Unsupervised Deep Learning-based End-to-end Network for Anomaly Detection and Localization

Bekhzod Olimov
Computer Science and Engineering Department
Kyungpook National University
Daegu, South Korea
bekhzod.olimov@knu.ac.kr

Barathi Subramanian
Computer Science and Engineering Department
Kyungpook National University
Daegu, South Korea
achu_samriti@yahoo.com

Jeonghong Kim
Computer Science and Engineering Department
Kyungpook National University
Daegu, South Korea
jhk@knu.ac.kr

Abstract — These days there is great demand for automatizing a visual inspection process in industrial companies since it is a tedious and time-consuming task. Recent progress in deep convolutional neural networks allowed to automatize visual inspection procedure. However, currently available supervised learning methods require large amount of labeled data, while the unsupervised learning techniques suffer from lack of accuracy. To address these problems, we propose a deep learning-based unsupervised learning method that exhibits fast and precise performance. The proposed unsupervised learning method based pseudo-labeling algorithm using graph Laplacian matrix that allows transferring computationally expensive autoencoder problem to classification task, the proposed system benefits from very fast convergence ability and significantly outperforms currently available deep learning-based AVI methods. In the conducted experiments using real-life fabric image datasets, the proposed method outperformed the currently available methods in terms of speed and accuracy.

Keywords—deep convolutional neural networks; fabric defect detection; industrial quality inspection; unsupervised learning;

I. Introduction

In manufacturing, the process of visual inspection is closely associated with quality of a product and prosperity of an enterprise since fast and accurate exploration of abnormal products eliminates the issue of visually qualitative defects. Consequently, it leads to customer satisfaction and high purchase rate by consumers. Traditional surface inspection methods use human labor to detect defected items. However, they are monotonous and tiresome activities. Moreover, they are prone to wrong decisions due to human-related characteristics, such as lack of concentration, optical illusion, subjective assessment, and inclination to exhaustion. Additionally, human inspectors require to obtain training and practice that demand certain amount of time. Employee turnover is another disadvantage of the manual labor. Thus, there has been a high inclination to equip manufacturing plants with embedded sensors to continuously monitor the equipment proper operation [1].

Due to the relevance of fabric AVI (Visual Inspection Process) techniques, there has been extensive research conducted on this topic using computer vision-related applications. Broadly, existing deep learning (DL)-based techniques can be categorized into supervised semi-supervised, and unsupervised approaches [2]. In general, the existing DL-based automated visual inspection (AVI) approaches have several limitations. Specifically, supervised learning approaches suffer from data scarcity and lack of correctly labeled training data [3, 4, 5]. Moreover, semi-supervised learning methods experience computational complexity and lack of generalizability [6, 7, 8, 9], while unsupervised learning techniques require enormous training and inference time [10, 11, 12]. All these factors negatively impact on the performance of AVI, which is supposed to be fast and precise.

Considering the aforementioned limitations of the currently available approaches, we developed a compound end-to-end model to detect anomalous fabric items based on unsupervised learning. There are three distinct stages in the proposed models pseudo-labeling of data with no annotations using an unsupervised algorithm, training of deep convolutional neural network (DCNN) to solve a binary classification problem by detecting abnormal items. In fact, the proposed method addresses the aforementioned problems of data shortage, lack of labeled instances, and computational complexity. Moreover, it is significantly faster than the existing DL-based approaches. In general, the proposed method has the following contributions.

- The proposed method employs an unsupervised clustering-based algorithm to distinguish the original data into normal and abnormal images and pseudo-label the data based on the clustering algorithm results; therefore, it requires no labeled data, which deals with the problem of data annotation.

- The proposed method substitutes anomaly detection (AD) solutions from autoencoders (AEs) with encoding and decoding parts and expansion parts with a simple binary image classification task. Also, it benefits from transfer learning, which makes it significantly more resource and memory intensive as well as fast for training and inference.

The rest of the manuscript has the following structure. Section II comprises detailed explanation of the proposed methodology. Section III provides comprehensive information on the conducted experiments and their outcomes. Finally, Section IV concludes this study and outlines future research directions.

II. Related Work

As discussed in Section I, there has been proposed great number of methods to improve AVI systems so far. The currently available approaches can be classified into supervised, semi-supervised, and unsupervised techniques.

