenue (Ave), and ShanghaiTech (Sh). Best and second best performances are highlighted as bold and underlined, in each category and dataset.

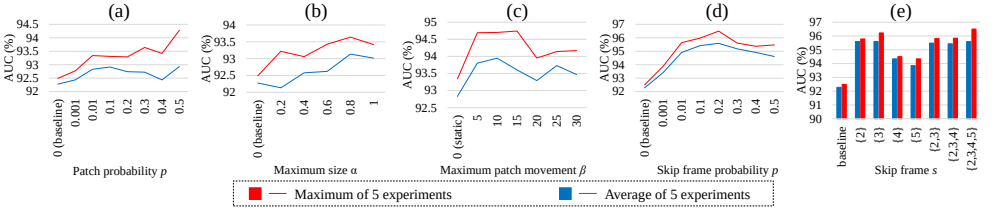

Figure 4: Evaluations to show the robustness of our method on a wide-range of hyperparameters values: (a) patch probability p given $\alpha = 0.5$ and $\beta = 0$; (b) maximum patch size α given $p = 0.01$ and $\beta = 0$; (c) maximum patch movement β given $p = 0.01$ and $\alpha = 0.5$; (d) skip frame probability p given $s = \{2,3,4,5\}$; (e) skip frame parameter s given $p = 0.2$.

Looking at the techniques in other categories, our proposed approach demonstrates a comparable performance. Particularly, compared with the architectures that are designed with complex components, such as attention, optical flow, adversarial training, LSTM, etc., for example, in BMAN [20], our method provides an overall comparable performance without any bells and whistles. In addition, most methods in the object-centric category [6, 7, 8, 17, 43, 52] require pre-trained object detectors which make their applicability limited to the set of predefined object categories. In contrast, while our method is generic and can be applied to a variety of AE based architectures, it is also not constrained by object detectors.

4.2.2 Hyperparameters Evaluation

To evaluate the robustness against the hyperparameters introduced in this work, we provide extensive analysis in Fig. 4 and Fig. 5. Only Ped2 is used to limit the span of experiments.

Fig. 4(a)-(c) show the evaluations for the hyperparameters used in patch based pseudo anomalies. Typically, the model trained using static patch based pseudo anomalies ($\beta = 0$) successfully outperforms the baseline (Fig. 4(a)-(b)) and moving the patch location ($\beta > 0$)

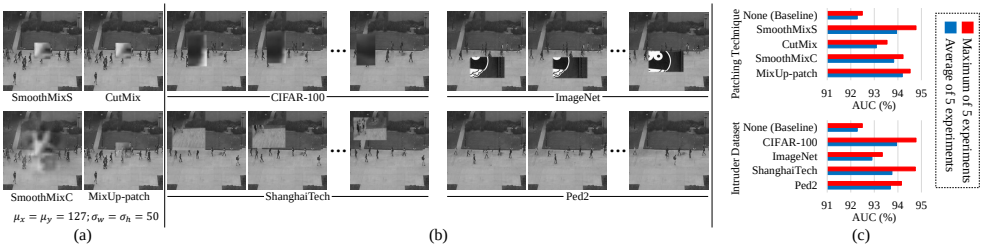

Figure 5: (a) Pseudo anomalous frame samples generated using different patch techniques: SmoothMixS [18], CutMix [53], SmoothMixC [18], and MixUp-patch (MixUp [56] with CutMix patch); (b) Visualizations of patch based pseudo anomaly sequence with SmoothMixS and different intruder datasets; (c) AUC comparisons between different patch techniques and intruder datasets.

further improves the overall performance (Fig. 4(c)). Additionally, we experiment using different patching techniques and intruder datasets. Pseudo anomalies generated using each of these methods can be seen in Fig. 5(a) & (b). Fig. 5(c) shows the robustness of our method across different types of patching techniques and intruder datasets in outperforming the baseline. Interestingly, using Ped2 itself as an intruder dataset can also elevate the performance. It can be attributed to the anomalous shapes remained after cropping normal objects as we take patches from the Ped2 (Fig. 5(b)). Note that, we utilize only training sets of the intruder datasets. More details on patching techniques are provided in the Supplementary.

Evaluations on a wide-range of hyperparameters used in skip frame based pseudo anomalies, i.e., p and s, can be seen in Fig. 4(d)-(e). All experiments show performance gains over the baseline, with maximum performance achieved with $p = 0.2$ and $s = \{2,3,4,5\}$.

4.3 Qualitative Results

For a deeper understanding on how our method improves the baseline, in this section, we provide a qualitative comparisons of the baseline, our models, and MemAE [9]. Fig. 6 shows several input test images from the three datasets, the reconstructions produced by different variants of our model, and the reconstruction error heatmaps. The heatmaps are generated by computing the squared error of each pixel between the input frame and its reconstruction, followed by min-max normalization. Based on the reconstruction error visualizations, it can be observed that both of our models successfully highlight the anomalous region more than the baseline by well reconstructing the normal data and poorly reconstructing the anomalous data, which results in the superior anomaly discrimination capability of our models. Furthermore, a few interesting observations may as well be noted. In the patch based method examples, the backpack being tossed up in Avenue is completely vanished from the reconstructions. Instead, our model reconstructs normal background. Similarly, reconstructions of bicycle riders produced by skip frame based model in ShanghaiTech and Ped2 are dislocated as if they were moving at a normal pedestrian pace. This demonstrates that, concurrent to our training objective, the learned model attempts to reconstruct normal even if the input is anomalous. More qualitative results are provided in the Supplementary.

