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

Yang Zou , Jongheon Jeong , Latha Pemula,
Dongqing Zhang , Onkar Dabeer

AWS AI Labs, KAIST

ECCV 2022

Report: Yu-Chen Lai

Data: 2022.11.30

Outline

- Introduction
- SPot-the-Difference (SPD) Regularization
  1. Self-supervised Contrastive Learning
  2. Augmentations for SPD
  3. Training with SPD
- Visual Anomaly (VisA) Dataset
- Experiments
  1. Datasets
  2. SPD in high-shot 1-class/2-class Regimes
  3. SPD in Low-shot 2-class Regime
  4. Ablation Study
- Conclusions

1.1. Introduction

- Anomaly detection (AD) and segmentation for industrial manufacturing.
  1. Anomalies are **rare**.
  2. Anomalies are often **small** See figure (a).
  3. Manufacturing usually requires **highly accurate** models.
  4. Inspection in manufacturing spans a **wide range of domains and tasks**.

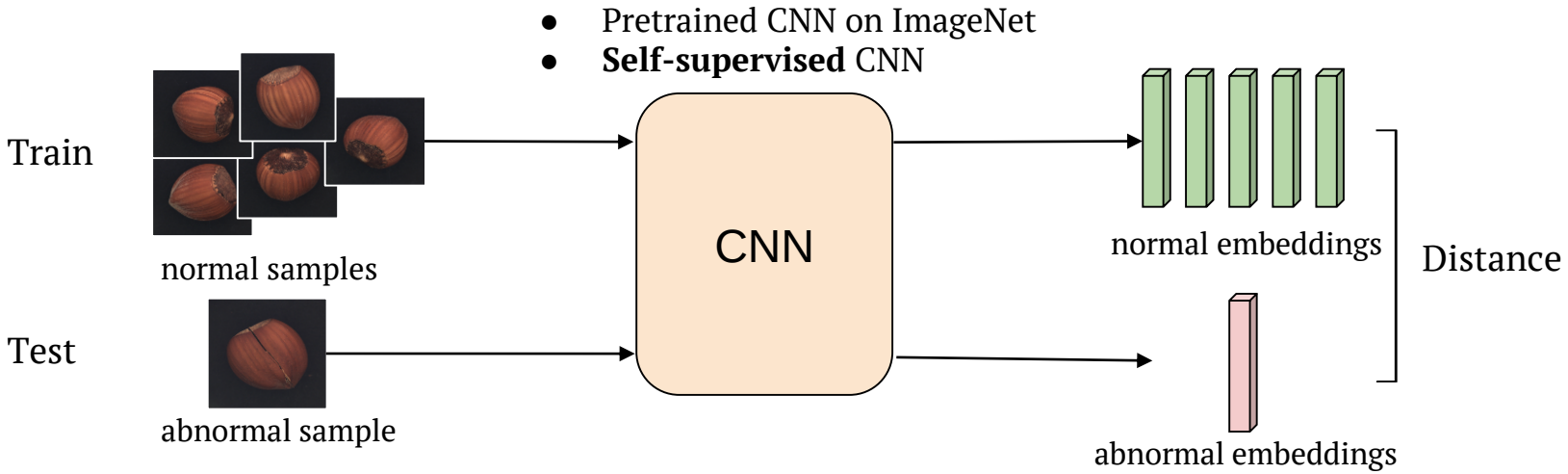
PCB1 – Normal

PCB1 – Anomaly

(a) Anomaly detection

1.1. Introduction

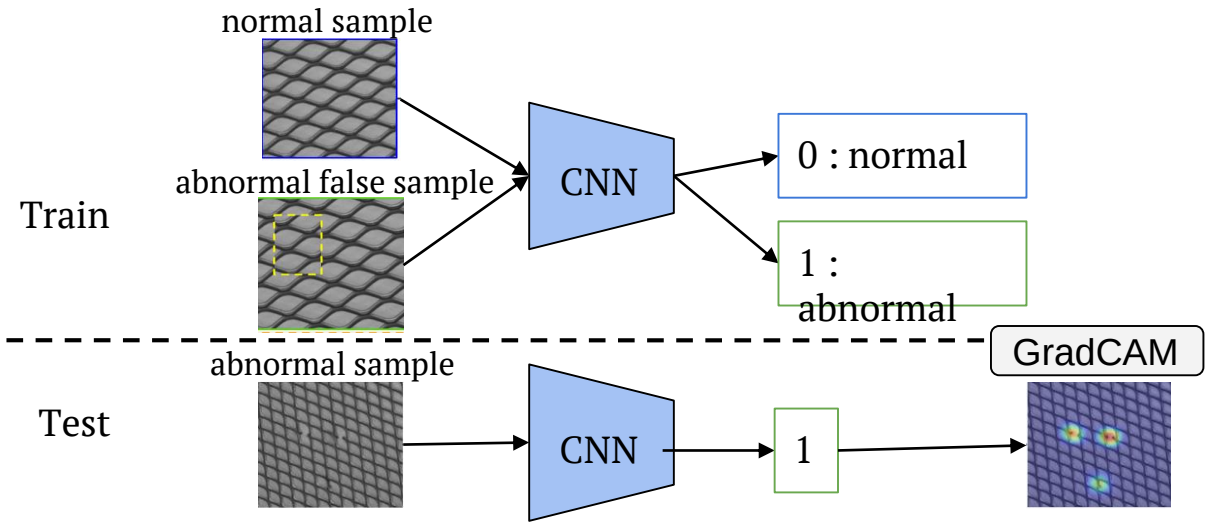
- **Embedding-based** AD model - e.g. PatchCore [33]