Supervised learning approaches using DL methods require availability of annotated datapoints. They learn to classify the data into normal and anomaly instances in a training stage and use the learned model to distinguish items in an inference phase. For example, Erfani et al. presented a supervised learning model that uses instances drawn from similar distributions and intrinsic structures of labeled data to identify landmarks that are supposed to be similar from different sources [3]. Liu et al. proposed an approach to establish the link between the heuristic unmasking procedure and multiple classifiers in statistical machine learning [4]. Despite providing accurate detection of anomaly points, the existing methods require a large amount of training time as input data become more complex. In addition, owing to the data-driven characteristics of DL models and shortage of real annotated data in the real-world, supervised learning approaches for AD are exploited less frequently compared with their semi-supervised and unsupervised counterparts [5].

Semi-supervised techniques require a single target label. These methods learn to distinguish normal and anomaly points based on the feature representation of the raw input data obtained from hidden layers of DCNN models. For example, Perera et al. have proposed a DL-based one-class transfer learning approach with two different loss functions[6]. Specifically, this method trains a CNN model based on compactness and descriptiveness loss functions and uses template matching during inference. Napoletano et al. have presented a method to assess the anomaly degree for each region of an image by computing DCNN-based visual similarity with respect to fault-free image feature representations in the training data [7]. It predicts anomaly images based on computation from convolutional neural network (CNN)-based similarity with respect to the trained model. Wang et al. have proposed abnormal event detection framework that is composed of a principal component analysis network and kernel principal component analysis to address the problem of anomaly detection [8]. Moreover, in[9] distribution-augmented contrastive learning extending training distributions via data augmentation has been proposed to obstruct the uniformity of contrastive representations. In[13], a method that leverages the descriptiveness of extracted features by CNNs to evaluate the density using normalizing flows and employ a multi-scale feature extractor to obtain better performance on high dimensionality of industrial images. Li et al. have proposed a two-step network to build anomaly detectors without using anomaly images [14]. First, the deep representations are learned using CutPaste data augmentation methods and then a generative one-class classifier is built based on the learned representations from the first stage. Although these techniques addressed the problem of labeled data shortage for training DCNN models, they inherit computational complexity from supervised learning methods. Moreover, these models are more likely to experience overfitting to normal instances due to a lack of training with anomaly data points.

DL-based unsupervised learning methods are widely exploited for AD due to the ability to generate powerful feature representations from unlabeled data. The most common form of unsupervised learning technique is AEs [15, 16, 17]. AE focuses on generating decision boundaries to separate normal data points from anomalous ones based on the feature representations obtained from training DCNN models. Haselmann et al. have proposed an AE model trained on image patches of normal samples, where central regions were cut out[10]. The model could detect anomalous images since it was trained using pixel-wise reconstruction loss on totally fault-free images. To address the issues with slight localization inaccuracies Begmann et al. have developed a model using a perceptual loss function examining local image region internal dependencies considering luminance, contrast, and structural information in contrast with a simple pixel-wise comparison loss functions employed in the existing methods[11]. Bergmann et al. have introduced a powerful student-teacher network for precise anomaly segmentation in industrial images using intrinsic uncertainty in the student networks as an extra scoring function that shows anomalies[12]. Although these models are commonly used in AVI systems, they suffer from enormous time for training and inference.

III. Proposed Methodology

In this section, we describe the proposed model. Fig. 1 shows a general overview of the proposed model. The model has two essential steps for both training and inference phases. Specifically, in training, input data pass through pre-processing, pseudo-labeling, and training process stages.

A. Pre-Processing Step

In the pre-processing stage, unlabeled fabric images experience integration stage, which means that raw data from different sources are unified under identical requirements to provide a smooth processing. Then, the images are extracted from directories and represented as tensors since they ensure more natural representations of multidimensional data. The resulting tensor is $4\mathrm{D} - \mathrm{X} \in \mathbb{R}^{M \times C \times H \times W}$, where *M, C, H,* and *W* are the total number of images, number of channels, image height, and image width, respectively.

