SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation - Supplementary Materials

Yang Zou[1], Jongheon Jeong[2⋆], Latha Pemula[1]
Dongqing Zhang[1], Onkar Dabeer[1]

[1]AWS AI Labs [2]KAIST
{yanzo,lppemula,zdongqin,onkardab}@amazon.com, jongheonj@kaist.ac.kr

A Scatter Plots for AU-ROC

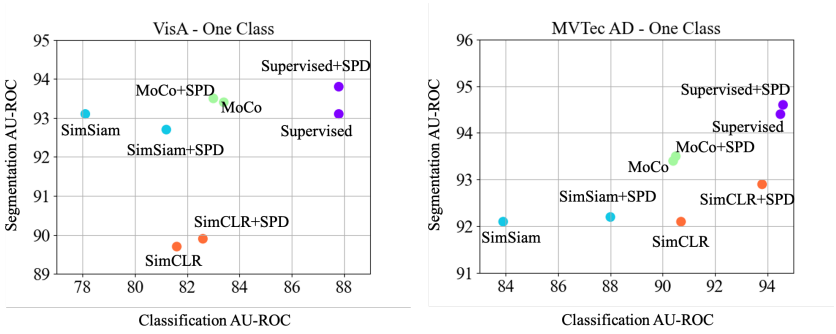

Fig. 8. Comparison of classification *vs.* segmentation AU-ROCs for various ImageNet pre-trained representations in the 1-class setup.

The scatter plots in terms of AU-PR for different ImageNet pre-trained models on 1-class training setups of VisA and MVTec-AD are shown in Fig. 7 of the main paper. In addition, we also show the scatter plots in terms of AU-ROC in Fig. 8. The SPD almost improves AU-ROC for all baselines on both classification and segmentation tasks of VisA and MVTec AD, demonstrating the effectiveness of the proposed SPD training.

B Further Discussion on AU-PR and AU-ROC

AU-ROC is a good metric for balanced dataset. However, as mentioned in several past works [4,5,8], in imbalanced dataset where minor class is more important, AU-ROC provides an inflated view of performance about the minor

⋆ work done during an Amazon internship

class and AU-PR is more informative. Imbalance is common in anomaly detection/segmentation datasets. Most experimental results in this work and [1] demonstrate the point. For example, considering the results of 1-cls segmentation on VisA, even when a model achieves $> 95\%$ AU-ROC, the AU-PR can be $< 10\%$. Moreover, the AU-ROC can be misleading about the performance on minor class in imbalanced dataset. Specifically, a model with a lower AU-ROC might be better than another model with higher AU-ROC in terms of the performance on anomaly class (reflected by AU-PR), although it might be worse in the major class. To give more intuition, we present the following toy example to demonstrate the above points.

A toy example: First, we denote P as ground truth positives, N as ground truth negatives, TP as True Positive, FP as False Positive, FN as False Negative, TN as True Negative. Then we define the following metrics.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} = \frac{N - TN}{N} \tag{3}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + 0.5 * (FP + FN)} \tag{4}$$

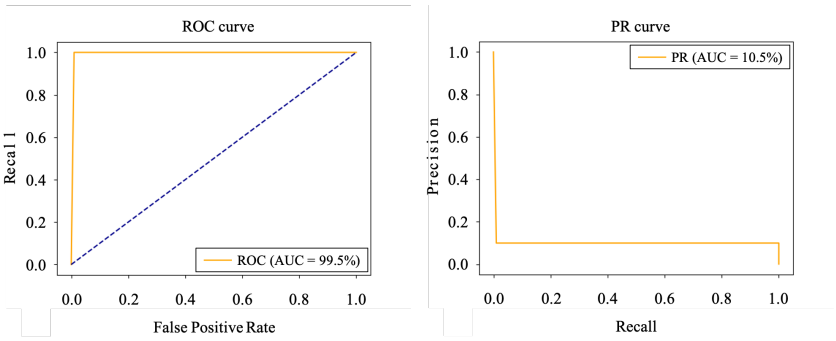

Fig. 9. ROC and PR curves for model A

Varying the operating thresholds, ROC curve measures the trade-off between recall and FPR and AU-ROC is the Area Under the ROC curve. PR curve measures trade-off between precision and recall and AU-PR is the Area Under the PR curve. Max F_1 is the best F_1 that can be obtained from the PR curve.

