AN ANOMALY DETECTION METHOD BASED ON SELF-SUPERVISED LEARNING WITH SOFT LABEL ASSIGNMENT FOR DEFECT VISUAL INSPECTION

Chuanfei Hu[1] *and Yongxiong Wang*[1,‡]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
chuanfei_hu@ieee.org, wyxiong@usst.edu.cn

ABSTRACT

Recently, local-editing-based transformations are introduced in anomaly detection for defect visual inspection, which construct a pretext task with the paradigm of self-supervised learning. However, supervised information of local-editing-based transformation may be incorrect when invalid transformation occurs in the pretext task. The reason is that the conventional method to generate labels ignores the differences of images between before and after the transformation. To address this issue, we propose soft label assignment (SLA) to construct soft labels via measuring the similarity between the original and transformed images. Meanwhile, a novel self-supervised learning-based anomaly detection method is proposed for defect visual inspection, which exploits local-editing-based transformation with SLA as a pretext classification task. A convolutional neural network (CNN) is trained to extract deep features of ambiguity and irregularity by the pretext classification task. In the main task, an anomaly detection is modeled via the deep representations to estimate defects regarded as anomalies. Experimental results demonstrate the effect of SLA, and the proposed method achieves superior performance than state-of-the-art methods in terms of the receiver operating characteristic curve (AUC-ROC).

Index Terms— Anomaly localization, deep learning, defect visual inspection, self-supervised learning

1. INTRODUCTION

Computer vision technology has been extensively applied to many defect visual inspections [1, 2] to assist, or even replace the manual inspectors. However, these methods are designed under supervised learning framework. It may cause supervised learning-based method to be ineffective when defects are unknown in prior categories or distributions [3, 4]. Consequently, unsupervised anomaly detection methods, which construct model by normal (defect-free) samples, have attracted much attention in visual defect inspections [5, 6].

For visual defect inspection, self-supervised learning has emerged as a powerful method [8, 9] to construct an unsupervised anomaly detection model. One of the recent methods is to conduct a pretext task by a local-editing-based transformation, editing the local content via image processing algorithms, such as Cutout (CO) [10], CutPaste (CP) and its variant CutPaste-Scar (CPS) [9]. Destination region of an image is first randomly selected where the pixels are edited based on the other content. Then, a CNN is trained to classify whether the image is transformed or not. After the pretext task, the deep representations extracted by trained CNN are more discriminative to model the anomaly detection for the visual defect inspection task. The motivation of local-editing-based transformation is to produce the ambiguity and irregularity to simulate as a coarse approximation of real defects [9]. However, local-editing-based transformations only use binary labels (1 and 0) as supervised information, resulting in insufficient representations of differences between before and after the transformation. Furthermore, the labels may be generated inappropriately while an invalid transformation is occurred. The invalid transformation means that an image is not changed obviously after a transformation. As illustrated in Fig. 1, the preliminary analysis reveal that the invalid transformation is a common case after certain local-editing-based transformation.

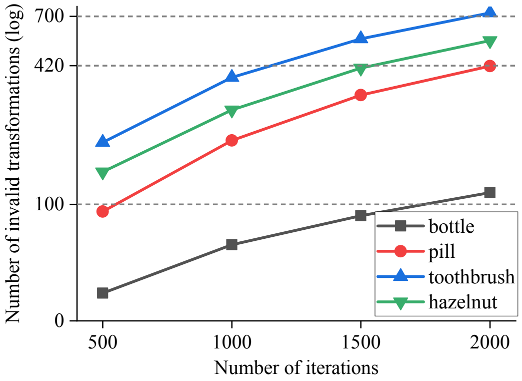

Fig. 1. We repeatedly implement CPS with 2000 times to four samples from different categories in the MVTec AD dataset [7] separately, and record the accumulations of invalid transformations every 500 times. The results statistically illustrate that the invalid transformation is a common case.

