CONSTRICTING NORMAL LATENT SPACE FOR ANOMALY DETECTION WITH NORMAL-ONLY TRAINING DATA

Marcella Astrid[1,2,3] **Muhammad Zaigham Zaheer**[4] **Seung-Ik Lee**[1,2*]
[1]Department of Artificial Intelligence,
University of Science and Technology, South Korea
[2]Field Robotics Research Section,
Electronics and Telecommunications Research Institute, South Korea
[3]Interdisciplinary Centre for Security, Reliability and Trust,
University of Luxembourg, Luxembourg
[4]Department of Computer Vision,
Mohamed Bin Zayed University of Artificial Intelligence, United Arab Emirates
marcella.astrid@uni.lu
zaigham.zaheer@mbzuai.ac.ae
the_silee@etri.re.kr [*]corresponding author

ABSTRACT

In order to devise an anomaly detection model using only normal training data, an autoencoder (AE) is typically trained to reconstruct the data. As a result, the AE can extract normal representations in its latent space. During test time, since AE is not trained using real anomalies, it is expected to poorly reconstruct the anomalous data. However, several researchers have observed that it is not the case. In this work, we propose to limit the reconstruction capability of AE by introducing a novel latent constriction loss, which is added to the existing reconstruction loss. By using our method, no extra computational cost is added to the AE during test time. Evaluations using three video anomaly detection benchmark datasets, i.e., Ped2, Avenue, and ShanghaiTech, demonstrate the effectiveness of our method in limiting the reconstruction capability of AE, which leads to a better anomaly detection model.

1 INTRODUCTION

Anomaly detection is a challenging problem due to the rarity of anomalous event occurrences. Therefore, it is typically approached as one class classification (OCC) problem where data instances of only one class, i.e., normal class, are used in the training (Hasan et al., 2016; Zaheer et al., 2020; Zhao et al., 2017; Luo et al., 2017a; Georgescu et al., 2021; Park et al., 2021). In this setting, an autoencoder (AE) is generally used to encode the normalcy representations in its latent space (Hasan et al., 2016; Zhao et al., 2017; Luo et al., 2017a; Park et al., 2020; Gong et al., 2019). Specifically, the AE is trained to reconstruct the normal data given in the training set. Since AE is not trained using real anomalies, it is expected to not reconstruct anomalies during test time. Therefore, the normal and anomalous data should be distinguishable by the difference in their reconstruction quality. However, as observed by several researchers (Astrid et al., 2021a;b; Gong et al., 2019), AE can often 'generalize' so well so it can start reconstruct any input, including the anomalous data. This property reduces the capability of an AE to discriminate between normal and anomalous data. It occurs because there is no mechanism constricting the latent space, which enables the latent space to grow too large. Consequently, the AE adapts and learns to reconstruct from this large latent space, which can cover the anomalous data as well.

To prevent this from happening, several researchers utilize memory mechanism to limit the latent space (Gong et al., 2019; Park et al., 2020). The AE is bound to reconstruct data using latent space within the span of memory vectors. However, as mentioned in (Gong et al., 2019), this requires additional computational costs to read the memory during test time. Rather than limiting the la-

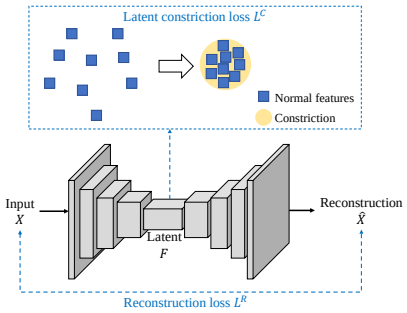

Figure 1: Overall configuration of our proposed method which consists of an autoencoder (AE) trained using reconstruction loss and our proposed latent constriction loss. The latent constriction loss restrains the normal features into smaller space in order to limit the reconstruction capability of the AE.

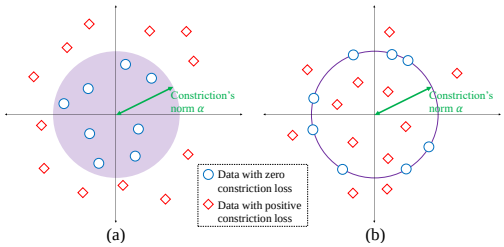

Figure 2: Two types of the proposed latent constriction losses: (a) constricting inside the norm sphere, (b) constricting on the surface of the norm sphere.

tent space using memory vectors, in this work, we propose to approach the problem directly by introducing a new training loss that constricts the latent space. In this way, AE learns to produce reconstructions using only the constricted latent space. Moreover, since we only add an additional loss during training time, there is no extra computational cost during test time.

