



Relation-Based Knowledge Distillation for Anomaly Detection

Hekai Cheng, Lu Yang^(✉), and Zulong Liu

School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, Sichuan, People's Republic of China
yanglu@uestc.edu.cn

Abstract. Anomaly detection is a binary classification task, which is to judge whether the input image contains an anomaly or not and the difficulty is that only normal samples are available at training. Due to this unsupervised nature, the classic supervised classification methods will fail. Knowledge distillation-based methods for unsupervised anomaly detection have recently drawn attention as it has shown the outstanding performance. In this paper, we present a novel knowledge distillation-based approach for anomaly detection (RKDAD). We propose to use the “distillation” of the “FSP matrix” from adjacent layers of a teacher network, pre-trained on ImageNet, into a student network which has the same structure as the teacher network to solve the anomaly detection problem, we show that the “FSP matrix” are more discriminative features for normal and abnormal samples than other traditional features like the latent vectors in autoencoders. The “FSP matrix” is defined as the inner product between features from two layers and we detect anomalies using the discrepancy between teacher’s and student’s corresponding “FSP matrix”. To the best of our knowledge, it is the first work to use the relation-based knowledge distillation framework to solve the unsupervised anomaly detection task. We show that our method can achieve competitive results compared to the state-of-the-art methods on MNIST, F-MNIST and surpass the state-of-the-art results on the object images in MVTecAD.

Keywords: Anomaly detection · Knowledge distillation · “FSP matrix”

1 Introduction

Anomaly detection (AD) is a binary classification problem and it has been approached in a one-class learning setting, i.e., the task of AD is to identify abnormal samples during testing while only normal samples are available at training. It has been an increasingly important and demanding task in many domains of computer vision, like in the field of visual industrial inspection [1, 2], the anomalies are rare events so it is usually required that we only train machine learning models on normal product images and detect abnormal images during inference. Moreover, anomaly detection is widely used in health monitoring [3], video surveillance [4] and other fields.

In recent years, a lot of studies have been done to improve the performance of anomaly detection [5–13] in images. Among them, some methods especially the methods based

This research was supported by NSFC (No. 61871074).

© Springer Nature Switzerland AG 2021

H. Ma et al. (Eds.): PRCV 2021, LNCS 13019, pp. 105–116, 2021.

https://doi.org/10.1007/978-3-030-88004-0_9

on deep learning have achieved great success. Most methods [5–7] model the normal data abstraction by extracting discriminative latent features, which are used to determine whether the input is normal or abnormal. Some others [8] detect anomalies using per-pixel reconstruction errors or by evaluating the density obtained from the model’s probability distribution. Some recent studies [5, 6] have shown that the knowledge distillation framework is effective for anomaly detection tasks. The main idea is to use a student network to distill knowledge from a pre-trained expert teacher network, i.e., to make the feature maps of certain layers of the two networks as equal as possible for the same input at training and anomalies are detected when the feature maps of certain layers of the two networks are very different at testing.

In this paper, we present a novel knowledge distillation-based approach (RKDAD) for anomaly detection. We use the “FSP matrix” as the distilled knowledge instead of the direct activation values of critical layers, the “FSP matrix” is defined as the inner product between features from two layers, i.e., the Gram matrix between two feature maps. By minimizing the L2 distance between the teacher’s and the student’s “FSP matrix” at training, the student network can learn the flow process of the normal sample features in the teacher network. We detect anomalies using the discrepancy between teacher’s and student’s corresponding “FSP matrix” at testing.

2 Related Work

Many previous studies have explored the anomaly detection tasks for images. In this section, we will provide a brief overview of the related works on anomaly detection tasks for images. We mainly introduce methods based on Convolutional Autoencoders (CAE), Generative Adversarial Networks (GAN) and methods based on knowledge distillation (KD).

2.1 CAE-Based Methods

The methods based on AE mainly use the idea that by learning the latent features of the normal samples, the reconstruction of abnormal inputs is not as precise as the normal ones, which results in larger reconstruction errors for abnormal samples, i.e., training on the normal data, the AE is expected to produce higher reconstruction error for the abnormal inputs than the normal ones and the reconstruction error can be defined as the L2 distance between the input and the reconstructed image. Abati et al. [7] proposed LSA [7] to train an autoregressive model at its latent space to better learn the normal latent features. MemAE [8] proposed the memory module to force the reconstructed image to look like the normal image, this mechanism increases the reconstruction error of abnormal images. CAVGA [9] proposed an attention-guided convolution adversarial variational autoencoder, which combines VAE with GAN to learn normal attention and abnormal attention in an end-to-end manner.