Furthermore, compared with MemAE [9] on Ped2, our model retains the quality of normal reconstructions. As discussed in Section 1 & 2, in addition to limiting reconstructions of anomalous regions, attributed to limited memory size, MemAE may also limit the reconstructions of normal parts. Therefore, it is less discriminative. It may be noted that, as MemAE official code provides only Ped2 trained model, we compare using only this dataset.

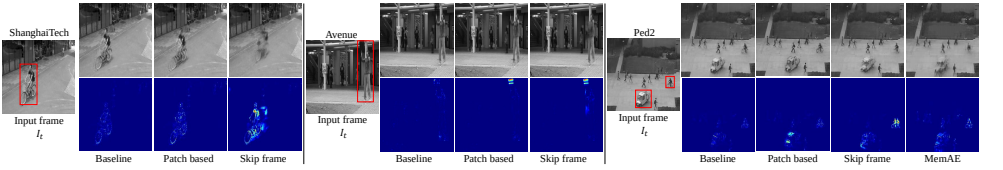

Figure 6: Visualizations of input test frames, reconstructions (first row), and reconstruction error heatmaps (second row) of the baseline, our model trained using patch based pseudo anomalies, and our model trained using skip frame based pseudo anomalies on ShanghaiTech, Avenue, and Ped2. Additionally, we provide comparison with MemAE [9] on Ped2. The anomalous regions are marked with red boxes.

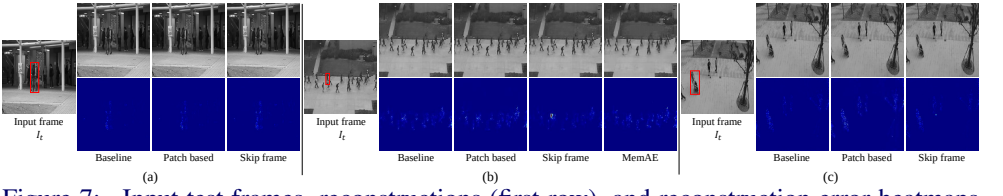

Figure 7: Input test frames, reconstructions (first row), and reconstruction error heatmaps (second row) on several examples when (a) both patch based and skip frame models failed, (b) patch based model failed, and (c) skip frame based model failed.

To further understand the behavior of our models, Fig. 7 shows various cases of failures produced by our model. Since the bicycle in Fig. 7(a) is too thin, both of our models, as well as the baseline, have difficulties in detecting the anomaly. Riding a skateboard in Fig. 7(b) is abnormal mainly because of its movement, but not because of its appearance as the skateboard is almost invisible. Our patch based pseudo anomaly model tends to have difficulties in such cases. The baseline and MemAE models also exhibit the same problem in this frame. Walking with a stroller in Fig. 7(c) is abnormal in its appearance, while the movement is normal. The skip frame based model tends to have the drawback in such case. Overall, these observations indicate that more careful choices of pseudo anomalies may lead to even better performances of the anomaly detection models.

5 Conclusion

We propose a training mechanism of an autoencoder (AE) assisted by pseudo anomalies for one-class classification with the objective to reconstruct only normal data even if the input is not normal. This consequently increases the reconstruction error of anomalous inputs without restraining normal reconstructions, which leads to highly discriminative anomaly scores. To carry out this training, we additionally propose two pseudo anomaly generation methods, i.e., patch and skip frame based. Extensive evaluations on three challenging video anomaly datasets demonstrate that our proposed training methodology is effective for improving the capability of an AE to detect anomalies.

Acknowledgements. This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01309, Development of AI Technology for Guidance of a Mobile Robot to its Goal with Uncertain Maps in Indoor/Outdoor Environments)

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.

[2] Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger-Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172. JMLR Workshop and Conference Proceedings, 2011.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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 & Machine Intelligence*, (01):1–1, 2021.

[9] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

[18] 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 Workshops*, pages 756–757, 2020.

[19] 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.

[20] 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.

[21] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Anomaly detection via online oversampling principal component analysis. *IEEE transactions on knowledge and data engineering*, 25(7):1460–1470, 2012.

[22] 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, 2013.

[23] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[24] 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.

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

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

[27] 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.

[28] 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.

[29] 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.

[30] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[31] 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.

[32] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009.

[33] Asim Munawar, Phongtharin Vinayavekhin, and Giovanni De Magistris. Limiting the reconstruction capability of generative neural network using negative learning. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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*, volume 33, pages 5216–5223, 2019.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] Yuki Yamanaka, Tomoharu Iwata, Hiroshi Takahashi, Masanori Yamada, and Sekitoshi Kanai. Autoencoding binary classifiers for supervised anomaly detection. In *Pacific Rim International Conference on Artificial Intelligence*, pages 647–659. Springer, 2019.

[52] 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.

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

[54] 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.

[55] 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 the European Conference on Computer Vision (ECCV)*, 2020.

[56] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

[57] 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.

[58] 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.

[59] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019.

[60] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.