In summary, our contributions are as follows: 1) We propose a novel constriction loss to limit the reconstruction capability of AE without any additional component added to the architecture; 2) We introduce and explore two types of the proposed constriction loss; 3) We evaluate the effectiveness of our method on three video anomaly benchmark datasets, i.e., Ped2 (Li et al., 2013), Avenue (Lu et al., 2013), and ShanghaiTech (Luo et al., 2017b).

2 Related Works

To limit the reconstruction capability of AE, in addition to the aforementioned memory-based methods (Park et al., 2020; Gong et al., 2019), several existing studies have utilized pseudo anomalies (Astrid et al., 2021a;b). However, such methods require prior knowledge to generate the pseudo anomaly. On the other hand, our method does not require it. We also acknowledge other methods which are not reconstruction-based methods. For example, binary classifier (Zaheer et al., 2020; Pourreza et al., 2021; Georgescu et al., 2021) and prediction task (Liu et al., 2018; Georgescu et al., 2021; Dong et al., 2020). Additionally, there are methods that utilize real anomalous data (Zaheer et al., 2023; Munawar et al., 2017; Sultani et al., 2018). Our method, on the other hand, uses reconstruction-based method and only normal data for the training.

3 Methodology

In this section, we discuss each component of our approach as seen in Figure 1, including conventional AE trained using reconstruction loss (Section 3.1), the novel constricting loss (Section 3.2), and the inference technique (Section 3.3).

3.1 Training AE to represent normal data

In OCC setting, an AE is typically used to create representation of normal data. The AE takes input X of size $T \times C \times H \times W$ and outputs its reconstruction $\hat{X}$ of the same size, where T, C, H, W are respectively the number of frames, number of channels, height, and width of the input. An AE consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. $\mathcal{E}$ encodes X into a latent feature F of size $T' \times C' \times H' \times W'$. $\mathcal{D}$ then decodes F into $\hat{X}$. To learn the normal representation in the latent space, conventionally, the AE is trained to reconstruct normal data available in the training data as:

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

where $\|\cdot\|_F$ is Frobenius norm.

3.2 CONSTRICTING THE NORMAL DATA LATENT SPACE

Since the AE is trained using only normal data, the latent features ideally consists of only normal representations. However, as nothing restricting the latent features, the latent space can grow too large, which may include lot of anomalous data features. This can lead to the problem where AE can reconstruct anomalous data as well as the normal data. Therefore, we propose to limit the latent space by adding a new constriction loss.

The overall loss in our method combines the reconstruction loss L^R in equation 1 and our proposed constriction loss L^C as:

$$L = L^R + \lambda L^C, \tag{2}$$

where λ is the loss weighting hyperparameter.

L^C specifically constricts the latent features $F_{j,i} \in \mathbb{R}^{T' \times C'}$ taken from F. We combine the channel C' and time T' dimension to incorporate both appearance and time information inside the latent features that we optimize. In this work, we propose two types of constriction losses using norm sphere which we discuss in details next.

3.2.1 CONSTRICTING INSIDE THE SPHERE

For the first type of constriction loss, we constrict the normal data to be inside a norm sphere. Therefore, L^C is defined as:

$$L^C = \frac{1}{H' \times W'} \sum_{j=1}^{H'} \sum_{i=1}^{W'} \max(0, \|F_{j,i}\|_2 - \alpha), \tag{3}$$

where $\|\cdot\|_2$ is L2-norm and α is a hyperparameter to control the size of the sphere. If $\|F_{j,i}\|_2 \leq \alpha$, the $F_{j,i}$ is already inside the norm sphere, hence the constriction loss is zero. Otherwise, the loss is positive. Illustration of the loss can be seen in Figure 2(a).

3.2.2 CONSTRICTING ON THE SURFACE OF THE SPHERE

Another way to limit the features is by constricting on the surface of a norm sphere as:

$$L^C = \frac{1}{H' \times W'} \sum_{j=1}^{H'} \sum_{i=1}^{W'} \left| \|F_{j,i}\|_2 - \alpha \right|, \tag{4}$$

where $|\cdot|$ is absolute operation. L^C is zero only if $\|F_{j,i}\|_2 = \alpha$. Outside of the surface, whether by having smaller or larger norm, the loss will be positive. Illustration of the loss can be seen in Figure 2(b).

3.3 INFERENCE

During test time, we follow the inference mechanism used in several reconstruction-based methods (Park et al., 2020; Astrid et al., 2021a;b) by utilizing PSNR value $\mathcal{P}_t$ of t-th input frame I_t and its reconstruction $\hat{I}_t$. $\mathcal{P}_t$ is then min-max normalized across each test video to produce a normalcy score $\mathcal{N}_t$ of range $[0, 1]$. Finally, the anomaly score the frame is calculated as:

$$\mathcal{A}_t = 1 - \mathcal{N}_t. \tag{5}$$

In this way, a higher value of $\mathcal{A}_t$ indicates a higher level of abnormality found in I_t.