After obtaining the images in tensors, they are resized to match the input size of DCNN later in a training process phase. Finally, the resized image pixel values are standardized to follow a standard normal distribution. In (1), X and X_{std} are original and standardized data; while i and M are a particular data point and the total number of instances, respectively.

$$X_{std} = \frac{X - \frac{1}{M}\sum_{i=1}^{M} x_i}{\sqrt{\frac{1}{M}\sum_{i=1}^{M}(x_i - \frac{1}{M}\sum_{i=1}^{M} x_i)^2}} \qquad (1)$$

B. Pseudo-Labeling Step

Considering that the training data have no annotated labels, we should bipartition it into normal and defective images. To attain this goal, we use the power of an unsupervised learning technique using the eigenvalues (EVL) and eigenvector (EVT) of a graph Laplacian matrix (GLM). Specifically, we construct adjacency (M_{adj}) and degree (M_{deg}) matrices that contain distances between samples and the sum of weights from

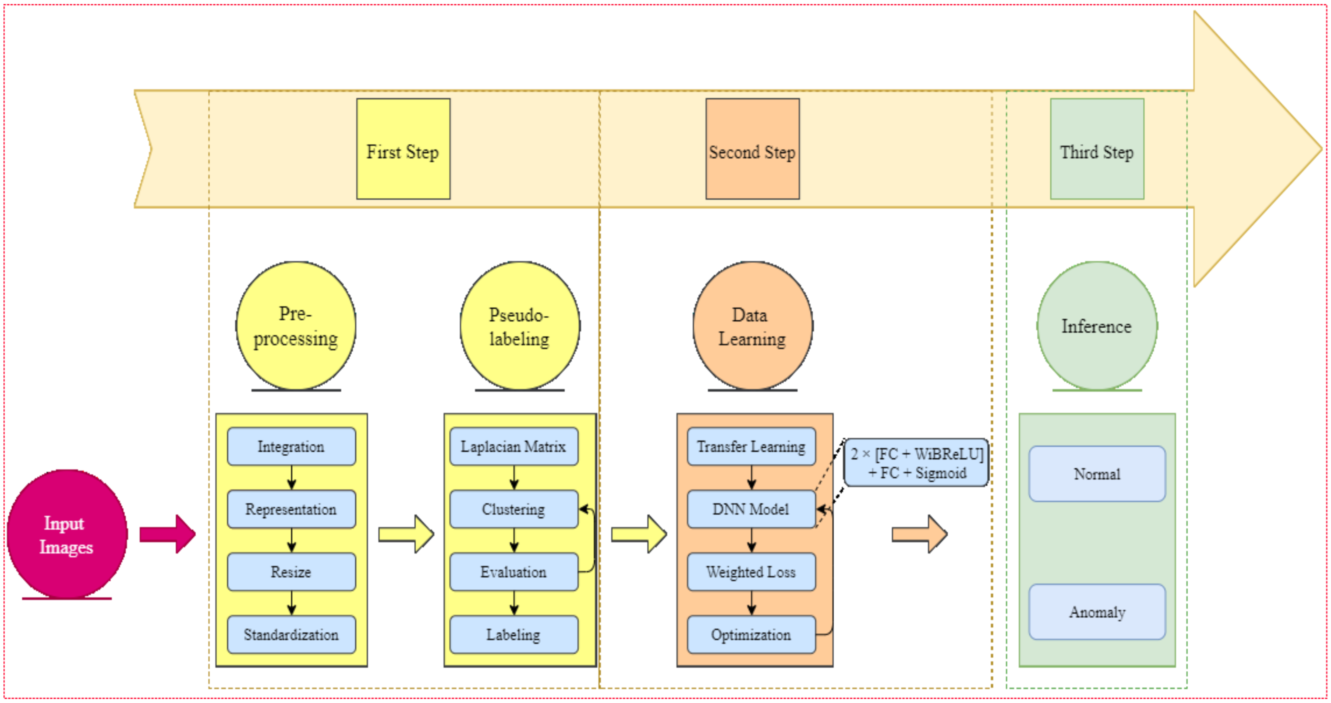


Figure 1: Graphical illustration of the proposed methodology. DNN stands for deep neural network.

instances, respectively. M_{GL} is the difference of the formulated matrices M_{deg} and M_{adj} .

In the last step, we obtain the largest EVL and its corresponding EVT of M_{GL}, which are equal to 0 and a constant value, respectively.

Then, the largest EVL and its corresponding EVT are stacked into a 1D vector and divided into two parts based on a certain threshold value. The threshold value is selected to be 0. This algorithm bi-partitions the input data into two groups, normal and defective samples. Then, the obtained clustering results are assessed using unsupervised learning evaluation metrics, such as the Silhouette coefficient. Once satisfactory scores are obtained from the evaluation metrics, the aforementioned sample groups are pseudo-labeled as positive and negative data points, respectively.