  - Only requires **normal images** for training.
  - Compare the **distance** between testing data and training(normal) data.



[33] *Roth, K., Pemula, L., Zepeda, J., Sch¨olkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2022)*

1.1. Introduction

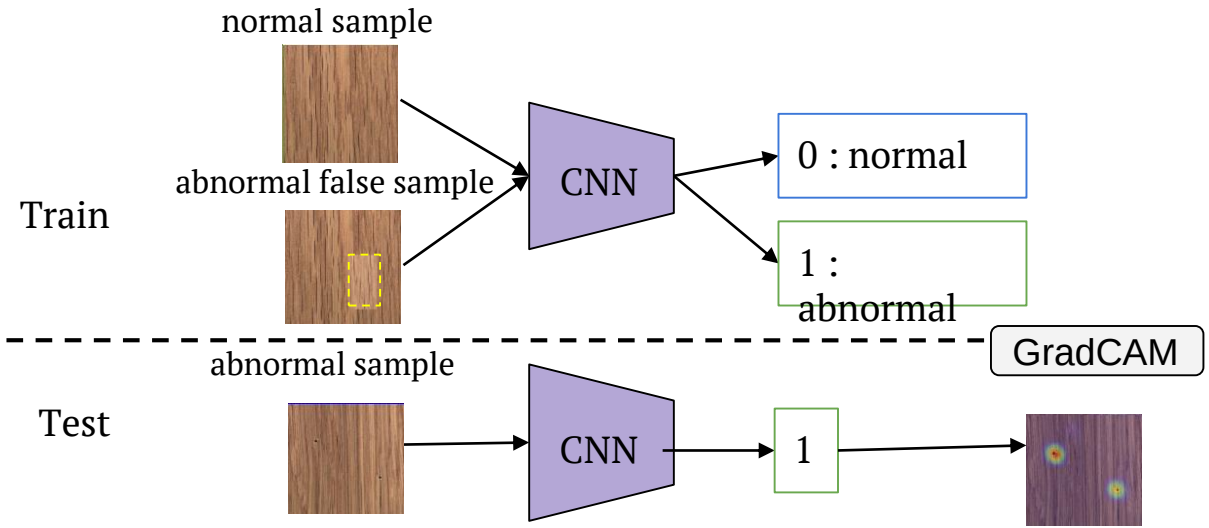
- Self-supervised Learning(SSL) for surface anomaly detection was explored in **CutPaste** [26] to learn representation from downstream images for each specific object.



[26] *Li, C.L., Sohn, K., Yoon, J., Pfister, T.: CutPaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674 (2021)*

1.1 Introduction

- Self-supervised Learning(SSL) for surface anomaly detection was explored in **CutPaste** [26] to learn representation from downstream images for each specific object.



[26] *Li, C.L., Sohn, K., Yoon, J., Pfister, T.: CutPaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674 (2021)*

1.1 Introduction

- Compare Self-supervised Contrastive Learning(SSCL) and SSL for surface anomaly detection.

