FOCUS YOUR DISTRIBUTION: COARSE-TO-FINE NON-CONTRASTIVE LEARNING FOR ANOMALY DETECTION AND LOCALIZATION

Ye Zheng**, Xiang Wang†, Rui Deng‡, Tianpeng Bao†, Rui Zhao†, Liwei Wu†

*Institute of Computing Technology, Chinese Academy of Sciences; *University of Chinese Academy of Sciences; †SenseTime Research; ‡University of California, Los Angeles.

ABSTRACT

The essence of unsupervised anomaly detection is to learn the compact distribution of normal samples and detect outliers as anomalies in testing. Meanwhile, the anomalies in real-world are usually subtle and fine-grained in a high-resolution image especially for industrial applications. Towards this end, we propose a novel framework for unsupervised anomaly detection and localization. Our method aims at learning dense and compact distribution from normal images with a coarse-to-fine alignment process. The coarse alignment stage standardizes the pixel-wise position of objects in both image and feature levels. The fine alignment stage then densely maximizes the similarity of features among all corresponding locations in a batch. To facilitate the learning with only normal images, we propose a new pretext task called non-contrastive learning for the fine alignment stage. Non-contrastive learning extracts robust and discriminating normal image representations without making assumptions on abnormal samples, and it thus empowers our model to generalize to various anomalous scenarios. Extensive experiments on two typical industrial datasets of MVTec AD and BenTech AD demonstrate that our framework is effective in detecting various real-world defects and achieves a new state-of-the-art in industrial unsupervised anomaly detection.

Index Terms— Unsupervised anomaly detection, non-contrastive, coarse-to-fine

1. INTRODUCTION

Image anomaly detection is the identification of unexpected or abnormal image patterns in the dataset, which has wide applications [1]. Different from classical supervised learning tasks that assume an even distribution among classes, the anomalies occur rarely in real-world and are often hard to collect and label. Moreover, the lack of prior knowledge about anomalous patterns imposes a great challenge for designing comprehensive anomaly detection algorithms. Due to the scarcity and uncertainty of abnormal images, existing anomaly detection methods usually follow the unsupervised or one-class classification setting. That is, models are provided with only normal data in training. During inference, the anomaly is spotted by the difference between the test data and learned normal features. Existing works [2, 3] are proven to be successful in abstracting semantically rich representation for isolating defect images; nonetheless, they lack the ability to explore the fine-grained structures for anomalies. For example, a common setting in previous works [3] is to set one category as the normal class and the rest as anomalies. In actual manufacturing or medical industries, however, the difference between normal images and anomalies is more fine-grained and subtle than these object class differences [1].

We thus designed a novel framework targeting the fine-grained anomalous patterns in actual industrial setting, where images are usually taken under a clean background and shared positions and defects are usually subtle. The intuition of our method is inspired by the human inference process. When asked to play the spot the difference game, human beings would usually first roughly align, or find the correspondence, between the global context of two images. Then, they closely examine the detailed local distinction underlying two patterns. Inspired by this, we design a two-stage coarse-to-fine framework that learns robust feature distributions for normal images.

We first apply a coarse alignment module to roughly extract and align global feature embedding. The module operates on both pixel-level for the input image and feature-level for each pyramid feature map. In the fine alignment stage, we apply self-supervised learning and propose a novel pretext task for learning the normal representation. The current state-of-the-art [4] in self-supervised anomaly detection designs augmentations that generate abnormal samples through mixing normal image patches. However, we lack sufficient prior knowledge of the real-world anomaly distributions, so the created defects cannot model the numerous real-life possibilities of anomalies. We thus define a new task for self-supervised learning called non-contrastive learning - using no abnormal samples and only normal images to train a robust feature encoder. By enforcing the similarity among each position’s feature from a minibatch, we capture the local fine-grained correspondence in every aligned position of images. The distribution of normal images thus becomes more compact and
more semantically meaningful, making the abnormal outliers more salient and easier to detect.