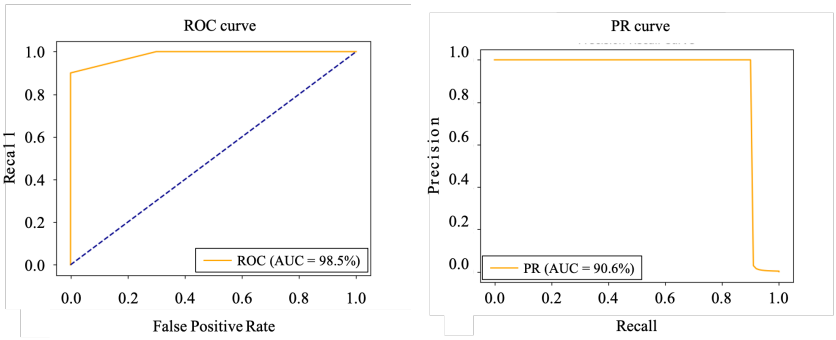

Fig. 10. ROC and PR curves for model B

Then we define a toy test set with 100 gt positives (anomaly), and gt 100,000 negatives (normal). This test set is imbalanced with neg/pos ratio = 1,000. Note that the scores of gt positive samples are the thresholds deciding the shapes of ROC and PR curves. In the following, we define model A and model B.

Model A has the following behaviors on the test set. To correctly predict each TP when threshold reduces, additional 10 FPs will be produced, leading to (TP, FP) pairs (1, 10), (2, 20), ..., (100, 1,000). We plot corresponding ROC and PR curves in Fig. 9. The AU-ROC=99.5% which seems to indicate the model is close to perfection. However, the AU-PR is just 10.5% and the Max F_1 is 18.2%. At the best threshold, model A only has 10% precision with 100% recall. Although the AU-ROC is inflatedly high, model A has a poor performance in predicting the positive class.

For model B, to correctly predict the first 90 TPs, there are no FPs. But in the remaining 10 positive samples, 3,000 FPs will be produced for each TP. So the (TP, FP) pairs are (1, 0), (2, 0), ..., (90, 0), (91, 3,000), (92, 6,000),...,(100, 30,000). We plot corresponding ROC and PR curves in Fig. 10.

Comparing with model A, model B has a worse AU-ROC (98.5% v.s. 99.5%). However, comparing at the best operating point, model B is much better than model A in predicting positive samples and achieves 100% precision and 90$ recall (v.s. 10% precision and 100% recall). Model B reaches 90.6% AU-PR and 94.7% Max F_1 which are much better than model A's 10.5% AU-PR and 18.2% Max F_1. In such case, the AU-ROC provides inflated and misleading view about model performance in positive predictions.

C Implementation Details

Pre-training: First, we set the SPD loss weight $\eta = 0.1$ for all the experiments unless specified otherwise. Second, for each baseline (SimSiam [3], MoCo [7], SimCLR [2]) with SPD, we follow exactly the same default hyperparamters in the baseline. Third, for SmoothBlend, the area of the cut patch is $0.5\% - 1\%$ in relative to the full image's size. The cut patch's aspect ratio ranges from 0.3

to 3. The standard deviations of kernel in Gaussian smoothing applied to the α mask are $(8, 8)$. Each image only has one smoothly blended patch. Fourth, for weak global augmentations, we choose random cropping with a ratio $[0.9, 1.0]$, color jittering with brightness= 0.1, contrast= 0.1, saturation= 0.1, hue= 0.05, Gaussian blur with standard deviations $(0.1, 0.3)$, horizontal flipping. Note that the color jittering and Gaussian blur are applied randomly with 0.8 and 0.5 probability.