	SSCL	SSL for surface anomaly detection
The features extracted from SSL	**high-level** semantic features	**low-level** texture features
Augmentation	grayscaling, large cropping, strong color jittering	
Sensitivity	**Global**	Global & **Local**

1.1 Introduction

- Inspired by the spot-the-difference puzzle, we propose a **contrastive SPot-the-Difference (SPD) training** to promote the local sensitivity of previous SSL methods.
- In the puzzle, players need to be sensitive to the subtle differences between the two globally alike images, which is similar to anomaly detection.

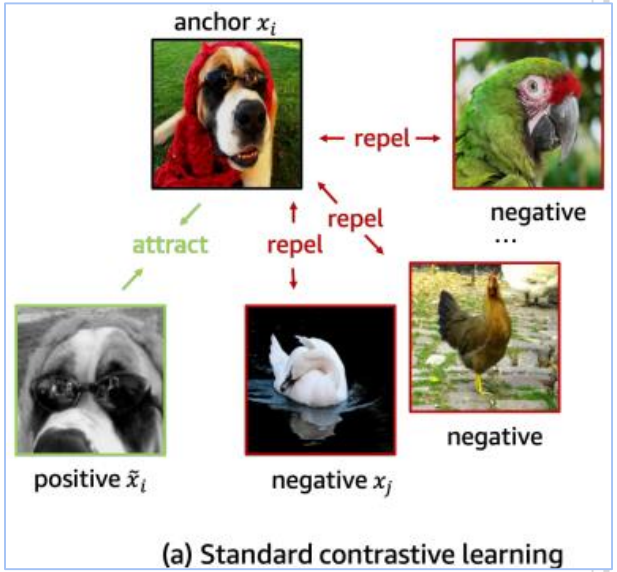
Image A

Image B

(b) Spot the difference

2.1 Self-supervised Contrastive Learning

- Many self-supervised learning methods, such as SimCLR [8] and MoCo [23], are based on contrastive learning.
  - **Maximize** the feature similarity between two
  - strongly augmented samples x_i and $\hat{x}_i$
  - **Minimizing** the similarities between the anchor x_i and other images x_j's.



[8] *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)*
[26] *Li, C.L., Sohn, K., Yoon, J., Pfister, T.: CutPaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674 (2021)*

2.1 Self-supervised Contrastive Learning

- Typically, an encoder extracts features h_i , $\hat{h}_i$ and h_j's which are inputs to a multilayer perceptron (MLP) head.
- The MLP head extracts the L2 normalized embeddings z_i , $\hat{z}_i$ and z_j's to compute the **InfoNCE loss** defined as follows.

$$\mathcal{L}_{\mathrm{NCE}}(x_i, \hat{x}_i) = -\log \frac{\exp\left(z_i \cdot \hat{z}_i / \tau\right)}{\exp\left(z_i \cdot \hat{z}_i / \tau\right) + \sum_{j=1}^{N} \mathbb{1}_{j \neq i} \exp\left(z_i, z_j / \tau\right)}$$

,where τ is a temperature scaling hyperparameter.

2.2 Augmentations for SPD

- Images augmented by most strong global transformations in SSL, such as grayscaling and large cropping, share semantics with anchor but with different local details.
- The features are forced to be **invariant about local details** and **capture the global semantics**.

anchor strong global aug.

2.2 Augmentations for SPD

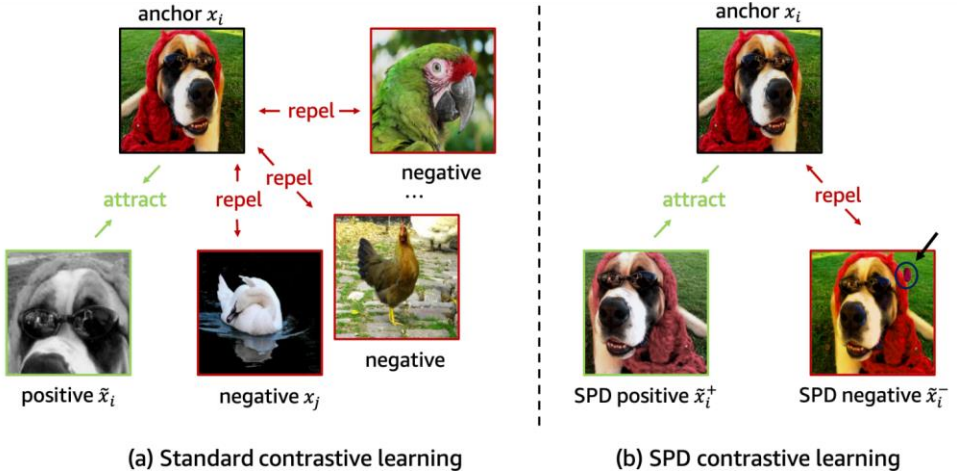
- Local augmentation
  - **SmoothBlend** is proposed to produce local deformations.
  - The augmented sample is obtained by $\bar{x} = (1 - \alpha) \odot x + \alpha \odot u$
    , where u is a cut patch with color jittering
    α is a mask with Gaussian blur corresponding to the pasted patch