To summarize, the main contributions of this paper are: (1) we propose a coarse-to-fine anomaly detection paradigm to detect and localize the fine-grained defects in real-world industrial dataset; (2) we propose a novel pretext task named dense non-contrastive learning for self-supervised learning of compact normal features without any assumption of abnormal samples; (3) we provide extensive experimental results and ablation studies to highlight the strength of our method, and the results in MvTec anomaly detection dataset [1] show that our method outperforms the previous state-of-the-art anomaly detection methods.

2. RELATED WORK

The mainstream unsupervised anomaly detection methods are either reconstruction-based or representation-based.

**Reconstruction-based method** applies autoencoders [1] or generative adversarial networks to encode and reconstruct the normal data. In inference, an anomaly is spotted when the reconstructed image diverges from the original one. The pixel-wise reconstruction error can be applied to localize anomalies [1], and the image level anomaly score is thus determined by aggregating pixel-wise errors. Despite the high interpretability, the pixel-wise difference fails to encode the global semantic meaning of images [4].

**Representation-based method**, on the other hand, extracts discriminative feature vectors from normal images [5] or normal image patches [6] and yields more promising results for anomaly detection. The anomaly score is calculated by the distance between the embedding of a test image and the distribution of normal image representations. The normal image distribution is typically characterized by the center of a n-sphere for the normal image [5], the Gaussian distribution of normal or the kNN for the entire normal image embedding [7].

To assist the learning of semantic vectors for images, many works [3, 8] employed self-supervised learning [9, 10] to discriminate normal data and outliers. Although these methods well capture the semantic object information in images, they fail to encode the fine-grained local irregularities in anomalies [4]. The method CutPaste [4] cuts a patch from one image and paste it on the other. However, the representation of created negative irregularities usually does not overlap with real-world anomalies [4], which limits the generalization potential of these methods in inference processes. In our paper, we eliminate any prior assumption about the anomalous data in training, and our model can therefore generalize to a variety of anomalies in real world. Furthermore, through our proposed coarse alignment of images and dense supervision of pixel-wise feature learning, we reduce the variances in normal data representation, enabling the learned compact distribution to predict robust distance estimates for outliers.

3. METHODOLOGY

In this section, we demonstrate our novel framework to detect and localize fine-grained anomalies. As indicated in Figure 1, our method consists of a coarse alignment stage and a fine alignment stage.

3.1. Coarse Alignment Stage

3.1.1. Image-Level Coarse Alignment (ICA)

The image-level coarse alignment (ICA) aims at regularizing the pixel distribution of normal images: it orients all images in a batch to a similar direction and position for dense comparison. Specifically, we regress the affine transformations \( T_0(D_i) \) on an input image \( D_i \in \mathbb{R}^{H \times W \times C} \):

\[
\begin{bmatrix} x'_1 \\ y'_1 \\ \vdots \\ x'_N \\ y'_N \\ 1 \end{bmatrix} = T_0(D_i) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ & & \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_N \\ y_N \\ 1 \end{bmatrix}
\]

(1)

Inspired by Spatial Transformer Network [11], we adopt its similar architecture to our ICA module, which uses a tiny network \( h_{\tau_0} \) to learn the above affine mapping from the original image (denoted by \( \{(x'_i, y'_i)\} \)) to a globally aligned representation (\( \{(x_i, y_i)\} \)). To train ICA module, we randomly pair the images in a batch and then minimize the \( \ell_2 \) distance between them:

\[
\mathcal{L}_{ICA}(D; \theta_h, T_0) = \sum_{A,B \in \mathcal{D}} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \| h_{\tau_0}(A_{i,j}) - h_{\tau_0}(B_{i,j}) \|_2.
\]

(2)

Note that we do not assign any standard orientation or alignment for our module to regress, so it learns a unified position in a self-supervised manner. Each time we random select two images \( A, B \in \mathcal{D} \), the \( \ell_2 \) loss in Equation 2 supervises the alignment of \( A \) and \( B \) toward the reduction of entropy in this system. Thus, given enough iterations, the system reaches a consensus on alignment and the entropy is thus reduced to a local minimum. The results of roughly unified positions are displayed by the coarsely-aligned images in Figure 2.