The output of the pseudo-labeling process cannot be employed in real-time inference; Instead, it provides annotations for the unlabeled data and ensures training a binary classification model in as a supervised learning method. Therefore, the output of this phase can be employed for training DCNN in the next step of the proposed methodology.

C. Data Learning Step

Afterward, DCNN is trained to solve a binary classification task and categorize them into normal and defective instances. We use power of transfer learning approach to acquire knowledge from DCNN trained on a significantly large database to solve a more sophisticated classification problem to obtain a fast, efficient, and accurate classification model. As a DCNN model, we select an 18-layer residual network model (ResNet-18) [18] pre-trained on the ImageNet dataset since it obtained the most optimal accuracy-time tradeoff compared with other popular DCNN models in the experiments. Precisely, we extract a vector of 512 values of the input features to the last fully connected (classification) layer of the ResNet-18 that exhibit complex and high-level representations of the ImageNet images and formulate a deep neural network model (DNNM) containing several fully connected layers. Specifically, the latent representation vector for the pre-trained model first is inputted into a fully connected (FC) layer and it outputs a 128D vector followed by the WIB-ReLU activation function [19]. This block of operations is repeated until the final FC layer with the sigmoid activation function outputs two values with probabilities for normal and defective items. Considering that the normal images significantly outnumbered abnormal ones, we use a weighted binary cross-entropy loss (L_{wbce}) function.

$$L_{wbce} = \frac{\sum_{i=1}^{M} \emptyset_{c_i} + \left(-i_c + \log\left(\sum_j e^{i_j}\right)\right)}{\sum_{i=1}^{M} \emptyset_{c_i}} \qquad (2)$$

In (2), *M, c,* and *Φ* are the total number of instances, a specific class, and weight value, respectively.

D. Inference Step

The trained model from the data learning step that contains the extracted features from the datasets, is used in the inference step to classify unseen data into normal and anomaly images.

IV. Experiments and Results

In this section, we represent comprehensive information about the conducted experiments and their results, as well as provide a comparison of the proposed method's experiment results with the ones of the existing state-of-the-art models.

A. Datasets Description

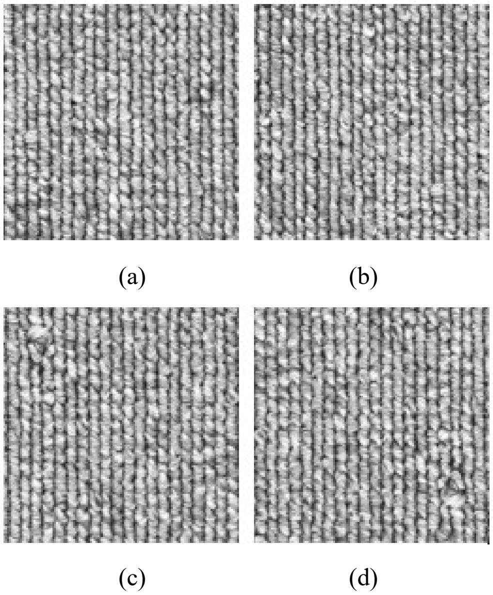
Although there are numerous open-source datasets for AD, we selected three real-life datasets containing normal and defective fabric/nanofabric material items. Two of the datasets were publicly available and open-source for research, namely NanoTWICE and MVTec AD datasets. The third dataset was provided by DWorld company located in Daegu, South Korea. Table 1 summarizes the description of the considered datasets.

TABLE I. Overall information on the datasets for the experiments

Dataset Name	**Image Type**	**Image Size**	**Number of Images**		
			Train	***Validation***	***Test***
NanoTWICE	Nanofabric	1024 × 700	30	5	10
MVTec AD*	Fabric	512 ×512	470	70	134
DWorld	Fabric	1024×1024	1750	250	495

* Carpet and tile classes only.

From Table 1, we divided the considered datasets in the ratio of 7 : 1 : 2 for training, validation, and test subsets to train, select hyperparameters, and test the generalizability of the proposed method, respectively. In addition, from Table 1, we can observe that the currently available open-source datasets contained very limited number of images. Specifically, NanoTWICE and MVTEC datasets comprised only 45 and 674 images, respectively. Because the existing datasets experience a shortage of images, we obtained a novel fabric materials AD dataset, called DWorld.


(a) (b)

(c) (d)

Fig 2. (a), (b) normal and (c), (d) defected images from DWorld dataset

The images were collected with a line-scan charge coupled device camera and exhibit various kinds of abnormalities. The DWorld dataset contains 2,495 unlabeled images each with 1024 × 1024 pixel size. Fig 2 shows normal and defected images from DWorld dataset.