anchor

Local aug.
- SmoothBlend

2.2 Augmentations for SPD

- Global augmentation
  - Using **weak augmentation**
  - It is a confusing if the network is designed to **maximize** the distance between negatives with only subtle changes while **minimizing** the distance between positives with largely global transformations.



2.2 Introduction

- Compare Self-supervised Contrastive Learning(SSCL) and for surface anomaly detection.

	SSCL	SSL for surface anomaly detection
The features extracted from SSL	**high-level** semantic features	**low-level** texture features
Augmentation	**Strong** augementation	**Weak** augementation & **SPD**(SmoothBlend)
Sensitivity	**Global**	Global & **Local**

2.3 Training with SPD

- x_i : an anchor image.
- $\tilde{x}_i^-$: the **negative** is generated by applying weak global augmentations followed by SmoothBlend.
- $\tilde{x}_i^+$: the **positive** is produced by weak global transformations only.

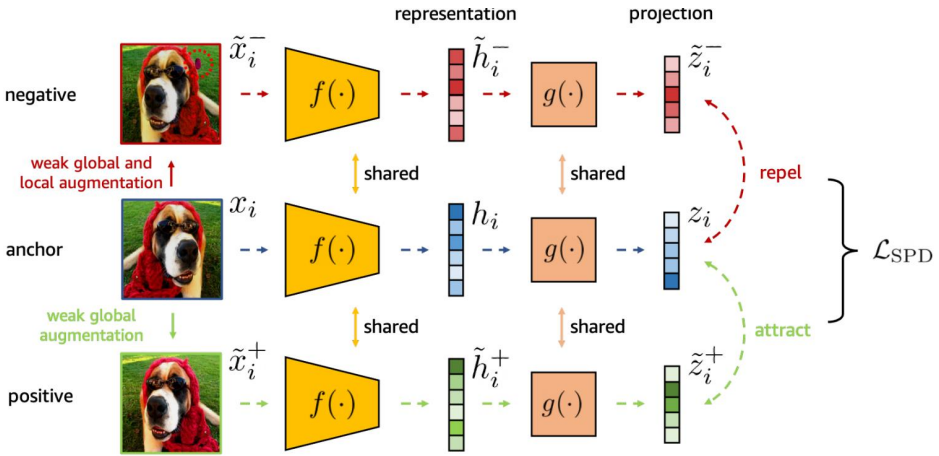


Fig. 4. The contrastive spot-the-difference learning

2.3 Training with SPD

- A shared feature extractor $f(\cdot)$ extracts the representations $h_i, \tilde{h}_i^-, \tilde{h}_i^+$.
- They're inputted into a shared MLP $g(\cdot)$ to get the projections $z_i, \tilde{z}_i^-, \tilde{z}_i^+$.