To overcome this problem, we propose soft label assign-

‡ Corresponding author: Yongxiong Wang (wyxiong@usst.edu.cn)

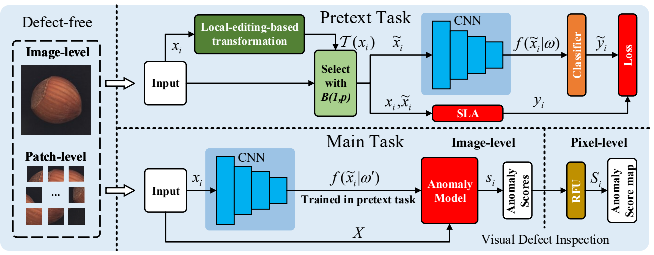

Fig. 2. The framework of the proposed method consists of the pretext task and the main task.

ment (SLA) to construct soft labels for local-editing-based transformations. The label is adapted via measuring the similarity between the original and transformed image, in which the magnitude of the transformation is presented appropriately. The advantage of SLA is that the soft labels can improve the robustness of deep representations, like label smoothing technique [11], and the incorrect labels caused by an invalid transformation can be avoided naturally. Simultaneously, a novel self-supervised learning-based anomaly detection method is proposed for defect visual inspection, which utilizes local-editing-based transformation with the soft labels as a pretext task. Specifically, we first train a CNN to classify the defect-free samples augmented by local-editing-based transformation with SLA. Then, the trained CNN is used to extract more discriminative features from samples. Finally, an anomaly score of a query sample can be estimated by anomaly model based on deep representations, where the sample with high anomaly score is regards as defects. It is noteworthy that the proposed method can be simply extended to localize the defect regions of the images, when the inputs of the model are image patches. In summary, our main contributions are as follows:

- We propose SLA to generate soft supervised information for local-editing-based transformations as a pretext task for visual defect inspection. To the best of our knowledge, we are among the first to measure the transformation magnitude of self-supervised learning for anomaly detection.

- A new self-supervised learning-based anomaly detection method is proposed to achieve image-level defect visual inspection task, which can be further extended to localize the pixel-level defect regions.

2. PROPOSED METHOD

2.1. Framework

The framework of the proposed method, following the paradigm of self-supervised learning, consists of the pretext task and the main task, as shown in Fig. 2. In the pretext task, a learnable deep model is trained to predict the local-editing-based transformation with SLA on the images, in which the model is guided to extract deep features of ambiguity and irregularity. Then, an anomaly model is constructed based on the deep features in the main task. The visual defects are denoted as anomalies can be detected via the anomaly model.

Specifically, the pipeline of the pretext task can be formulated as follows:

$$\tilde{y}_i = g(f(\tilde{x}_i|\omega)|\phi), \tag{1}$$

where ω and ϕ represent the parameters of deep model f and classifier g, respectively. $\tilde{y}_i$ refers to a prediction of local-editing-based transformation for $\tilde{x}_i$ which is an input of the deep model generated as follows:

$$\tilde{x}_i = bx_i + (1-b)\mathcal{T}(x_i), \tag{2}$$

where x_i is a defect-free image, $\mathcal{T}(\cdot)$ is the local-editing-based transformation, and $b \sim Bernoulli(1,p)$ with probability $p = 0.5$.

The backbone of the deep model f is ResNet-18 [12] which is efficient and has been widely used in many computer vision tasks [13,14]. The output of the deep model f is a vector whose length is same as the dimensions of the features. After a classifier g, cross-entropy loss function is used to measure the error between the prediction $\tilde{y}_i$ and the soft label y_i. The trained parameters ω' of f are fixed to extract the deep representations in the main task. Then, an anomaly model is constructed based on the deep representations of defect-free samples. Finally, unknown defects, regarded as anomalies, can be estimated as follows:

$$s_i = z(\hat{x}_i|f_{\omega'}, X), \tag{3}$$

where z is denoted as the anomaly model to estimate the anomaly score s_i of query samples $\hat{x}_i$ which may include defects. X is defect-free set $\{f(x_i|\omega')|i = 1, ..., N\}$, where N is the number of x_i in the dataset.