B. Training Details

1) Experiment Settings: We formulated the baseline and proposed methods using Python version 3.6.9 and PyTorch library version 1.4.0. The weight parameters were initialized based on a standard normal distribution with a mean and standard deviation of 0 and 1, respectively, to meet the requirements of the WIB-ReLU activation function. All bias parameters were zero-initialized. We used binary crossentropy as the minimizing function and stochastic gradient descent as the parameter optimizer for the proposed method. The experiments were conducted using 32 GB NVIDIA Tesla V100-SXM2 GPU with CUDA 10.0 with a mini-batch size of eight for MVTec and DWorld and four for NanoTWICE datasets.

2) Evaluation Metrics: We exploit several evaluation metrics to assess the performance of the proposed model compared with the one of baseline methods from different angles. Specifically, we define accuracy score (AS), false positive rate (FPR), true positive rate or recall score (TPR or RS), precision score (PS), and F1 score (F1) for the evaluation of the models' performance.

3) We selected three existing methods as the baseline models: Same Same But DifferNet (SSBD) [13], learning deep features for one-class classification (LDFC)[6], and Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings (UNST) [12]. All baseline models and proposed method were trained under the same circumstances.

C. Experiment Results

We also checked the baseline and proposed models' generalization ability on the test set of the considered datasets and the results are provided in Table II. From the table II, the proposed model achieved the best performance in the AS, PS, F1, and AUC evaluation metrics on the NanoTWICE dataset by taking the least amount of time for inference. Regarding MVTec AD, the proposed model has the best performance in all of them, significantly outperforming its counterparts, which barely achieved 90% accuracy, except for SSBD, LDFC, which scores were somewhere at 95%. LDFC was also the fastest model to train and outperformed the proposed method in terms of speed. Finally, the proposed model largely outperformed the baseline models in the test set of the DWorld dataset. Specifically, our model achieved 99.6% for accuracy related metrics, such as AS and F1. Also, it was the fastest model in terms of training time by taking on average 5.530 seconds for an epoch.

D. Computational Complexity

Table III represents execution time requirements of the proposed method's steps in mean and standard deviation (STD) using the test sets of the considered datasets.

TABLE II. PERFORMANCE COMPARISON OF THE BASELINE AND PROPOSED METHODS ON THE TEST SET OF THE CONSIDERED DATASETS*

Dataset Name	Model Name	AS	PS	RS	F1	AUC	Time (s)
NT	SSBD	0.942	0.926	0.978	0.952	0.870	2.034
	LDFC	0.958	0.932	**0.976**	0.954	0.886	1.901
	UNST	0.819	0.822	0.842	0.832	0.809	3.335
	Ours	**0.984**	**0.971**	0.973	**0.972**	**0.899**	**1.224**
MVAD	SSBD	0.970	0.968	0.960	0.964	0.910	7.920
	LDFC	0.968	0.960	0.940	0.960	0.906	**5.240**
	UNST	0.892	0.865	0.835	0.850	0.843	12.827
	Ours	**0.990**	**0.992**	**0.990**	**0.991**	**0.913**	7.164
DW	SSBD	0.972	0.921	0.939	0.930	0.892	8.802
	LDFC	0.980	0.899	**0.997**	0.948	0.905	9.124
	UNST	0.902	0.945	0.901	0.923	0.878	17.092
	Ours	**0.996**	**0.997**	0.996	**0.996**	**0.946**	**5.530**

* Results of the experiments on 32 GB NVIDIA Tesla V100-SXM2 GPU

Observably, training and inference phase stages execution times vary significantly. Specifically, in training stage, training process took the longest time that was equal to 282.6 seconds. Pseudo-labeling phase was relatively faster by requiring in average 44.28 seconds. Regarding inference, the most time-consuming phases in training, such as pseudo-labeling and training process were not activated because inference time employed trained DCNN model from the training stage; therefore, inference phase was considerably faster than training counterpart by requiring approximately 8 seconds for completion. Considering that this short amount of time detected and illustrated defected regions of the hundreds of abnormal textile product items, the proposed method not only accurate but also efficient and fast enough for its employment in real-time textile manufacturing visual inspection applications. Notably, the values in Table III were computed based on the experimental results on the datasets from Section V-A and they may vary depending on the number and complexity of the images in different datasets.