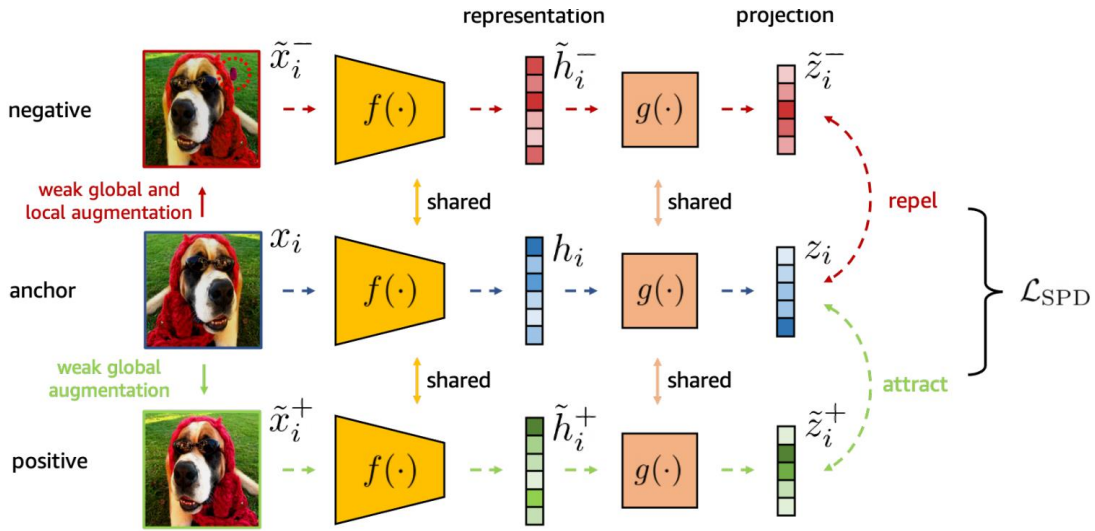


Fig. 4. The contrastive spot-the-difference learning

2.3 Training with SPD

- In summary, the SPD learning minimizes the following SPD loss

$$\mathcal{L}_{\text{SPD}}(x_i, \tilde{x}_i^-, \tilde{x}_i^+) = \cos(z_i, \tilde{z}_i^-) - \cos(z_i, \tilde{z}_i^+)$$

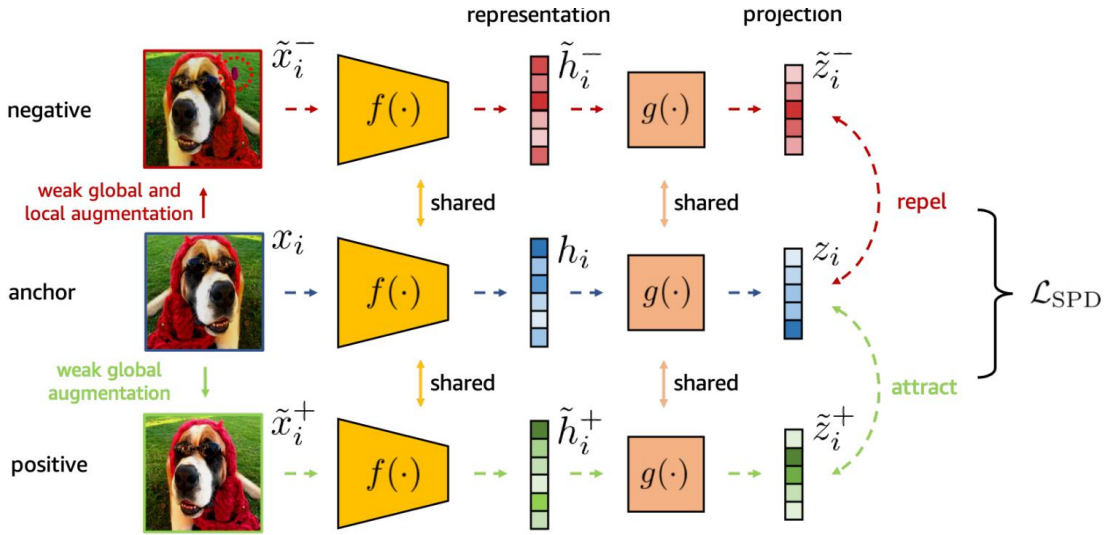


Fig. 4. The contrastive spot-the-difference learning

2.3 Standard contrastive SSL with SPD

- Example: SimCLR
  - the **anchor** x_i
  - **positive** $\hat{x}_i$ via strong global augmentations
  - other images x_j's in the same batch as **negatives**
- The network is trained by the following combined loss

$$\mathcal{L}(x_i, \hat{x}_i, \tilde{x}_i^-, \tilde{x}_i^+) = \mathcal{L}_{\text{NCE}}(x_i, \hat{x}_i) + \eta \cdot \mathcal{L}_{\text{SPD}}(x_i, \tilde{x}_i^-, \tilde{x}_i^+)$$

