# Learning Transferable Representations for Image Anomaly Localization Using Dense Pretraining—Supplementary Material

Haitian He[1] Sarah Erfani[1] Mingming Gong[1] Qiuhong Ke[2]
[1]The University of Melbourne [2]Monash University
{haitianh@student., sarah.erfani@, mingming.gong@}unimelb.edu.au qiuhong.ke@monash.edu

| | Ins. Model | object | texture | **all** |
|---|---|---|---|---|
| (b) | -* | **96.5**±0.1 | **90.8**±0.4 | **94.6**±0.1 |
| | BYOL | 95.6±0.1 | 88.9±0.5 | 93.4±0.2 |
| | MoCoV2 | 95.5±0.4 | 88.9±1.1 | 93.3±0.7 |
| | RotNet(MLP head) | 96.0±0.2 | 90.0±0.5 | 94.0±0.3 |
| | DistAug | 95.5±0.1 | 87.9±0.2 | 93.0±0.1 |
| | Augmentation | object | texture | **all** |
| (c) | cj | **96.5**±0.1 | **91.0**±0.2 | **94.6**±0.1 |
| | gb | 95.8±0.1 | 90.5±0.2 | 94.0±0.0 |
| | cj+gb* | **96.5**±0.1 | 90.8±0.4 | **94.6**±0.1 |
| | cj+gb+gs | 96.0±0.0 | 89.9±0.3 | 93.9±0.1 |
| | cj+gb+so | 96.4±0.1 | 89.5±0.2 | 94.1±0.1 |
| | cj+gb+gs+so | 95.8±0.0 | 88.8±0.2 | 93.5±0.0 |
| | Batch Size | object | texture | **all** |
| (d) | 32 | 96.3±0.1 | 89.9±0.5 | 94.2±0.2 |
| | 64 | **96.6**±0.0 | 90.7±0.2 | **94.6**±0.1 |
| | 128 | **96.6**±0.1 | 90.7±0.1 | **94.6**±0.1 |
| | 256* | 96.5±0.1 | **90.8**±0.4 | **94.6**±0.1 |
| | 512 | 96.2±0.1 | **90.8**±0.2 | 94.4±0.1 |
| | Scales | object | texture | **all** |
| (e) | c5 | 96.2±0.2 | 90.4±0.4 | 94.3±0.2 |
| | c4c5* | **96.5**±0.1 | **90.8**±0.4 | **94.6**±0.1 |
| | c3c4c5 | 96.3±0.1 | 90.1±0.2 | 94.2±0.0 |

Table 1. The localization performance of DS2 under (b) incorporating different instance pretraining models, (c) different augmentation choices, (d) different batch sizes, and (e) different scales of feature maps. The best results are bold-faced, and the choice used in DS2 is marked with *. ("**-**": no instance branch; **cj**: color jitter; **gb**: Gaussian blur; **gs**: grayscale; **so**: solarize)

## 1. Detailed Results for Ablation Studies (b)-(e)

The detailed results for ablation studies (b)-(e) are presented in Table 1.

## 2. Implementation Details of DS2

We pretrain DS2 for 400 epochs. We use step scheduler with learning rate ($lr$) decay epochs set at 120, 160, and 200; the decay rate is 0.1. We choose LARS optimizer as it is commonly used in SSL works [1, 5, 14]. The initial $lr$ is linearly scaled with the batch size ($bs$): $lr = lr_{base} \times bs/256$, with the base learning rate ($lr_{base}$) of 2.0 and $bs$ of 256. The weight decay is set to 1e-5. The data augmentation $\mathcal{T}$ includes random resized crop, horizontal flip, color jitter, and Gaussian blur. All augmented views are resized to 224 × 224 before being processed by the encoder. The positive-pair distance threshold $\delta$ is set to 0.1.

## 3. Re-implementation Details of CutPaste(3-way)

We follow the reported implementation details in [9] and reference the publicly available code[1]. The re-implementation choices are listed below.