2.2 GAN-Based Methods

GAN-based methods attempt to find a specific latent feature space by training on normal samples. Then during testing, anomalies are detected based on the reconstruction or the

feature error. AnoGAN [10] is the first research to use GAN for anomaly detection. Its main idea is to let the generator learn the distribution of normal images through adversarial training. When testing, the L1 distance between the generated image and the input image and the feature error will be combined to detect anomalies. GANomaly [11] proposed the Encoder1-Decoder-Encoder2 architecture with a discriminator. What is used to detect anomalies is not the difference between the input image and the reconstructed image but the difference between the features of the two encoders. Skip-GANomaly [12] improves the generator part and uses the U-net [13] architecture that is with stronger reconstruction capabilities.

2.3 KD-Based Methods

Recently, KD-based methods for anomaly detection have drawn attention as it has shown the outstanding performance. Bergmann et al. [5] proposed Uniformed Students [5] which is the first anomaly detection method based on knowledge distillation. In this method, several student networks are trained to regress the output of a descriptive teacher network that was pretrained on a large dataset. Anomalies are detected when the outputs of the student networks differ from that of the teacher network, and the intrinsic uncertainty in the student networks is used as an additional scoring function that indicates anomalies. Salehi et al. [6] proposed to use the “distillation” of features at various layers of the pre-trained teacher network into a simpler student network to tackle anomaly detection problem. The Uniformed Students solely utilizes the last layer activation values in distillation, the second method mentioned above shows that considering multiple layers’ activation values leads to better exploiting the teacher network’s knowledge and more distinctive discrepancy. However, the methods mentioned above only consider the direct activation values as the knowledge of distillation without considering the relations between layers, which is more representative of the essential characteristics of the normal samples.

3 Method

In this section, we will first introduce the details of the gram matrix and show how to use the Gram matrix to define the “FSP matrix”. Then, we will introduce our approach to use the “FSP matrix” from two adjacent layers as the “distilled knowledge” to solve unsupervised anomaly detection tasks.

3.1 Gram Matrix and the “FSP Matrix”

As Eq. (1) shows, the matrix composed of the inner product of any k vectors in n -dimensional Euclidean space is defined as the Gram matrix of the k vectors. Obviously, Gram matrix of k vectors is a symmetric matrix. Gram matrix is often used in style transfer tasks, specifically, the feature map of the content image in a certain layer will be flattened into a one-dimensional feature vector according to the channel, and then a gram matrix composed of C vectors can be obtained, C is the number of channels in the feature map. Use the same operation to calculate the Gram matrix of the style image,

then minimize the distance of the Gram matrix of the two images. The Gram matrix is used to measure the difference in the style of two images. If the distance between the Gram matrix of the feature vectors of the two images is very small, it can be determined that the styles of the two images are similar. Essentially, the Gram matrix can be regarded as the eccentric covariance matrix between feature vectors, i.e., the diagonal elements reflect the information of the different feature vectors themselves, that is, the intensity information of the feature vectors, and the off-diagonal elements provide correlation information between different feature vectors.

$$\Delta(\alpha_1, \alpha_2, \dots, \alpha_k) = \begin{pmatrix} (\alpha_1, \alpha_1) & (\alpha_1, \alpha_2) & \dots & (\alpha_1, \alpha_k) \\ (\alpha_2, \alpha_1) & (\alpha_2, \alpha_2) & \dots & (\alpha_2, \alpha_k) \\ \dots & \dots & \dots & \dots \\ (\alpha_k, \alpha_1) & (\alpha_k, \alpha_2) & \dots & (\alpha_k, \alpha_k) \end{pmatrix} \quad (1)$$

The ‘‘FSP matrix’’ [14] proposed by Yim [14] et al. As Fig. 1 shows, the FSP matrix is generated by the features from two layers instead of being generated by the feature map of a single layer like the standard Gram matrix. By computing the inner product which represents the direction, to generate the FSP matrix, the flow between two layers can be represented by the FSP matrix.