Downstream anomaly detection and segmentation models: For PaDiM, we follow exactly the same hyperparameters in [6]. For two-class supervised networks, in high-shot setups, we fine-tune the models for 80 epochs with SGD with learneable backbone parameters. In few-shot setups, we train the models for $1,000$ iterations for few-shot classification and 500 iterations for few-shot segmentation. We choose a fixed learning rate policy with $lr = 0.0001$.

D Full results for each subset of VisA

In this section, we present the results for each subset of VisA w.r.t. SimSiam, SimSiam+SPD, supervised and supervised+SPD pre-training. Tables 7 and 8 provide the results for 1-class classification and segmentation training setups with PaDiM. Tables 9 and 10 present the results for 2-class high-shot classification and segmentation training setups. Tables 11 and 13 show the results for 2-class 5-shot classification and segmentation training setups. Tables 12 and 14 give the results for 2-class 10-shot classification and segmentation training setups. Generally speaking, the VisA subsets with multiple instances (Macaroni1, Macaroni2, Capsules, Candles) are the most difficult cases with lowest scores. The VisA subsets with complex structures (PCB1, PCB2, PCB3, PCB4) are relatively easier than the multiple instances cases with better scores. The VisA subsets with single instance (Cashew, Chewing gum, Fryum, Pipe Fryum) are easier than the complex structure cases.

Table 7. 1-class anomaly detection on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	83.5	85.4	83.5	86.8	89.5	92.0	90.4	92.7
	PCB2	76.1	76.6	76.9	76.6	88.5	89.1	87.0	87.9
	PCB3	73.0	75.0	72.0	72.2	87.8	87.8	86.9	85.4
	PCB4	92.3	93.9	93.8	95.2	98.3	98.5	98.9	99.1
Multiple instances	Macaroni1	67.7	72.2	74.4	75.7	81.3	84.8	82.2	85.7
	Macaroni2	56.1	59.2	62.3	66.8	63.6	69.8	66.7	70.8
	Capsules	70.4	58.1	72.5	62.0	74.8	65.8	78.6	68.1
	Candles	81.1	83.9	83.7	85.3	86.5	88.9	86.2	89.1
Single instance	Cashew	90.5	79.7	93.8	86.6	95.7	90.6	95.8	90.5
	Chewing gum	92.8	90.1	98.2	96.7	99.6	99.2	99.7	99.3
	Fryum	89.0	81.5	89.4	83.6	94.2	89.8	93.7	89.8
	Pipe fryum	89.9	81.6	93.6	87.1	98.5	97.2	97.6	95.6
	Mean	80.2	78.1	82.84	81.2	88.2	7.80	88.6	87.8

Table 8. 1-class anomaly segmentation on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	9.8	94.8	13.1	96.4	18.2	97.3	21.1	97.7
	PCB2	9.8	96.3	9.6	96.3	12.4	97.4	11.8	97.2
	PCB3	10.2	96.0	10.5	96.2	14.8	96.1	15.7	96.7
	PCB4	6.4	88.2	8.5	86.7	12.0	88.4	11.0	89.2
Multiple instances	Macaroni1	2.9	97.9	3.4	97.7	5.0	98.8	4.0	98.8
	Macaroni2	0.3	93.9	0.6	94.3	0.8	95.6	1.0	96.0
	Capsules	1.3	84.3	2.7	87.5	2.8	83.8	3.2	86.3
	Candles	3.5	95.6	3.5	93.7	6.7	96.8	7.3	97.3
Single instance	Cashew	9.9	88.8	9.5	86.3	10.0	85.5	8.7	86.1
	Chewing gum	29.7	97.3	28.5	97.0	29.2	96.1	31.3	96.9
	Fryum	11.6	90.1	11.7	89.0	12.1	88.2	11.9	88.0
	Pipe fryum	13.2	94.4	11.7	91.6	12.5	93.3	16.7	95.4
	Mean	9.1	93.1	9.4	92.7	11.4	93.1	12.0	93.8