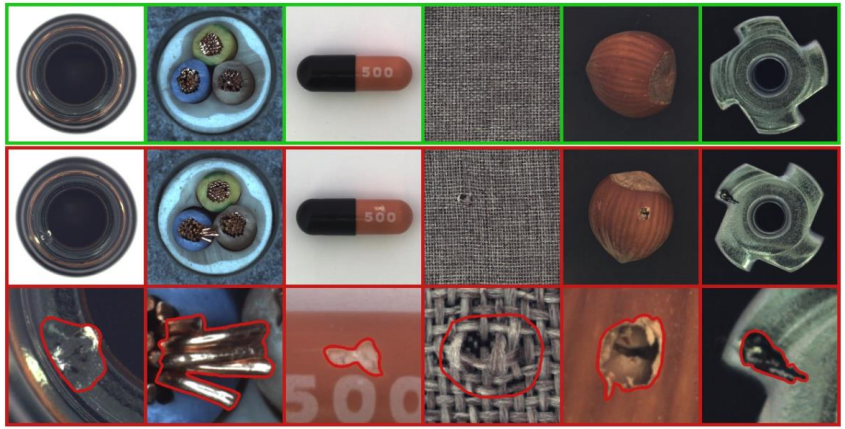
2.3 Standard supervised pre-training with SPD

- SPD could improve **local sensitivity**.
- Implement
  - **Auxiliary classifier**
    - Add on top of the last feature layer of the standard supervised model (ResNet-50)
    - To classify if an augmented SPD image has a local perturbation or not
  - **Cross-entropy**

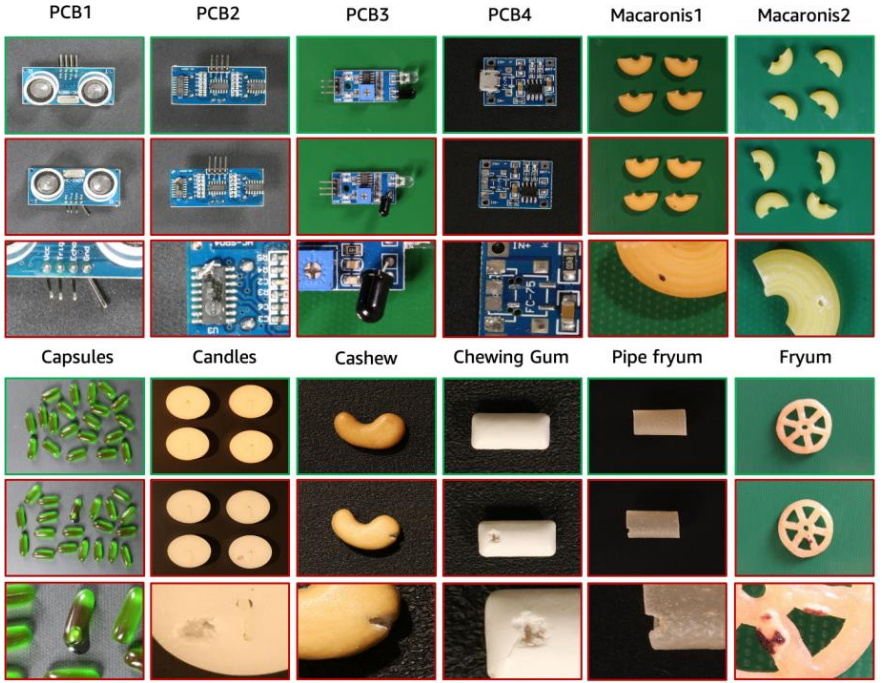
3. Visual Anomaly (VisA) Dataset

- Dataset Description
  - There are **10,821** images with 9,621 normal and 1,200 anomalous samples.
  - It spans 12 objects across 3 domains.
  - All images were acquired using a 4, 000 × 6, 000 high-resolution RGB sensor
  - **Larger** and **more complex** than MVTec-AD

3. Visual Anomaly (VisA) Dataset



MVTec-AD Dataset



VisA Dataset

4. Experiments

- Datasets
  - For self-supervised as well as supervised pre-training.
    - ImageNet 2012 classification dataset
  - For downstream tasks.
    - VisA dataset
    - MVTec-AD dataset
- Evaluation Metrics
  - AU-ROC
    - Seems to be close to perfection.
  - AU-PR
    - Far-from satisfactory.

$$FPR = FP / (FP + TN)$$
$$TPR = TP / (TP+ FN)$$

$$Recall(TPR) = TP / (TP+ FN)$$
$$Precision = TP / (TP+FP)$$

4. Experiments

- Anomaly detection and segmentation algorithms
  - 1-class anomaly classification/segmentation
    - PaDiM [14]
  - 2-class anomaly classification
    - ResNet
  - 2-class anomaly segmentation
    - U-Net
- Implementation details
  - Backbone : **ResNet-50**
  - Adopt exactly the same hyperparameters in **SimSiam**, **MoCo**, **SimCLR** and supervised learning for pre-training.