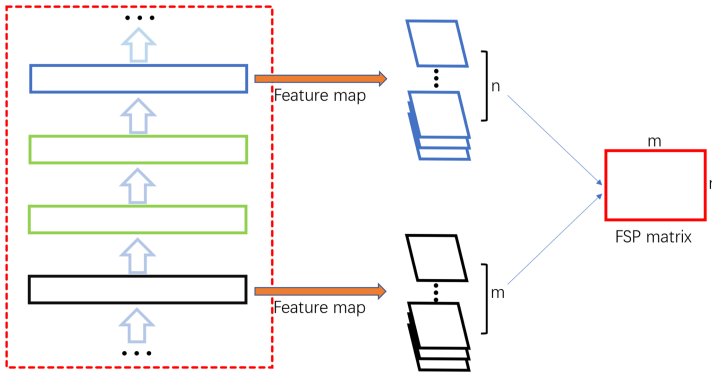


Fig. 1. The generation process of FSP matrix between different layers.

The calculation process of the FSP matrix is shown in Eq. (2), where F^1 and F^2 are two feature maps from different layers, h and w are the height and width of the feature map respectively, i and j are the channel indexes of the two feature maps, x and W are the input image and network parameters. It can be seen from Eq. (2) that the premise of calculating the FSP matrix is that the height and width of the two feature maps are equal.

$$FSP_{i,j}(x; W) = \sum_{s=1}^h \sum_{t=1}^w \frac{F^1_{s,t,i}(x; W) \times F^2_{s,t,j}(x; W)}{h \times w} \quad (2)$$

By letting the student network learn the flow of normal samples’ knowledge in the teacher network during training, if the input is normal sample during testing, the flow

of the teacher and the student network will be similar, while for the abnormal input, the flow will be very different.

3.2 The Proposed Approach

Given a training dataset $D_{train} = \{x_1, \dots, x_n\}$ containing only normal images, we train a student network with the help of a pre-trained teacher network, and the teacher network remains the same throughout the training process. Given a test dataset D_{test} , we utilize the discrepancy of the “FSP matrix” between the teacher and the student network to detect anomalies during the test, therefore, the student network must be trained to mimic the behavior of the teacher network, i.e., the student network should learn the “FSP matrix” of the teacher network during training. Earlier KD-based works for anomaly detection such as [5], which strive to teach just the activation values of the final layer of the teacher to the student, and in [6] they encourage the student to learn the teacher’s knowledge on normal samples through conforming its intermediate representations in a number of critical layers to the teacher’s representations. Because the feature maps of different layers of the neural network correspond to different levels of abstraction, mimicking different layers leads to a more thorough understanding of normal data than using only the final layer. In the methods mentioned above, the knowledge taught by the teacher to the student is the direct activation values of the critical layers, considering that a real teacher teaches a student the flow for how to solve a problem, in [14], Yim et al. proposed to define high-level distilled knowledge as the flow for solving a problem. Because the Gram matrix can be generated by computing the inner product of feature vectors, it contains the directionality between features, the flow can be represented by using Gram matrix consisting of the inner products between features from two layers. As Fig. 1 shows, the Gram matrix across layers is defined as the “FSP matrix”, i.e., the “relational knowledge” in our approach RKDAD, which is the higher-level knowledge for the normal images than the activation values of critical layers. We encourage the student to learn this higher-level knowledge of the teacher when training with normal image samples, during the test, if the input is an abnormal image the “FSP matrix” of the teacher network and the student network will be very different. Compared with the activation values of critical layers, the “FSP matrix” is more difficult to learn but it is also more discriminative for normal and abnormal samples.

In what follows, we refer to the “FSP matrix” between the i -th and the $(i + 1)$ -th layers in the networks as $F_{i,i+1}$, and the “FSP matrix” between the two adjacent layers of the teacher network as $F_{i,i+1}^t$ and the student’s one as $F_{i,i+1}^s$. The feature maps of the i -th layer and the $(i + 1)$ -th layer should have the same resolution. The loss function of our approach can be defined as Eq. (3)

$$L_{FSP} = \frac{1}{N} \sum_{i=1}^N \lambda_i \times \left\| F_{i,i+1}^t - F_{i,i+1}^s \right\|_2^2 \quad (3)$$

where N represents the number of all convolutional layers in the network (The convolutional layer refers to a module with a convolution operator, an activation function and an optional pooling operation), and λ_i is used to control the number of “FSP matrix” in the loss function, i.e., when λ_i is equal to 0, the two adjacent convolutional layers starting from the i -th convolutional layer are not used.