1. Model architecture
   - The feature extractor is a ResNet-18 [7] ending with average-pooling but without the last fully-connected layer.
   - The classifier consists of three fully-connected layers, with batch normalization [8] and ReLU activation in-between. The input dimension for classifier is 512, the output dimension is 3, and the intermediate dimensions are 512 and 128, respectively.
   - The training loss is three-way cross entropy loss.
2. Scheduling
   - The optimizer is momentum SGD with learning rate of 0.03, momentum of 0.9, and weight decay of 0.00003.
   - The scheduler is a single cycle of cosine annealing.
   - The batch size is 96, and the model is trained for 256 epochs, with 256 steps in each epoch.

[1] https://github.com/Runinho/pytorch-cutpaste

3. Data augmentation for 3-way classification

    - The inputs for the model are $64 \times 64$ patches from $256 \times 256$ images. Color jitter is randomly applied to the patch.
    - For normal CutPaste augmentation, the area ratio of the patch with respect to the image is randomly chosen from the range $(0.02, 0.15)$. The aspect ratio between the height and width of the patch is randomly chosen from the range $(0.3, 1) \cup (1, 3.3)$.
    - For CutPaste-Scar augmentation, we make slight adjustment to the patch size. The original paper mentions that the patch size should be between $[2, 16]$ in width and $[10, 25]$ in height. However, this absolute size is selected for $256 \times 256$ image-based training. As for 3-way classification, the input is $64 \times 64$ patch, so we accordingly reduce the width to $[1, 4]$ by a reduction ratio of $1/4$, and adjust the height to $[5, 6]$ so that the patch holds the scar shape. The patch is randomly rotated between $[-45, 45]$ degrees before being pasted into the image.
    - For both augmentation methods, color jitter is applied to the patch with maximum intensity of $0.1$, and the paste-back location of the patch is selected in a way that the patch can be fitted into the image without being cut off.

## 4. Details on the Usage of Evaluation Datasets

- The MVTec AD is an industrial defect dataset, containing five texture categories (*e.g.*, leather, carpet) and ten object categories (*e.g.*, hazelnut, screw). The training set contains 3,629 anomaly-free (normal) images, and the testing set contains 1,725 images that are either anomaly-free or with anomalous regions. For each testing image with defect, a pixel-accurate anomaly ground-truth mask is provided. In our experiments, we pretrain one model using all the 3,629 training images regardless of their categories. After pretraining, the learned model is evaluated against each category's test images.

- The MVTec LOCO dataset introduces logical anomaly. However, since the DS2 and the baselines are not designed for detecting logical anomalies, we evaluate them on structural-anomalous and anomaly-free test images only.

- The KSDD2 is a dataset of defective production items with challenging "near-in-distribution" anomalies. It provides anomalous images in the training set for supervised training. However, as DS2 and the baselines all perform self-supervised training using normal images, we train them only with the normal images from the training set and then evaluate them on the whole test set.

- The MTD dataset contains normal and five defective types of magnetic tiles. We follow the common practice [10, 13] by taking 80% normal images for training and then evaluating on the rest 20% normal images plus all the anomalous ones.

## 5. Related Works on Self-supervised Learning

Recently, SSL overtakes the dominance of ImageNet-supervised [4] pretraining and provides a feasible solution to learn representations without the need of manually labeled data. One can categorize SSL into two categories: (1) instance-level representation learning and (2) dense-level representation learning.

In the first category, each image is deemed as an individual class, and the goal is to maximize the similarity in the representation space between two views from the same image (*i.e.,* positive pair), while minimize the feature similarity between views from different images (*i.e.,* negative pair) [1, 2, 6]. Grill et al. [5] proposed the BYOL model which gets rid of negative pairs without causing model collapsing. The authors owed its success to the additional predictor in the online branch and the moving-average parameter update fashion of its target branch. Later on, Chen et al. [3] discovered that it is the stop-gradient design in the target branch that prevents BYOL from collapsing, while the moving-average design is more related to accuracy improvement than collapsing prevention.