Table 9. 2-class high-shot anomaly detection on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	79.8	96.7	84.9	98.4	89.9	98.8	93.4	99.4
	PCB2	95.9	98.8	96.9	99.3	98.2	99.7	96.9	99.3
	PCB3	86.0	98.0	94.0	99.1	99.8	100.0	99.4	99.9
	PCB4	98.7	99.8	99.7	100.0	99.1	99.9	99.9	100.0
Multiple instances	Macaroni1	88.2	98.5	95.0	99.3	99.5	99.9	99.9	100.0
	Macaroni2	82.4	97.2	93.2	98.8	99.9	100.0	100.0	100.0
	Capsules	70.4	91.6	76.6	93.7	88.4	97.0	94.3	98.6
	Candles	89.5	98.4	89.8	98.1	97.8	99.6	98.2	99.7
Single instance	Cashew	80.0	97.0	92.7	98.7	99.1	99.8	98.4	99.7
	Chewing gum	98.1	99.4	98.4	99.5	99.6	99.8	100.0	100.0
	Fryum	99.5	99.9	99.6	99.9	99.5	99.9	99.6	99.9
	Pipe fryum	95.8	99.5	97.7	99.7	99.2	99.9	99.3	99.9
	Mean	88.7	97.9	93.2	98.7	97.5	99.5	98.3	99.7

Table 10. 2-class high-shot anomaly segmentation on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	79.6	98.6	85.1	99.4	59.4	93.6	91.3	99.2
	PCB2	32.2	95.2	31.5	95.4	52.1	98.2	66.6	98.4
	PCB3	14.4	91.6	33.7	96.8	51.1	98.1	59.0	99.3
	PCB4	57.6	99.2	66.5	99.4	71.4	98.7	77.4	99.2
Multiple instances	Macaroni1	32.9	99.8	35.1	99.6	48.1	99.5	52.6	99.9
	Macaroni2	16.0	95.8	22.5	96.2	25.0	89.6	30.2	94.9
	Capsules	74.2	98.1	80.1	98.6	81.9	98.4	90.2	99.2
	Candles	18.6	93.6	22.7	93.4	46.5	98.1	51.5	98.3
Single instance	Cashew	76.4	98.0	83.3	99.5	85.5	97.5	86.8	98.8
	Chewing gum	84.1	99.4	86.2	99.4	89.2	99.4	90.1	99.3
	Fryum	81.9	98.7	89.0	99.8	85.1	98.8	85.4	98.6
	Pipe fryum	77.8	99.2	81.1	99.6	86.1	97.8	81.7	97.8
	Mean	53.8	97.3	59.7	98.1	65.1	97.3	71.9	98.5

Table 11. 2-class 5-shot anomaly detection on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	40.3	78.5	47.0	89.8	55.6	91.7	59.8	92.7
	PCB2	47.3	74.8	46.4	79.2	42.4	77.6	65.0	83.7
	PCB3	27.4	71.0	24.7	68.0	24.5	67.4	30.6	71.5
	PCB4	51.4	92.7	71.3	96.6	77.9	96.6	69.8	95.9
Multiple instances	Macaroni1	43.8	86.9	51.9	89.9	50.3	90.6	40.8	85.9
	Macaroni2	12.5	59.3	12.2	58.1	13.2	61.2	14.8	63.4
	Capsules	33.6	69.8	32.3	65.9	32.9	65.6	29.3	63.2
	Candles	53.8	88.9	67.5	92.0	67.7	92.5	69.7	93.0
Single instance	Cashew	79.9	96.9	80.2	96.8	81.0	96.8	80.7	96.8
	Chewing gum	49.1	72.5	52.3	76.2	74.1	89.0	67.5	87.5
	Fryum	97.4	99.2	98.6	99.6	97.8	99.4	96.8	99.3
	Pipe fryum	86.6	96.7	88.5	96.3	93.5	98.1	92.3	97.6
	Mean	51.9	82.3	56.1	84.0	59.2	85.5	59.8	85.9