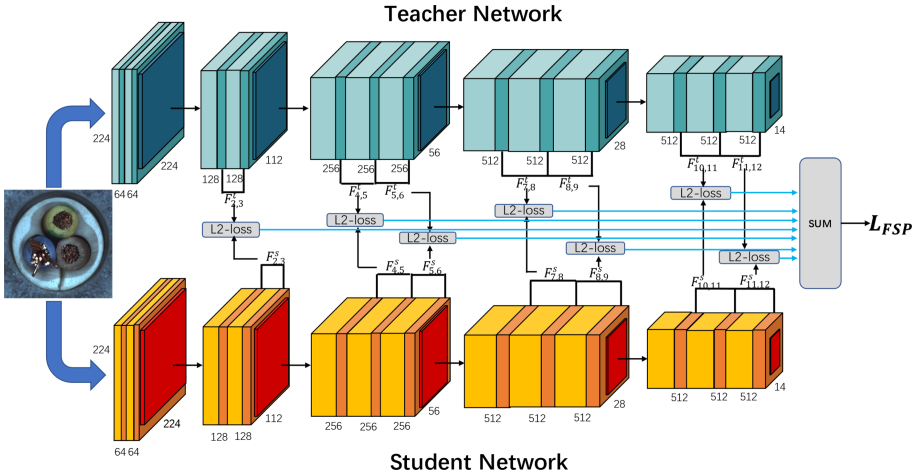


Fig. 2. Complete architecture of our proposed method. The student network shares the same structure with the teacher network pre-trained on a large dataset and learns the “FSP matrix” between two adjacent layers on normal data from the teacher. The discrepancy of their “FSP matrix” is formulated by a loss function and used to detect anomalies test time.

It should be noted that the teacher network and the student network share the same network architecture, While the teacher network should be deep and wide enough to learn all necessary features to perform well on a large-scale dataset, like ImageNet [15], and the teacher network should be pre-trained on a large-scale dataset. The goal of the student is to acquire the teacher’s knowledge of “FSP matrix” of the normal data.

Anomaly Detection: To detect anomalous images, each image in D_{test} is fed to both the teacher and the student network, i.e., we need two forward passes for anomaly detection. L_{FSP} , the loss function of RKDAD, is also the final anomaly score. As the student only learned the knowledge of “FSP matrix” of the normal data from the teacher, when the input is an abnormal sample, the “FSP matrix” of the teacher and the student will be very different, the anomaly score can be thresholded for anomaly detection.

4 Experiments

In this section, we have done extensive experiments to verify the effectiveness of our method. We will first introduce the implementation details of our approach, and then we introduce the datasets used. At last, we will show the anomaly detection results on the datasets introduced. Specially, we report an average result sampled every 10 consecutive epochs instead of reporting the maximum achieved result in many methods. The average result is a better measure of a model’s performance.

4.1 Implementation Details

VGG [16] has shown outstanding performance on classification tasks. In our approach, we choose the VGG-16 pre-trained on ImageNet as the teacher network, and a randomly

initialized VGG-16 as the student network. Of course, there are many other excellent network structures that can be used. However, it is required that the feature map resolution of some adjacent layers in the network structure used is the same, so that the FSP matrix can be calculated. Similar to [6], we avoid using bias terms in the student network. The model architecture of our approach is described in Fig. 2. We add 7 pairs of FSP matrix to the loss function in total, the loss function is also the anomaly score that is ultimately used to detect anomalies.

In all experiments, we use Adam [17] for optimization. The learning rate is set to be 0.001 and the batch size is 64. We only use normal images which are fed to both the teacher and the student network, and the parameter weights of the teacher network remain unchanged while that of the student network are updated during the training process, i.e., there are only forward propagation in the teacher network, and both forward and back propagation in the student network. Because the FSP matrix is more difficult to learn during the training process than the direct activation value of feature maps for the student network, it is required to train many epochs, such as 1000 or more.