4 EXPERIMENTS

4.1 DATASETS

We evaluate our method on three video anomaly benchmark datasets, i.e., Ped2 (Li et al., 2013), Avenue (Lu et al., 2013), and ShanghaiTech (Luo et al., 2017b). The training set of each of the dataset consists only of normal frames, whereas each video in the test set consists of one or more anomalous portions.

Ped2. It contains 16 training and 12 test videos. Normal behavior is defined as walking pedestrians. Whereas abnormal behaviors include riding bicycles, skateboards, and vehicles.

Methods	Ped2	Ave	Sh
AE-Conv2D (Hasan et al., 2016)	90.0%	70.2%	60.85%
AE-Conv3D (Zhao et al., 2017)	91.2%	71.1%	-
AE-ConvLSTM (Luo et al., 2017a)	88.10%	77.00%	-
TSC (Luo et al., 2017b)	91.03%	80.56%	67.94%
StackRNN (Luo et al., 2017b)	92.21%	81.71%	68.00%
STEAL Net (Astrid et al., 2021b)	**98.4%**	**87.1%**	<u>73.7%</u>
LNTRA - Patch (Astrid et al., 2021a)	94.77%	<u>84.91%</u>	72.46%
LNTRA - Skip frame (Astrid et al., 2021a)	<u>96.50%</u>	84.67%	**75.97%**
MemAE (Gong et al., 2019)	94.1%	83.3%	71.2%
MNAD-Reconstruction (Park et al., 2020)	90.2%	82.8%	69.8%
Baseline	92.49%	81.47%	71.28%
Ours-Inside Sphere	94.66%	82.93%	71.35%
Ours-On Sphere Surface	94.05%	82.14%	71.39%

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

Avenue. It consists of 16 training and 21 test videos, where walking people are considered normal. Whereas, examples of anomalous behaviors are throwing bags and running.

ShanghaiTech. It is by far the largest one-class anomaly detection dataset consisting of 330 training and 107 test videos recorded at 13 different locations with various camera angles and lighting conditions. In total, there are 130 anomalous events in the test set including running, riding bicycle, and fighting.

4.2 Experimental setup

Evaluation metric. We utilize a widely used frame-level area under the ROC curve (AUC) to evaluate our method. The ROC curve is computed for all videos in each test set, i.e., one ROC curve for each dataset.

Implementation details. As the baseline AE, we use the baseline in (Astrid et al., 2021a) and train it using only reconstruction loss (i.e., $\lambda = 0$ in equation 2). In the AE trained using our constriction loss, we remove the last LeakyReLU of the encoder to let the latent space covers the negative and positive space equally before constricting it with our proposed loss. The input and output of each model have size of $16 \times 1 \times 256 \times 256$. The training configurations, such as mini batch size, learning rate, and optimizer, follow (Astrid et al., 2021a). During inference (Section 3.3), we compute anomaly score using the 9th frame. By default, for all experiments, we set the loss weighting hyperparameter to $\lambda = 0.0001$. As for constricting inside the sphere, we set $\alpha = 0.1$ for Ped2 and $\alpha = 1$ for the other datasets. Whereas for constricting on the surface of the sphere, we set $\alpha = 10$ for Ped2 and $\alpha = 100$ for the other datasets. We repeat our experiments on both the baseline and our method five times and select the maximum performance.

4.3 Comparisons with the baseline

To find out the effectiveness of our method in limiting reconstruction capability of AE, we compare our method with the baseline AE. The AUC comparisons can be seen in the last three rows of Table 1. Each of our method, using either of the constriction methods, successfully outperforms the baseline which demonstrates the effectiveness of our approach. Moreover, as seen qualitatively in Figure 3, our method successfully limits the reconstruction capability of AE which consequently produces higher reconstruction error in the anomalous regions compared to the baseline.

4.4 Comparisons with other methods

For fair comparisons, we compare our approach only to reconstruction-based methods (i.e., methods that use reconstruction as the sole decision factor of anomaly score) in Table 1.

As seen, our method is on par with memory-based networks, i.e., MemAE and MNAD-Reconstruction. With similar performance, our method does not required any additional computation cost on top of the baseline AE during test time, while memory based networks require memory reading operation (Gong et al., 2019). This also demonstrates that directly limiting the latent space using constriction loss performs similarly to limiting using memory.