Table 12. 2-class 10-shot anomaly detection on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	74.7	95.5	74.5	95.5	68.0	94.2	76.7	96.1
	PCB2	53.7	79.4	60.7	85.8	73.3	89.2	71.5	88.7
	PCB3	37.1	81.0	48.2	87.9	47.4	86.2	41.2	82.9
	PCB4	62.1	95.6	73.0	97.0	76.2	97.0	71.4	96.6
Multiple instances	Macaroni1	59.2	92.3	60.9	93.5	69.9	95.3	68.4	94.2
	Macaroni2	19.9	65.6	18.3	67.9	16.8	69.5	30.4	76.9
	Capsules	50.2	82.0	45.9	78.5	48.8	80.4	45.1	80.6
	Candles	69.7	93.8	72.5	94.0	79.2	95.7	79.5	95.9
Single instance	Cashew	78.3	96.6	79.0	96.8	78.5	96.6	81.4	97.1
	Chewing gum	80.4	92.7	83.7	94.2	91.2	97.1	92.9	97.3
	Fryum	98.4	99.3	98.1	99.5	98.7	99.6	98.5	99.7
	Pipe fryum	96.8	99.5	96.5	99.4	96.9	99.3	97.3	99.5
	Mean	65.0	89.4	67.6	90.8	70.4	91.7	71.2	92.1

Table 13. 2-class 5-shot anomaly segmentation on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	13.4	69.8	13.8	73.3	2.3	66.8	2.4	66.4
	PCB2	2.9	69.2	1.0	67.6	2.5	55.3	2.9	64.7
	PCB3	8.5	69.8	11.3	65.0	14.4	72.7	9.7	66.9
	PCB4	34.9	81.4	31.9	82.3	31.1	83.1	39.1	84.1
Multiple instances	Macaroni1	2.3	80.1	6.5	83.8	8.7	84.8	7.1	86.4
	Macaroni2	0.2	80.4	1.1	81.9	0.7	80.0	0.3	80.2
	Capsules	11.3	68.5	14.0	67.0	7.3	63.1	12.4	70.1
	Candles	5.8	75.9	7.2	71.0	4.3	73.8	6.6	73.8
Single instance	Cashew	24.9	78.6	23.3	87.2	21.6	77.3	22.8	76.9
	Chewing gum	70.0	90.7	70.7	93.0	72.7	96.7	71.0	96.0
	Fryum	6.3	67.2	5.8	58.3	3.9	55.7	7.5	62.1
	Pipe fryum	26.9	71.1	31.6	81.3	44.4	86.2	42.2	83.2
	Mean	17.3	75.2	18.2	76.0	17.8	74.6	18.7	75.9

Table 14. 2-class 10-shot anomaly segmentation on VisA.

		SimSiam		+SPD		Supervised		+SPD	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Complex structure	PCB1	17.1	72.1	24.5	80.7	6.5	66.5	5.3	58.5
	PCB2	12.5	63.3	7.7	81.7	11.4	72.9	14.3	75.7
	PCB3	23.3	80.2	18.1	73.6	21.7	75.8	25.5	78.6
	PCB4	45.2	92.1	46.9	86.2	41.3	88.6	50.1	91.4
Multiple instances	Macaroni1	10.3	83.2	12.8	86.2	20.8	92.4	14.4	88.1
	Macaroni2	7.3	89.0	7.0	78.4	8.8	87.6	8.8	87.4
	Capsules	36.1	88.0	42.4	83.7	36.7	83.6	29.9	74.9
	Candles	11.6	71.9	18.0	82.0	13.1	79.9	14.9	84.0
Single instance	Cashew	32.0	85.1	30.1	86.6	43.7	84.4	37.1	86.2
	Chewing gum	76.0	96.0	78.9	94.8	77.0	97.5	81.7	97.1
	Fryum	29.7	74.3	27.9	76.5	20.2	70.5	32.2	69.3
	Pipe fryum	40.5	84.5	42.4	88.5	38.9	81.3	53.3	90.9
	Mean	28.5	81.6	29.7	83.2	28.3	81.8	30.6	81.8

E Qualitative Results

Attention maps: In Fig. 11, we show the qualitative results to demonstrate the effectiveness of SPD regularization. Based on GradCAM [9], we generate attention maps of anomalous images by regarding negative cosine similarity as the distance (loss) between the defective image and its nearest normal sample. High energy regions contribute mostly to the feature cosine distance. Compared to SimSiam, SPD helps the baseline model to be more sensitive to the defective regions, demonstrating the validity of proposed SPD learning.