Intuitively, such image-level framework can be extended to localize the defect regions on the image when the input samples, such as $\hat{x}_i$ and $\tilde{x}_i$, are image patches cropped from the images. Following [9], the image patches are first extracted with a fixed stride and patch size. Then, anomaly scores of each location can aggregate a low-resolution score map S_i^{low}, which are estimated via the same pipeline of the image-level framework. To generate a score map S_i with the size of the holistic image, receptive field upsampling (RFU) [15] is utilized as follows:

$$S_i = RFU(S_i^{low}|k, r, s), \tag{4}$$

where s and r are the stride and the size of image patches. k is a Gaussian filter with size 32 × 32. The area on S_i with high anomaly scores reveal the defect regions of the original image.

2.2. Local-Editing-Based Transformation With SLA

Conventionally, a supervised annotation of local-editing-based transformation $\mathcal{T}(\cdot)$ is generated based on whether the transformation is implemented for the sample, which can be formulated as follows:

$$y'_i = \begin{cases} 1, & \text{if } \mathcal{T}(\cdot) \text{ is implemented} \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where y'_i is a pretext task label of x_i. Obviously, such hard label y'_i may present incorrect supervision information when $\mathcal{T}(\cdot)$ occurs an invalid transformation. Therefore, a method of measuring the similarity between transformed and original image should be considerable in the process of generating labels. The labels can be regarded as presenting the magnitude of the transformation. The negative impact of the invalid transformation would be naturally circumvented.

Here, soft label assignment (SLA) is proposed which is a simple method to compute the number of pixels adjusted via a transformation. For local-editing-based transformation, we argue the insight that the statistics of the adjusted pixels can represent the magnitude of the transformation more efficiently than the numerical comparison of pixels. Specifically, the statistics of the adjusted pixels are first stated as follows:

$$V_i = \sum_{w} \sum_{h} \mathbb{1}\{\Delta(x_i, \tilde{x}_i) > \epsilon\}_{w \times h}, \tag{6}$$

where w and h are the width and height of x_i. ϵ is set to 0.02 ($x_i, \tilde{x}_i \in [0, 1]$), which is a small threshold to tolerate imperceptible changes to human observers. $\mathbb{1}\{\cdot\}$ is an indicator function applied element-wise to derive an indicator matrix belongs to Hamming space $\mathbb{H}^{w \times h}$. $\Delta(\cdot, \cdot)$ is denoted to compare the differences of two multi-dimensional samples:

$$\Delta(A, B) = \begin{bmatrix} \sum\limits_{c} |A^c_{11} - B^c_{11}| & \cdots & \sum\limits_{c} |A^c_{1h} - B^c_{1h}| \\ \vdots & \ddots & \vdots \\ \sum\limits_{c} |A^c_{w1} - B^c_{w1}| & \cdots & \sum\limits_{c} |A^c_{wh} - B^c_{wh}| \end{bmatrix}, \tag{7}$$

where $A, B \in \mathbb{R}^{c \times w \times h}$, A^c_{wh} is the $(w$,$h)$ entry of A in the dimension c. Then, the normalization of V_i can be formulated as follows:

$$V'_i = \frac{V_i}{w' \times h'}, \tag{8}$$

where w' and h' are the width and height of the destination region. Here, we utilize the size of the destination region $w' \times h'$ to replace the size of holistic images $w \times h$ for normalization. The insight is that only the destination region would be adjusted during the transformation, while V_i can be normalized into $[0, 1]$ simply. Furthermore, logistic function, which is shifted and scaled, with an adjustable parameter α is used to alter the level of soft labels flexibly.The soft label y_i can be finally derived as follows:

$$y_i = \frac{2}{1 + e^{-\alpha V'_i}} - 1, \tag{9}$$

where α is set to 10.