In the second category, feature comparisons are performed on the dense feature vector level to learn representations that are more compatible with dense downstream tasks (*e.g.,* semantic segmentation). Wang et al. [11] proposed the DenseCL model, which extends SimCLR [1] and defines positive and negative pairs on the granularity of individual feature vectors. Xie et al. [14] extended BYOL and proposed PixPro model. Different from BYOL, PixPro establishes positive pairs according to the geometric proximity of feature vectors. Wang et al. [12] further enhanced dense pretraining robustness with SetSim, where positive and negative pairs are set-based in order to include more semantic and structural information.

## 6. More Qualitative Results of DS2 on MVTec AD Dataset

We include more localization visual examples of DS2 on each category of MVTec AD dataset in Figures 1–5. In each figure, the first six rows show successful cases, and the last two rows show failed cases.

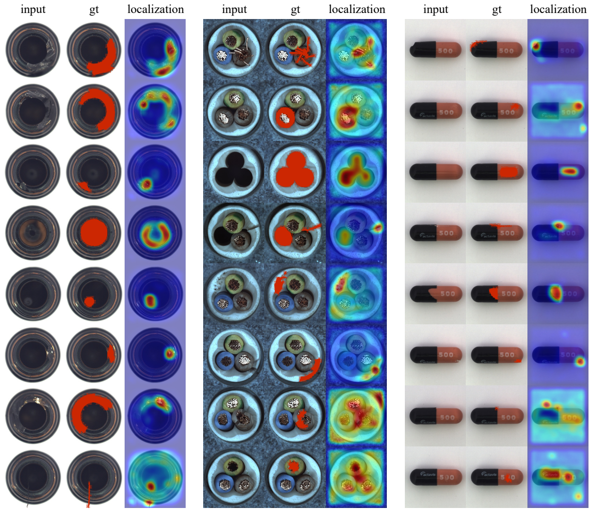

Figure 1. More localization visual examples of DS2 on, from left to right, *bottle*, *cable*, and *capsule* categories. The first six rows show successful cases, and the last two rows show failed cases. The *gt* stands for ground truth, where anomalous parts are highlighted by a red mask.

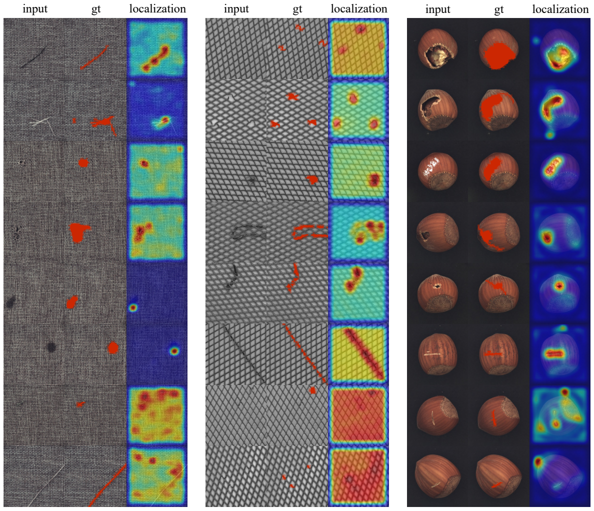

Figure 2. More localization visual examples of DS2 on, from left to right, *carpet*, *grid*, and *hazelnut* categories. The first six rows show successful cases, and the last two rows show failed cases. The *gt* stands for ground truth, where anomalous parts are highlighted by a red mask.