Anomaly segmentation results: In Fig. 12, we show the segmentation results for PaDiM with SimSiam, SimSiam+SPD, supervised and supervised+SPD pre-

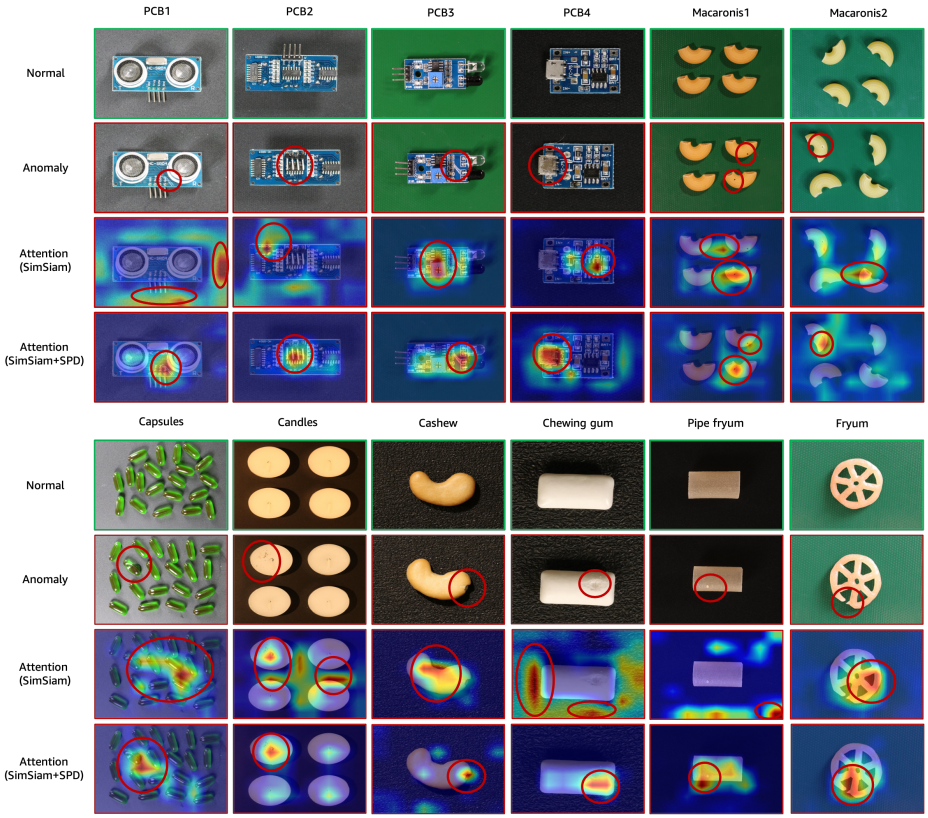

Fig. 11. Attention maps generated by GradCAM. 1st row: normal images; 2nd row: anomalous images; 3rd row: GradCAM based on SimSiam; 4th row: GradCAM based on SimSiam+SPD. Defects and high energy (red) parts in attentions are highlighted. Best viewed by zooming in.

trained ResNet-50. We can see the SPD gives better qualitative segmentation results than each baseline.