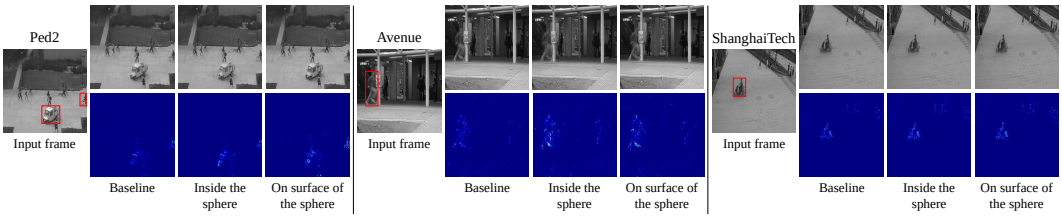

Figure 3: Qualitative comparisons of the baseline and our method in test samples from each dataset. The top and bottom rows are the outputs of the AE and the respective reconstruction error heatmaps. Reconstruction error heatmaps are computed using reconstruction error in the frame followed by min-max normalization in a frame. Red boxes mark the anomalous regions. Our method successfully distorts the anomalous regions better than the baseline.

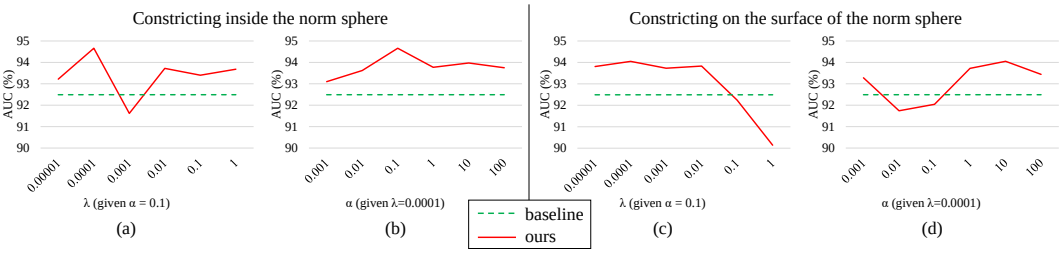

Figure 4: Evaluation of the hyperparameters used in our method on Ped2 dataset in constricting (a)-(b) inside the norm sphere and (c)-(d) on the surface of the norm sphere. Different loss weighting λ (equation 2) and constricting norm α (equation 3 & equation 4) are used. As mostly our method (red solid line) outperforms the baseline (green dotted line), our method is robust towards different hyperparameter values.

However, compared to method utilizing pseudo anomalies, such as STEAL Net (Astrid et al., 2021b) and LNTRA (Astrid et al., 2021a), our method is inferior. In spite of that, our method does not require any prior knowledge while still achieving competitive performance. On the other hand, STEAL Net (Astrid et al., 2021b) and LNTRA (Astrid et al., 2021a) require prior knowledge on what anomalies are, such as anomalous speed and appearance.

4.5 Hyperparameters evaluation

In our proposed method, we introduce two new hyperparameters, i.e., loss weighting λ and constricting norm α. Figure 4 shows the results of our experiments using different values of hyperparameters. To limit the span of experiments, the experiments are conducted in Ped2 only. In most values, our method is robust and shows better performances compared to the baseline. However, constricting on the surface generally does not work with smaller α values because it is already very constricting compared to constricting inside the sphere. Using a smaller α can lead to too much constrictions, which can also hinder the AE to reconstruct any data, including the normal data itself.

5 Conclusion

In this paper, we propose a novel constriction loss to limit the reconstruction capability of AE in anomaly detection. The loss is applied to the latent space between the encoder and the decoder of AE. We introduce two types of constriction losses, i.e., constricting the latent inside norm sphere and on the surface of norm sphere. The evaluations on Ped2, Avenue, and ShanghaiTech datasets demonstrate the effectiveness of our method in improving the capability of the AE in distorting the anomalous inputs. Moreover, compared to memory-based networks (Gong et al., 2019; Park et al., 2020) that require the computational burden of reading memory, our method does not require any additional computation cost during test time.

ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions)

REFERENCES

Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. *British Machine Vision Conference*, 2021a.

Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 207–214, October 2021b.

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

Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12742–12752, 2021.

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*, pp. 1705–1714, 2019.

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*, pp. 733–742, 2016.

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.

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*, pp. 6536–6545, 2018.

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*, pp. 2720–2727, 2013.

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)*, pp. 439–444. IEEE, 2017a.

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*, pp. 341–349, 2017b.

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)*, pp. 1–6. IEEE, 2017.

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*, pp. 14372–14381, 2020.

YeongHyeon Park, Won Seok Park, and Yeong Beom Kim. Anomaly detection in particulate matter sensor using hypothesis pruning generative adversarial network. *ETRI Journal*, 43(3):511–523, 2021.

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*, pp. 2003–2012, 2021.

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*, pp. 6479–6488, 2018.

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*, pp. 14183–14193, 2020.

Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Clustering aided weakly supervised training to detect anomalous events in surveillance videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

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*, pp. 1933–1941, 2017.