Fig. 3. Object (cable, toothbrush, capsule) and texture (wood, grid, leather) images in MVTEcAD [2]. The images in the upper row are normal samples and the ones in the lower row are abnormal samples.

4.2 Datasets

We verified the effectiveness of our method on three datasets as follows.

MNIST [18]: a handwritten digit images dataset, which consists of 60k training and 10k test 28×28 Gy-scale images, includes numbers 0 to 9.

Fashion-MNIST [19]: a more complex image dataset proposed to replace MNIST, which covers images of 70 k different products in 10 categories, such as T-shirt, dress and coat etc.

MVTEcAD [2]: a dataset dedicated to anomaly detection with more than 5 k images, which includes 5 categories of texture images and 10 categories of object images. For each category, the dataset train contains only normal images, and the test sets contain a variety of abnormal images and some normal images. In our experiment, the images will be scaled to the size 128×128 . Some object and texture samples are shown in Fig. 3.

Note that for MNIST and Fashion-MNIST, we regard one class as normal and others as anomaly at training while at testing the whole test set is used. For MVTEC-AD, the datasets train and test are used.

4.3 Results

We use the area under the receiver operating characteristic curve (AUROC) for evaluation. The results are shown in Table 1, 2 for MNIST, Fashion-MNIST and Table 3, 4, 5 for MVTEC-AD. Note that the AUROC values in the tables are the average of 10 consecutive epochs, not the maximum value. We have compared our method with many approaches, and the Tables show that our method can achieve competitive results with to state-of-the-art method on the datasets we used.

The Results on MNIST and Fashion-MNIST. As Table 1 and Table 2 show, our method RKDAD can achieve competitive results on mnist and fashion-mnist compared to the state-of-the-art methods with using only the sum of 12 distances of the “FSP matrix” as the loss function and the anomaly score, which verifies the effectiveness of our relation-based knowledge distillation framework for anomaly detection task.

Table 1. AUROC in % on MNIST.

Method	0	1	2	3	4	5	6	7	8	9	Mean
ARAE [20]	99.8	99.9	96.0	97.2	97.0	97.4	99.5	96.9	92.4	98.5	97.5
OCSVM [21]	99.5	99.9	92.6	93.6	96.7	95.5	98.7	96.6	90.3	96.2	96.0
AnoGAN [10]	96.6	99.2	85.0	88.7	89.4	88.3	94.7	93.5	84.9	92.4	91.3
DSVDD [22]	98.0	99.7	91.7	91.9	94.9	88.5	98.3	94.6	93.9	96.5	94.8
CapsNetpp [23]	99.8	99.0	98.4	97.6	93.5	97.0	94.2	98.7	99.3	99.0	97.7
OCGAN [24]	99.8	99.9	94.2	96.3	97.5	98.0	99.1	98.1	93.9	98.1	97.5
LSA [7]	99.3	99.9	95.9	96.6	95.6	96.4	99.4	98.0	95.3	98.1	97.5
CAVGA [9]	99.4	99.7	98.9	98.3	97.7	96.8	98.8	98.6	98.8	99.1	98.6
U-Std [5]	99.9	99.9	99.0	99.3	99.2	99.3	99.7	99.5	98.6	99.1	99.35
Mul-KD [6]	99.8	99.8	97.8	98.7	98.4	98.2	99.4	98.4	98.4	98.1	98.71
OURS	99.91	99.96	98.74	99.41	98.32	98.18	99.12	98.97	97.99	98.34	98.89

The Results on MVTEC-AD. Note that Table 3 shows the AUROC results of 10 categories of object images in MVTEC-AD, and it can be seen that the performance of the proposed approach RKDAD can surpass other methods to obtain the state-of-the-art results. Table 4 shows the results of 5 categories of texture images and Table 5 shows the average performance across all categories in MVTEC-AD. It can be seen from Table 4 that our method is not very good for the texture images but it still exceeds the performance of many methods. Table 5 shows that the performance of our method on MVTEC-AD is second only to the state-of-the-art method, and is significantly better than other methods.

Table 2. AUROC in % on Fashion-MNIST.