2.3. Anomaly Estimation Model

In the main task, we model the anomaly estimation via kernel density estimator (KDE) [16], which is an effective density-based anomaly model. The anomaly scores of the samples can be estimated based on the prior density distribution of defect-free samples. The high anomaly score can be regarded that the defect patterns exist for the sample with the high possibility. Since nonparametric KDE is sensitive the scale of the samples [17], parametric Gaussian density estimator (GDE) [18] is considerable in our task.

3. EXPERIMENT

3.1. Experimental Setup

1) Dataset: The proposed method is evaluated on MVTec AD dataset [7], which is recently constructed to benchmark unsupervised anomaly detection algorithms for industrial products. The dataset consists of 10 object and 5 texture categories containing among 5300 images. The training sets of each category are only composed of defect-free images, and the testing sets include both defect-free and defect images. The precise pixel-level annotations of defect regions are provided to evaluate the performance of algorithms.

2) Implementation Details: The experiments are conducted on a station with a single RTX 2080Ti GPU. The deep model is implemented via deep learning framework PyTorch. Following the common protocol of unsupervised anomaly detection, we train the model for the each category on the corresponding training sets. Meanwhile, the different pretext tasks with local-editing-based transformations including CO, CP, and CPS are implemented independently, The size of images is resized to 256×256, in which the other hyper parameters of local-editing-based transformations are same as the original literature [9]. Stochastic Gradient Descent (SGD) is used as the optimizer with 0.9 momentum, batch size of 64, and initial learning rate of 0.03. The parameters of deep model are regularized via ℓ_2 regularization with a weight of 0.00003. The number training epochs is 256, while the learning rate is adjusted by a single cycle of cosine learning decay [19]. For the training procedure of localization task, the image patches are extracted with stride $s = 4$ and size $r = 32$ as the input of the proposed framework.

3) Evaluation Metric: We adopt the area under the receiver operating characteristic curve (AUC-ROC) as the metric to evaluate the proposed method, which is widely used for anomaly detection tasks. The anomaly samples and regions are treated as positive in AUC-ROC.

3.2. Comparisons With Other Methods

We first implement the proposed method for CPS which is compared with other anomaly detection methods for image-level defect visual inspection. Self-supervised learning-based

Table 1. Comparison of our methods with other anomaly detection methods for image-level AUC-ROC (%) on MVTec AD dataset. The best results are highlighted in **bold**.

Category	DOCC [20]	U-Student [21]	P-SVDD [8]	CPS [9]	CPS-SLA
Carpet	90.6	95.3	92.9	94.6	**95.4**
Grid	52.4	**98.7**	94.6	95.5	95.9
Leather	78.3	93.4	90.9	**100**	99.9
Tile	96.5	95.8	**97.8**	89.4	90.7
Wood	91.6	95.5	96.5	98.7	**99.1**
Bottle	**99.6**	96.7	98.6	98.0	98.2
Cable	**90.9**	82.3	90.3	78.8	80.5
Capsule	91.0	92.8	76.7	95.3	**95.7**
Hazelnut	95.0	91.4	92.0	96.7	**97.3**
Metal nut	85.2	94.0	94.0	97.9	**98.3**
Pill	80.4	86.7	86.1	85.8	**87.9**
Screw	86.9	**87.4**	81.3	83.7	85.6
Toothbrush	96.4	98.6	**100**	96.7	97.8
Transistor	90.8	83.6	91.5	91.1	**92.0**
Zipper	92.4	95.8	97.9	**99.5**	99.1
Average	87.8	92.5	92.0	93.4	**94.2**

Table 2. Comparison of our methods with other anomaly detection methods for pixel-level AUC-ROC (%) on MVTec AD dataset. The best results are highlighted in **bold**.