TABLE III. AVERAGE TIME OF EXECUTION FOR EACH STAGE OF THE PROPOSED METHOD*

Step	**Mean (Training)**	**STD (Training)**	**Mean (Inference)**	**STD (Inference)**
Pre-Processing	0.973	0. 011	0. 418	0. 007
Pseudo-Labeling	44.28	0.782	0.000	0.000
Training Process	282.6	1.477	0.000	0.000
Anomaly Interpretation	0.000	0.000	7.639	0.215

* Results of the experiments on 32 GB NVIDIA Tesla V100-SXM2 GPU

V. CONCLUSION AND FUTURE WORK

This paper studied AVI system applications using deep learning-based methods. Considering the limitations of the existing methods, we developed an end-to-end unsupervised deep-learning-based image classification system to detect anomalies in fabric images. The proposed system comprised of two main stages, such as pseudo-labeling and data learning. The main contribution of the proposed method is that it does not require annotated data, which saves significant amount of time and money. Based on the experimental results obtained from training and inference on three fabric and nanofabric material databases, the proposed method performed considerably better than the currently available techniques in terms of accuracy and time. In the future, we will focus on further improving the proposed method by adding anomaly interpretation technique to automatically detect and visualize the defected part of the anomalous data.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A3043970). This study was also supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394).

REFERENCES

[1] Y. Liang, S. Wang, W. Li, X. Lu, Data-driven anomaly diagnosis for machining processes, Engineering 5 (4) (2019) 646-652.

[2] A. Castellani, S. Schmitt, S. Squartini, Real-world anomaly detection by using digital twin systems and weakly supervised learning, IEEE Transactions on Industrial Informatics 17 (7) (2020) 4733-4742.

[3] S. Erfani, M. Baktashmotlagh, M. Moshtaghi, V. Nguyen, C. Leckie, J. Bailey, K. Ramamohanarao, From shared subspaces to shared land-marks: A robust multi-source classification approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.

[4] Y. Liu, C.-L. Li, B. Poczos, Classifier two sample test for video anomaly detections., in: British Machine Vision Conference, 2018, p. 71.

[5] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: A survey, arXiv preprint arXiv:1901.03407 (2019).

[6] P. Perera, V. M. Patel, Learning deep features for one-class classification, IEEE Transactions on Image Processing 28 (11) (2019) 5450-5463.

[7] P. Napoletano, F. Piccoli, R. Schettini, Anomaly detection in nanofibrous materials by cnn-based self-similarity, Sensors 18 (1) (2018) 209.

[8] T. Wang, Z. Miao, Y. Chen, Y. Zhou, G. Shan, H. Snoussi, Aed-net: An abnormal event detection network, Engineering 5 (5) (2019) 930-939.

[9] K. Sohn, C.-L. Li, J. Yoon, M. Jin, T. Pfister, Learning and evaluating representations for deep one-class classification, in: International Conference on Learning Representations, 2021.

[10] M. Haselmann, D. P. Gruber, P. Tabatabai, Anomaly detection using deep learning based image completion, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 1237-1242.

[11] P. Bergmann, S. Lowe, M. Fauser, D. Sattlegger, C. Steger, Improving unsupervised defect segmentation by applying structural similarity to autoencoders, arXiv preprint arXiv:1807.02011 (2018).

[12] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings, in: Proceedings of the IEEE/CVF Conference on Computer Vi-sion and Pattern Recognition, 2020, pp. 4183-4192.

[13] M. Rudolph, B. Wandt, B. Rosenhahn, Same same but differnet: Semi-supervised defect detection with normalizing flows, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1907-1916.

[14] C.-L. Li, K. Sohn, J. Yoon, T. Pfister, Cutpaste: Self-supervised learning for anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9664-9674.

[15] B. Olimov, S. Karshiev, S. Din, A. Ahmad, A. Paul, J. Kim, FU-Net: fast biomedical image segmentation model based on bottleneck convolution layers, Multimedia Systems, 2021, pp. 637-650.

[16] B. Olimov, S. Koh, J. Kim, AEDCN-Net: Accurate and efficient convolutional neural network model for medical segmentation, IEEE Access (9), 2021, pp. 154194-154203.

[17] B. Olimov, A. Paul, J. Kim, REF-Net: robust, efficient, and fast network for semantic segmentation applications using devices with limited computational resources, IEEE Access (9), 2021, pp. 15084-15098.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[19] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, J. Kim, Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model, Concurrency and Computation: Practice and Experience (2020) e61