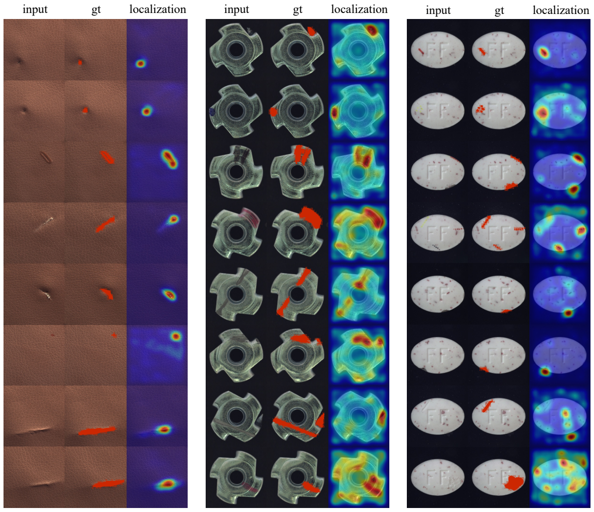

Figure 3. More localization visual examples of DS2 on, from left to right, *leather*, *metal nut*, and *pill* categories. The first six rows show successful cases, and the last two rows show failed cases. The *gt* stands for ground truth, where anomalous parts are highlighted by a red mask.

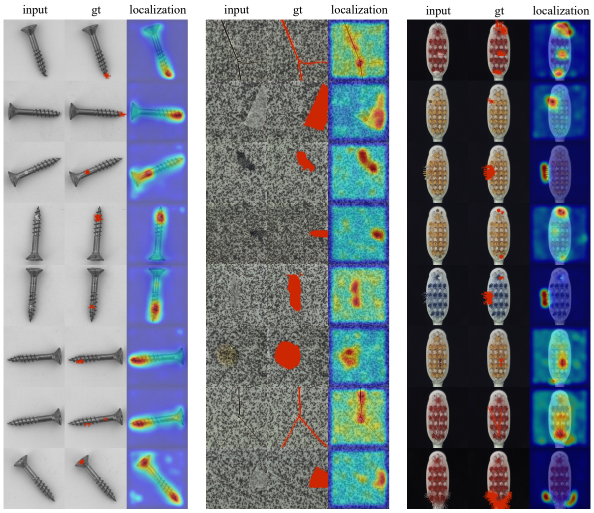

Figure 4. More localization visual examples of DS2 on, from left to right, *screw*, *tile*, and *toothbrush* categories. The first six rows show successful cases, and the last two rows show failed cases. The *gt* stands for ground truth, where anomalous parts are highlighted by a red mask.

## References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 2

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch,

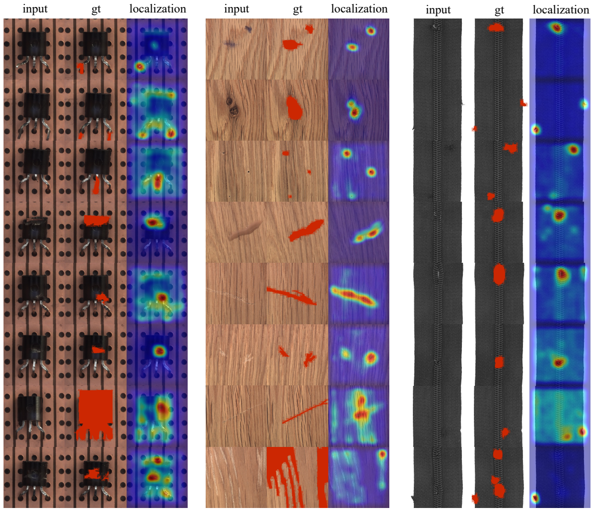

Figure 5. More localization visual examples of DS2 on, from left to right, *transistor*, *wood*, and *zipper* categories. The first six rows show successful cases, and the last two rows show failed cases. The *gt* stands for ground truth, where anomalous parts are highlighted by a red mask.

Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 1

[9] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, 2021. 1

[10] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022. 2

[11] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021. 2

[12] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *CVPR*, pages 16590–16599, 2022. 2

[13] Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Learning unsupervised metaformer for anomaly detection. In *ICCV*, pages 4369–4378, October 2021. 2

[14] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, pages 16684–16693, 2021. 1, 2