Category	SSIM-AE [22]	AnoGan [23]	FCDD [15]	P-SVDD [8]	CPF [9]	CPF-SLA
Carpet	87	54	96	92.6	98.3	**98.7**
Grid	94	58	91	96.2	97.5	**98.2**
Leather	78	64	98	97.4	99.5	**99.8**
Tile	59	50	91	**91.4**	90.5	91.0
Wood	73	62	88	90.8	95.5	**96.3**
Bottle	93	86	97	98.1	97.6	**98.2**
Cable	82	78	90	**96.8**	90.0	90.5
Capsule	94	84	93	95.8	97.4	**97.9**
Hazelnut	97	87	95	97.5	97.3	**98.1**
Metal nut	89	76	94	**98.0**	93.1	93.4
Pill	91	87	81	95.1	95.7	**95.9**
Screw	96	80	86	95.7	96.7	**97.2**
Toothbrush	92	90	94	98.1	98.1	**98.3**
Transistor	90	80	88	**97.0**	93.0	93.6
Zipper	88	78	92	95.1	**99.3**	99.1
Average	86.8	74.2	91.6	95.7	95.9	**96.4**

methods include P-SVDD [8], CPS [9] and CPS-SLA, while CPS and CPS-SLA are based on local-editing-based transformation. "-SLA" represents the local-editing-based transformation with SLA. The quantitative results are reported in Tab. 1, where the results of each category are presented and the advantage performances of corresponding methods are listed in the last row. It shows that the proposed CPS-SLA achieves a superior performance in terms of AUC-ROC.

Furthermore, the proposed method is evaluated to localize anomaly regions for pixel-level defect visual inspection. We implement SLA for CutPaste-Fusion (CPF) [9] which includes CP and CPS. It means that the deep model is trained to distinguish the transformations of CP and CPS simultaneously on image patches in the pretext classification task. The proposed method based on CPF-SLA is compared with other methods including SSIM-AE [22], AnoGan [23], FCDD [15], and P-SVDD [8]. As listed in Tab. 2, the proposed CPF-SLA surpasses other methods in terms of pixel-level AUC-ROC. The effect of local-editing-based transformations with SLA is demonstrated for constructing self-supervised learning-based defect visual inspection, which can both detect and localize the defects on the samples under the unsupervised anomaly detection framework.

Table 3. Image-level results of ablation study for SLA on MVTec AD dataset. "w/o" and "w" mean the transformation "without" and "with" SLA, respectively.

Methods	AUC-ROC (%)		
	w/o	w	RImp (%)
CO	71.3	72.6	1.82
CP	90.9	91.8	0.99
CPS	93.5	94.2	0.74

Table 4. Image-level results of ablation study for α on MVTec AD dataset.

α	1	2	5	7	10	15
AUC-ROC (%)	67.9	79.5	91.2	92.7	94.2	94.1

3.3. Ablation studies

1) Effect of soft label assignment: To clarify the effect of SLA for local-editing-based transformation in our task, we design the pretext tasks via CO, CP, and CPS, respectively. The average results of each pretext task are reported in Tab. 3, where the local-editing-based transformations with SLA possess the superior performances. There are among 0.7 to 1.8% relative improvements (RImp) for each local-editing-based transformation. It demonstrates the effect of SLA, which can enhance the discrimination of the deep representations via avoiding invalid transformation and soft assignment.

2) Effect of parameter α: To reveal the effect of parameter α, the value of α is selected heuristically with constants. CPS-SLA with different values of α is used to conduct the pretext tasks whose performances are presented in Tab. 4. It can be seen that the best value of α is 10 for the image-level defect visual inspection, while the higher α can not further improve the performance.

4. CONCLUSION

In this paper, we propose a novel method, soft label assignment (SLA), to construct soft labels for local-editing-based transformations in a paradigm of self-supervised learning. Incorrect supervised information can be avoided naturally, and deep features are represented more discriminative in the pretext task. Meanwhile, a new self-supervised learning-based anomaly detection method is proposed for defect visual inspection. The defect regions are detected and localized via unsupervised anomaly model based on deep representations. Experimental results on MVTec AD dataset demonstrate that the proposed method achieves the superior performance compared to other competitive methods.