3.1.2. Feature-Level Coarse Alignment (FCA)

For transformed images in Figure 2, their positions are not strictly aligned to a unified representation. We thus introduce a feature-level coarse alignment (FCA) module to further adjust the high-level image representations. We use a pre-trained ResNet-18 as the feature extractor. As shown in Figure 1, we insert the FCA between succeeding layers to align the embedding distribution with feature-level affine transformation. FCA has a similar implementation as ICA; nonetheless, this module enforces the backbone to extract generalizable features to align positions in a global high-level embedding. The FCA module is only governed by self-supervised loss in the fine
alignment stage. More implementation details of image-level and feature-level alignment are in Appendix.

### 3.2. Fine Alignment Stage

Non-contrastive learning is the training of a normal image representation without leveraging its distance with anomalies. To detect the fine-grained anomalies, we propose the pixel-wise alignment module, which maximizes the feature similarity across every embedding position for all normal images.

As indicated by the Algorithm 1, we first randomly shuffle the minibatch and sample two feature maps. Let $W, V \in \mathbb{R}^{H' \times W' \times C'}$ denote the two encoded image feature maps belonging to two different images from the last FCA. Then, for every position $0 \leq i < H', 0 \leq j < W'$ in these two features, we extract the corresponding feature vectors $w_{ij} \sim W$ and $v_{ij} \sim V$. We aim at encoding a unique vector representation for each position in the feature map, as well as narrowing its distribution for all normal images. Thus, we use $1 \times 1$ convolutional operator instead of fully-connected layers in the feature extractor. The vectors are passed to a shared 3-layer $1 \times 1$ conv encoder $f$. Only $w_{ij}$ is processed through a 2-layer $1 \times 1$ conv predictor $g$ to project its feature to the vector space of $v_{ij}$. Given the two output vectors from the encoder $m_{ij} \triangleq f(w_{ij})$ and $n_{ij} \triangleq f(v_{ij})$, we minimize their negative cosine similarity:

$$
L_{ij}(m_{ij}, n_{ij}; \theta_g, \theta_f) = - \frac{(g(m_{ij}), n_{ij})}{\|m_{ij}\|_2 \cdot \|n_{ij}\|_2},
$$

where $\theta_g$ and $\theta_f$ are the parameters of encoder and predictor, respectively. We conduct the above minimization for $w$ and $v$ at all positions $i, j$ respectively to densely supervise the positionally-aligned feature distribution.
To avoid model collapsing when training with normal data only, we introduce the stop-gradient operation. That is, $\nabla L_{ij}$ is only allowed to descent backward through the upper branch of the network w.r.t. $m_{ij}$, and it updates no information of $n_{ij}$ to the encoder $f$. A symmetry operation is applied to further supervise the learning of robust and generalizable features. The aggregated loss for every position thus becomes:

$$L_{FAS}(m,n; \theta_g, \theta_f) = \sum_{i,j} \sum_{D} \frac{1}{2} L_{ij}(m_{ij}, \text{stop}\_\text{grad}(n_{ij})) + \frac{1}{2} L_{ij}(n_{ij}, \text{stop}\_\text{grad}(m_{ij})).$$

Its implementation is detailed in Algorithm 1. We train the above stages in an end-to-end manner and adjust the weight between coarse and fine alignment with $\lambda_1$ and $\lambda_2$. Hence, the final loss in our framework is:

$$L_{\text{total}}(\cdot; \theta_h, f, g, T_\theta) = \lambda_1 \cdot L_{ICA} + \lambda_2 \cdot L_{FAS}.$$  

$L_{FAS}$ is the dominant loss function supervising all parameters and $L_{ICA}$ is the auxiliary loss function used to guarantee the convergence of ICA. By optimizing the coarse and fine alignment module collectively, we allow the network to self-adjust and learn meaningful correlations of normal image embeddings. Please refer to our appendix for more principle analysis.