Method	0	1	2	3	4	5	6	7	8	9	Mean
ARAE [20]	93.7	99.1	91.1	94.4	92.3	91.4	83.6	98.9	93.9	97.9	93.6
OCSVM [21]	91.9	99.0	89.4	94.2	90.7	91.8	83.3	98.8	90.3	98.2	92.8
DAGMM [25]	30.3	31.1	47.5	48.1	49.9	41.3	42.0	37.4	51.8	37.8	41.7
DSEBM [26]	89.1	56.0	86.1	90.3	88.4	85.9	78.2	98.1	86.5	96.7	85.5
DSVDD [22]	98.2	90.3	90.7	94.2	89.4	91.8	83.4	98.8	91.9	99.0	92.8
LSA [7]	91.6	98.3	87.8	92.3	89.7	90.7	84.1	97.7	91.0	98.4	92.2
Mul-KD [6]	92.5	99.2	92.5	93.8	92.9	98.2	84.9	99.0	94.3	97.5	94.5
OURS	90.72	99.35	90.79	94.91	91.84	94.78	85.81	99.16	90.57	96.14	93.41

All experimental results show that although the model architecture of RKDAD is very simple, it still achieves excellent results, which proves that the proposed relation-based knowledge distillation framework for anomaly detection has great potential.

Table 3. AUROC in % on object images in MVTecAD. We surpass the SOTA method.

Method	Bottle	Hazelnut	Capsule	Metal Nut	Pill	Cable	Transistor	Toothbrush	Screw	Zipper	Mean
AVID [27]	88	86	85	63	86	64	58	73	66	84	75.3
AESSIM [28]	88	54	61	54	60	61	52	74	51	80	63.5
AEL2 [28]	80	88	62	73	62	56	71	98	69	80	73.9
AnoGAN [10]	69	50	58	50	62	53	67	57	35	59	56.0
LSA [7]	86	80	71	67	85	61	50	89	75	88	75.2
CAVGA [9]	89	84	83	67	88	63	73	91	77	87	80.2
DSVDD [22]	86	71	69	75	77	71	65	70	64	74	72.2
VAE-grad [29]	86	74	86	78	80	56	70	89	71	67	75.7

(continued)

Table 3. (continued)

Method	Bottle	Hazelnut	Capsule	Metal Nut	Pill	Cable	Transistor	Toothbrush	Screw	Zipper	Mean
GT [30]	74.29	33.32	67.79	82.37	65.16	84.70	79.79	94.00	44.58	87.44	71.34
Mul-KD [6]	99.39	98.37	80.46	73.58	82.70	89.19	85.55	92.17	83.31	93.24	87.80
OURS	99.05	96.82	73.08	82.79	78.70	92.71	87.75	88.33	94.67	89.97	88.39

Table 4. AUROC in % on texture images in MVTEcAD.

Method	Leather	Wood	Carpet	Tile	Grid	Mean
AVID [27]	58	83	73	66	59	67.8
AESSIM [28]	46	83	67	52	69	63.4
AEL2 [28]	44	74	50	77	78	64.6
AnoGAN [10]	52	68	49	51	51	54.2
LSA [7]	70	75	74	70	54	68.6
CAVGA [9]	71	85	73	70	75	74.8
DSVDD [22]	73	87	54	81	59	70.8
VAE-grad [29]	71	89	67	81	83	78.2
GT [30]	82.51	48.24	45.90	53.86	61.91	58.49
Mul-KD [6]	73.58	94.29	79.25	91.57	78.01	83.34
OURS	60.90	92.37	77.71	78.79	66.42	75.24

Table 5. Mean AUROC in % on MVTEcAD.

Method	AVID	AESSIM	AEL2	AnoGAN	LSA	CAVGA	DSVDD	VAE-grad	GT	Mul-KD	OURS
Mean	73	63	71	55	73	78	72	77	67.06	87.87	84.00

5 Conclusion

In this paper, we have presented a novel knowledge distillation-based approach for anomaly detection (RKDAD). We have further explored the possibility of anomaly detection using the “relational knowledge” between different layers when the knowledge of normal samples flows in the network, and we show that using the “distillation” of the “FSP matrix” from adjacent layers of a teacher network, pre-trained on ImageNet, into a student network which has the same structure as the teacher network, and then using the discrepancy between teacher’s and student’s corresponding “FSP matrix” at testing can achieve competitive results compared to the state-of-the-art methods. We have verified the effectiveness of our method on many datasets. In this paper, we only consider the

“FSP matrix” from adjacent layers as the “relational knowledge”, more forms of “relational knowledge” can be explored to improve the performance of anomaly detection task in the future.