5. REFERENCES

[1] Binyi Su, Haiyong Chen, and Zhong Zhou, “Baf-detector: An efficient cnn-based detector for photovoltaic cell defect detection,” *IEEE Trans. Ind. Electron.*, pp. 1–1, 2021.

[2] C. Hu and Y. Wang, “An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images,” *IEEE Trans. Ind. Electron.*, vol. 67, no. 12, pp. 10922–10930, 2020.

[3] Victoria J. Hodge and Jim Austin, “A survey of outlier detection methodologies,” *Artif Intell Rev*, vol. 22, pp. 85–126, 2004.

[4] D. Lu and Q. Weng, “A survey of image classification methods and techniques for improving classification performance,” *Int J Remote Sens*, vol. 28, no. 5, pp. 823–870, 2007.

[5] Jian Zhou, Dimitri Semenovich, Arcot Sowmya, and Jun Wang, “Dictionary learning framework for fabric defect detection,” *The Journal of The Textile Institute*, vol. 105, no. 3, pp. 223–234, 2014.

[6] Jian Zhou, Jun Wang, and Honggang Bu, “Fabric defect detection using a hybrid and complementary fractal feature vector and fcm-based novelty detector,” *Fibres & Textiles in Eastern Europe*, vol. 25, no. 6, pp. 46–52, 2017.

[7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, “Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection,” in *CVPR*, 2019, pp. 9592–9600.

[8] Jihun Yi and Sungroh Yoon, “Patch svdd: Patch-level svdd for anomaly detection and segmentation,” in *ACCV*, 2020, pp. 375–390.

[9] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister, “Cutpaste: Self-supervised learning for anomaly detection and localization,” in *CVPR*, June 2021, pp. 9664–9674.

[10] Terrance DeVries and Graham W Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv:1708.04552*, 2017.

[11] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li, “Bag of tricks for image classification with convolutional neural networks,” in *CVPR*, 2019, pp. 558–567.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, June 2016, pp. 770–778.

[13] Bo Dong, Yan Zhou, Chuanfei Hu, Keren Fu, and Geng Chen, “Bcnet: Bidirectional collaboration network for edge-guided salient object detection,” *Neurocomputing*, vol. 437, pp. 58–71, 2021.

[14] Bo Dong, Mingchen Zhuge, Yongxiong Wang, Hongbo Bi, and Geng Chen, “Towards accurate camouflaged object detection with mixture convolution and interactive fusion,” 2021.

[15] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus Robert Muller, “Explainable deep one-class classification,” in *International Conference on Learning Representations*, 2021.

[16] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister, “Learning and evaluating representations for deep one-class classification,” in *ICLR*, 2021.

[17] Artur Gramacki, *Nonparametric kernel density estimation and its computational aspects*, Springer, 2018.

[18] Benjamin Nachman and David Shih, “Anomaly detection with density estimation,” *Phys. Rev. D*, vol. 101, pp. 075042, Apr 2020.

[19] Ilya Loshchilov and Frank Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” in *2017 International Conference on Learning Representations*, 2016.

[20] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller, “A unifying review of deep and shallow anomaly detection,” *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, 2021.

[21] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings,” in *CVPR*, 2020, pp. 4182–4191.

[22] Paul Bergmann., Sindy L5̈owe., Michael Fauser., David Sattlegger., and Carsten Steger., “Improving unsupervised defect segmentation by applying structural similarity to autoencoders,” in *VISAPP*, 2019, pp. 372–380.

[23] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth, “f-anogan: Fast unsupervised anomaly detection with generative adversarial networks,” *Med. Image Anal.*, vol. 54, pp. 30 – 44, May. 2019.