[14] *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)*

4.2. SPD in high-shot 1-class/2-class Regimes

- The results of **PaDiM** with various **pre-training options.**

Table 2. 1-class performance evaluation of various ImageNet pre-training options on VisA and MVTec-AD with PaDiM. Bold numbers refers to the highest score. In the brackets are the gaps to the ImageNet supervised/self-supervised pre-training counterpart. In green are the gaps of at least +0.5 point.

	ImageNet labels	VisA (1-class)				MVTec-AD (1-class)			
		Classification		Segmentation		Classification		Segmentation	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	✓	88.2	87.8	11.4	93.1	97.4	94.5	35.2	94.4
SimSiam	✗	80.2	78.1	9.1	93.1	92.6	83.9	29.7	92.1
+SPD	✗	82.8 (+2.6)	81.2 (+3.1)	9.4 (+0.3)	92.7 (-0.4)	94.1 (+1.5)	88.0 (+4.1)	32.0 (+2.3)	92.2 (+0.1)
MoCo	✗	83.6	83.4	10.5	93.4	95.0	90.4	33.2	93.4
+SPD	✗	84.1 (+0.5)	83.0 (-0.4)	11.0 (+0.5)	93.5 (+0.1)	95.6 (+0.6)	90.5 (+0.1)	33.5 (+0.3)	93.5 (+0.1)
SimCLR	✗	82.7	81.6	8.8	89.7	94.7	90.7	29.8	92.1
+SPD	✗	83.9 (+0.8)	82.6 (+1.0)	8.7 (-0.1)	89.9 (+0.2)	96.8 (+2.1)	93.8 (+3.1)	31.7 (+1.9)	92.9 (+0.8)
Sup. pre-train+SPD	✓	**88.6** (+0.4)	**87.8** (+0.0)	**12.0** (+0.6)	**93.8** (+0.7)	**97.5** (+0.1)	**94.6** (+0.1)	**36.3** (+1.1)	**94.6** (+0.2)

4.2. SPD in high-shot 1-class/2-class Regimes

- The results of **PaDiM** with various **pre-training options.**

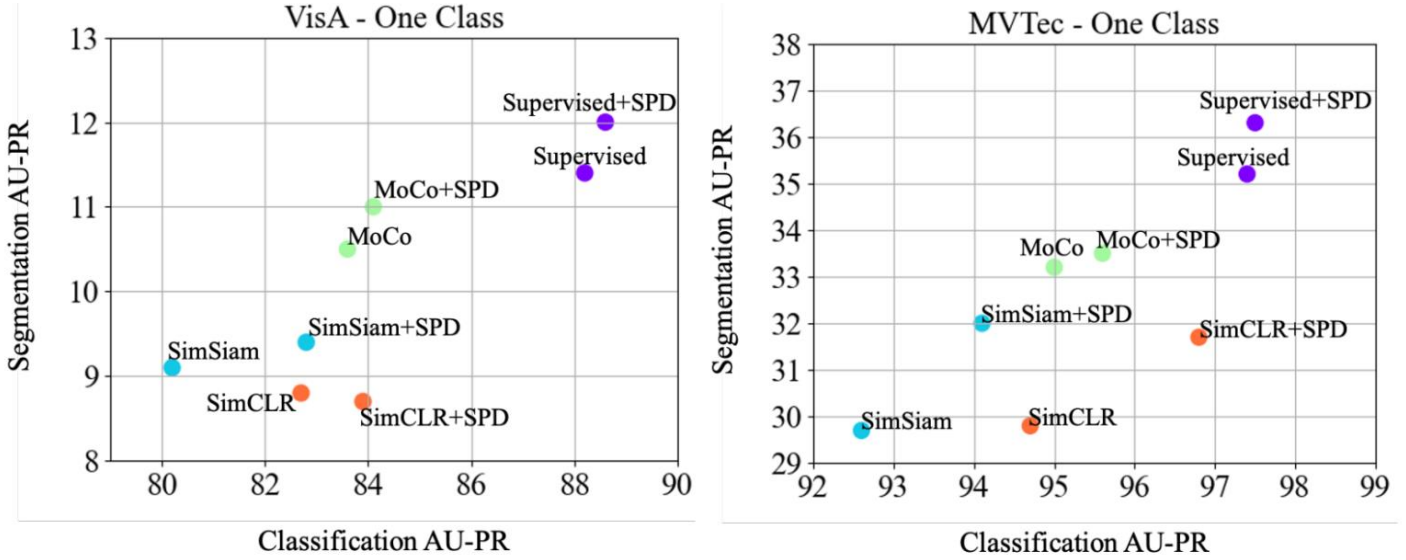


Fig. 7. Scatter plots for various ImageNet pre-training models in 1-class setup.