References

1. Shuang, M., Yudan, W., Guojun, W.: Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model. *Sensors* **18**(4), 1064 (2018)
2. Bergmann P., Fauser M., Sattlegger D., et al.: MVTEC AD — a Comprehensive Real-World dataset for unsupervised anomaly detection. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9592–9600 (2019)
3. Zhe L., et al.: Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8290–8299 (2018)
4. Zhou, J.T., Du, J., Zhu, H., et al.: AnomalyNet: an anomaly detection network for video surveillance. *IEEE Trans. Inf. Forensics Secur.* **14**(10), 2537–2550 (2019)
5. Bergmann P., Fauser M., Sattlegger D., et al.: Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4182–4191 (2020)
6. Salehi M., Sadjadi N., Baselizadeh S., et al.: Multiresolution knowledge distillation for anomaly detection. arXiv preprint arXiv: 2011.11108 (2020)
7. Abati D., Porrello A., Calderara S., Cucchiara R.: Latent space autoregression for novelty detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 481–490 (2019)
8. Gong D., Liu L., Le V., et al.: Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1705–1714 (2019)
9. Venkataramanan, S., Peng, K.-C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly localization in images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 485–503. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_29
10. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
11. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: semi-supervised anomaly detection via adversarial training. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 622–637. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_39
12. Akay, S., Atapour-Abarghouei, A., Breckon, T.P.: Skip-GANomaly: skip connected and adversarially trained encoder-decoder anomaly detection. In: IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2019)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Yim, J., Joo, D., Bae, J., et al.: A Gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7130–7138 (2017)

15. Jia, D., Wei, D., Richard, S., Li-Jia, L., Kai, L., Fei-Fei L.: Imagenet: a large-scale hierarchical imagedatabase. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
16. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks For Large-Scale Image Recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
17. Diederik, P.K., Jimmy, B.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
18. LeCun, Y., Cortes, C., et al. <http://yann.lecun.com/exdb/mnist>. Accessed 05 April 2021
19. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset For Benchmarking Machine Learning Algorithms. arXiv preprint [arXiv:1708.07747](https://arxiv.org/abs/1708.07747) (2017)
20. Salehi, M., Arya, A., Pajoum, B., et al.: Arae: Adversarially robust training of autoencoders improves novelty detection. arXiv preprint [arXiv:2003.05669](https://arxiv.org/abs/2003.05669) (2020)
21. Chen, Y., Xiang, S.Z., Huang, T.S.: One-class svm for learning in image retrieval. In: Proceedings 2001 International Conference on Image Processing, pp. 34–37 (2001)
22. Ruff, L., Vandermeulen, R.A., et al.: Deep one-class classification. In: International Conference on Machine Learning, pp. 4393–4402 (2018)
23. Li, X., Kiringa, I., Yeap, T., et al.: Exploring deep anomaly detection methods based on capsule net. In: ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning, pp. 375–387 (2020)
24. Perera, P., Nallapati, R., Bing, X.: Ocgan: one-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2898–2906 (2019)
25. Zong, B., Song, Q., Martin Renqiang, M., et al.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations (2018)
26. Shuangfei, Z., Yu, C., Weining, L., Zhongfei, Z.: Deep structured energy-based models for anomaly detection. In: International Conference on Machine Learning, pp. 1100–1109 (2016)
27. Sabokrou, M., Pourreza, M., Fayyaz, M., et al.: Avid: adversarial visual irregularity detection. In: Asian Conference on Computer Vision, pp. 488–505 (2018)
28. Bergmann, P., Lwe, S., Fauser, M., et al.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), pp. 372–180 (2019)
29. Dehaene, D., Frigo, O., Combrexelle, S., et al.: Iterative energy-based projection on a normal data manifold for anomaly localization. In: International Conference on Learning Representations (2020)
30. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: Advances in Neural Information Processing Systems, pp. 9758–9769 (2018)