### 3.3. Anomaly Score Computation in Inference

With the densely extracted features, we model the representation of normal images with the Gaussian distribution for every pixels on feature map following [12]. We extract the normal image representation at position $(i,j)$ by concatenating the three pyramid layers of features of CNN at $(i,j)$. Let $X_{ij} \in \mathbb{R}^{(C_1+C_2+C_3) \times N}$ denote the aggregated feature from the CNN for all images of training set. We model a distinctive Gaussian distribution $N(M_{ij}, \Sigma_{ij})$ for each pixel $(i,j)$ on the feature map by:

$$\mu_{ij} = \frac{1}{N} \sum_{k} x_{ij}^k; \Sigma_{ij} = \frac{1}{N-1} \sum_{k} (x_{ij}^k - \mu_{ij})(x_{ij}^k - \mu_{ij})^T$$

During inference, we compute the anomaly score by taking the Mahalanobis distances at every pixel between the test images and the normal distribution:

$$D(x_{ij}) = \sqrt{(x_{ij} - \mu_{ij})^T \Sigma_{ij}^{-1}(x_{ij} - \mu_{ij})}$$

Then, the distance matrix $D$ is an anomaly map with dense pixel-wise anomaly scores. A greater score indicates a severer anomalous signal. We thus use the maximum anomaly score map to represent the anomaly score for the entire image.

### 4. EXPERIMENTS

#### 4.1. Datasets and Metrics

We perform experiments on two industrial anomaly detection datasets MVTec AD dataset [1], BeanTech AD dataset [13] and Disturbed MVTec AD dataset. In all datasets, the training set consists of only normal images, while the testing set has a mixture of both normal and abnormal images. These datasets provide both anomaly types and anomaly masks as test image labels for evaluation. During inference, we evaluate our method with image-level AUC and pixel-level AUC. Considering that the current MVTec AD dataset does not contain products with multiple appearances but only has spatially aligned products, we build a new dataset called Disturbed MVTec to simulate more challenging real life detection situations and verify the robustness and effectiveness of our framework. Limited by the space, please refer our appendix for the details and results of BeanTech AD dataset and Disturbed MVTec AD dataset.

#### 4.2. Anomaly Detection and Localization for MVTec

In Table 1, we compare our method with the state-of-the-art one-class anomaly detection approaches in MVTec AD dataset, including deep one-class classifier (DOCC) [14], FCDD [15], uninformed student (V-S) [6], patch SVDD [16], SPADE [17], PaDiM [12], Cut Paste [4] under the metrics of image-level AUC and pixel-level AUC, we give the results of the mean and standard deviation of 5 repeated experiments. With our proposed coarse-to-fine non-contrastive learning method, we achieve the best result among all existing works and make notable improvements on both of texture and object defects. Our method surpasses the current state-of-the-art by a margin of 2.1, yielding 97.7 image-level AUC and 98.2 pixel-level AUC. Some comprehensive results of defect localization are provided in Appendix.

#### 4.3. Quantitative Analysis on Distribution

To quantify the improvement of distribution compactness in our work, we compare the distance variance score of our method to the baseline (ImageNet pre-trained model). Given a class of total $N$ normal image representations $X \in \mathbb{R}^{N \times H \times W \times C}$, the distance variance score is calculated by first computing each pixel’s distance to its alignment center:

$$\mu_{ij} = \frac{1}{N} \sum_{n} x_{nij}, \forall 0 \leq i < H, 0 \leq j < W$$

$$D_{nij} = \sqrt{(x_{nij} - \mu_{ij})^T \Sigma_{ij}^{-1}(x_{nij} - \mu_{ij})}$$

Then, we get the distance map $D \in \mathbb{R}^{N \times H \times W}$ at every location for all normal images. We compute the distance variance...
We investigate the contributions of the main components for our method in Table 2. "Baseline" only uses the ImageNet pre-trained ResNet-18 to model the Gaussian distribution in Equation 6 in inference. Adding a single non-contrastive learning block improves the image-level AUC to 95.7 and pixel-level AUC to 96.8. This demonstrates the effectiveness of our designed non-contrastive learning module, because it eliminates the abnormal samples in training and shrinks the distribution of normal samples. Then, adding the coarse alignment module further enhances our advantage over the current state-of-the-art. This is very intuitive, since without first aligning the coarse locations of images, the densely minimized distance among pixels may not be correctly associated. The ablation study shows the additive effect of each module and the comprehensiveness of our framework.

### 5.2. Effects of Coarse Alignment Stage

We give the qualitative results of a specific class to show the effects of the coarse alignment stage in Table 3. We can observe that the coarse align stage can improve most of the categories in MVTec AD dataset, especially for "screw", and we will discuss this improvement is achieved below.
We investigate the effects of the position of the feature-level coarse alignment module in Table 4. The baseline is the clean backbone without feature-level alignment. We find that inserting FCA after a single feature layer has limited improvement over the baseline, while inserting it in all three layers gives a thorough boost in AUC and PRO. We speculate that although adding FCA in a single layer enables the network to adjust the feature’s positions, it limits the flexibility and reception field of the alignment to a single scope. Adding them to all three layers successively reinforces the feature alignment process and allows the self-supervised signal in Equation 4 to descent backward without information loss.

6. CONCLUSION

In this paper, we propose a coarse-to-fine non-contrastive learning framework for unsupervised anomaly detection. The key to our success is the dense non-contrastive learning with coarse alignment and fine alignment module, which encourages the model to learn and narrow down the distribution of normal patterns. Our method achieves high performance on the industrial defect dataset and surpasses the state-of-the-art approaches in both anomaly detection and localization tasks.

7. REFERENCES

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, “Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9592–9600.

[2] Bernhard Schölkopf, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt, et al., “Support vector method for novelty detection,,” in NIPS, Citeseer, 1999, vol. 12, pp. 582–588.

[3] Izhak Golan and Ran El-Yaniv, “Deep anomaly detection using geometric transformations,” arXiv preprint arXiv:1805.10917, 2018.

[4] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister, “Cutpaste: Self-supervised learning for anomaly detection and localization,” arXiv preprint arXiv:2104.04015, 2021.

[5] Lukas Ruff, Robert Vandermeulen, Nico Goerntz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft, “Deep one-class classification,” in International conference on machine learning. PMLR, 2018, pp. 4393–4402.

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4183–4192.

[7] Liron Bergman, Niv Cohen, and Yedid Hoshen, “Deep nearest neighbor anomaly detection,” arXiv preprint arXiv:2002.10445, 2020.

[8] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister, “Learning and evaluating representations for deep one-class classification,” International Conference on Learning Representations (ICLR), 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, “Spatial transformer networks,” arXiv preprint arXiv:1506.02025, 2015.

[12] Thomas Defard, Aleksandr Setkov, Anelique Loesch, and Romaric Audigier, “Pdim: a patch distribution modeling framework for anomaly detection and localization,” arXiv preprint arXiv:2011.08785, 2020.

[13] Pankaj Mishra, Riccardo Verc, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti, “Vt-adl: A vision transformer network for image anomaly detection and localization,” arXiv preprint arXiv:2104.10036, 2021.

[14] 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,” Proceedings of the IEEE, 2021.

[15] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller, “Explainable deep one-class classification,” arXiv preprint arXiv:2007.01760, 2020.

[16] Jihan Yi and Sungroh Yoon, “Patch svdd: Patch-level svdd for anomaly detection and segmentation,” in Proceedings of the Asian Conference on Computer Vision, 2020.

[17] Niv Cohen and Yedid Hoshen, “Sub-image anomaly detection with deep pyramid correspondences,” arXiv preprint arXiv:2005.02257, 2020.