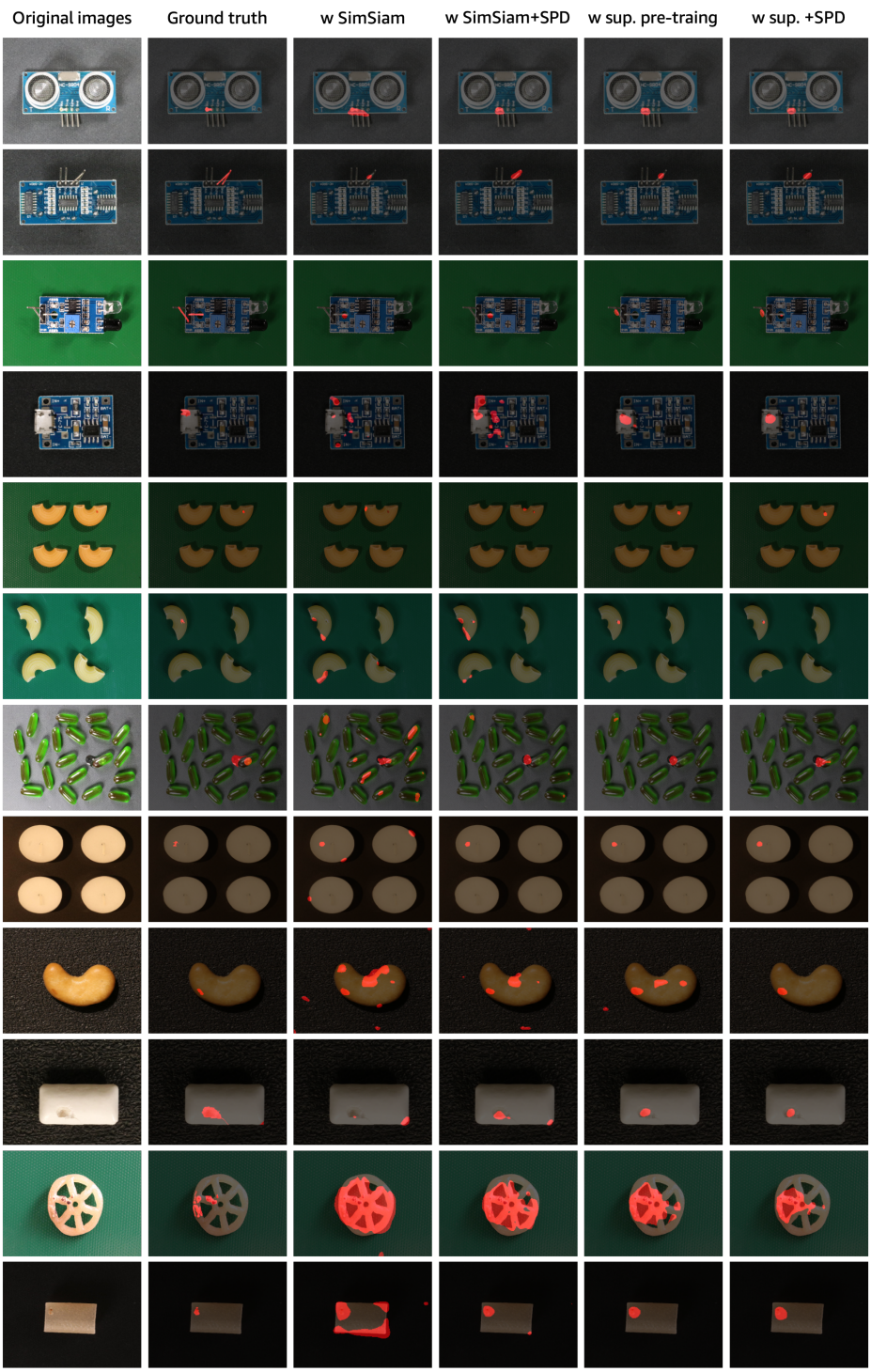

Fig. 12. Segmentation results from PaDiM with various pre-trained models.

References

1. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. International Journal of Computer Vision **129**(4), 1038–1059 (2021)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
3. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
4. Cook, J., Ramadas, V.: When to consult precision-recall curves. The Stata Journal **20**(1), 131–148 (2020)
5. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
6. Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDim: a patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition. pp. 475–489. Springer (2021)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
8. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), e0118432 (2015)
9. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)