4.2. SPD in high-shot 1-class/2-class Regimes

- 2-class anomaly classification/segmentation

Table 3. 2-class fine-tuning with different pre-training on VisA high-shot setup.

	ImageNet labels	VisA (2-class, high-shot)			
		Classification		Segmentation	
		AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	✓	97.5	99.5	65.1	97.3
SimSiam	✗	88.7	97.9	53.8	97.3
+SPD	✗	93.2 (+4.5)	98.7 (+0.8)	59.7 (+5.9)	98.1 (+0.8)
MoCo	✗	93.9	98.8	62.4	98.0
+SPD	✗	94.2 (+0.3)	98.8 (+0.0)	64.4 (+2.0)	97.9 (-0.1)
SimCLR	✗	93.4	98.5	67.7	95.3
+SPD	✗	92.7 (-0.7)	98.6 (+0.1)	68.2 (+0.5)	95.7 (+0.4)
Sup. pre-train+SPD	✓	**98.3** (+0.8)	**99.7** (+0.2)	**71.9** (+6.8)	**98.5** (+1.2)

4.3. SPD in Low-shot 2-class Regime

- Low-shot anomaly classification/segmentation
  - A 2-class U-Net with ResNet-50 encoder

Table 4. Low-shot anomaly detection and segmentation on VisA.

	ImageNet labels	Classification (2-class, low-shot)				Segmentation (2-class, low-shot)			
		5-shot		10-shot		5-shot		10-shot	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	✓	59.2	85.5	70.4	91.7	17.8	74.6	28.3	81.8
SimSiam	✗	51.9	82.3	65.0	89.4	17.3	75.2	28.5	81.6
+SPD	✗	56.1 (+4.2)	84.0 (+1.7)	67.6 (+2.6)	90.8 (+1.4)	18.2 (+0.9)	**76.0** (+0.8)	29.7 (+1.2)	83.2 (+1.6)
MoCo	✗	56.1	83.8	68.7	90.6	21.5	80.5	32.3	**85.7**
+SPD	✗	56.4 (+0.3)	83.9 (+0.1)	68.0 (-0.7)	90.1 (-0.5)	**22.1** (+0.6)	78.5 (-2.0)	**32.8** (+0.5)	84.9 (-0.8)
SimCLR	✗	48.4	79.6	58.2	86.0	18.4	71.2	23.0	75.1
+SPD	✗	47.4 (-1.0)	79.9 (+0.3)	59.0 (+0.8)	86.1 (+0.1)	18.9 (+0.5)	74.5 (+3.3)	25.1 (+2.1)	78.2 (+3.1)
Sup. pre-train+SPD	✓	**59.8** (+0.6)	**85.9** (+0.4)	**71.2** (+0.8)	**92.1** (+0.4)	18.7 (+0.9)	75.9 (+1.3)	30.6 (+2.3)	81.8 (+0.0)

4.4. Ablation Study

- Based on ImageNet **SimSiam** pre-training
- **PaDiM** as the anomaly detection and segmentation algorithms

Table 5. Ablation study

	VisA (1-class)				MVTec-AD (1-class)			
	Classification		Segmentation		Classification		Segmentation	
	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
SimSiam w/ Res50	80.2	78.1	9.1	93.1	92.6	83.9	29.7	92.1
+SPD ($\eta = 0.1$)	82.8	81.2	9.4	92.7	94.1	88.0	32.0	92.2
+SPD ($\eta = 0.5$)	80.5	79.3	8.7	93.0	93.3	84.9	30.1	91.9
+SPD ($\eta = 1.0$)	81.5	79.8	9.4	92.8	93.4	85.8	30.0	92.0
+SPD w/ CutPaste	78.8	77.0	9.7	93.1	93.5	85.2	28.2	91.3
+SPD w/ Xent	71.4	66.6	2.7	84.8	86.3	71.0	15.2	82.6
SimSiam w/ WideRes50	80.3	77.7	9.9	93.6	93.0	84.7	31.3	92.2
+SPD	81.9	80.4	10.5	93.7	93.4	85.4	32.5	92.8

4.4. Ablation Study

- Sensitivity analysis on SPD **loss weight** η
- Comparison between SPD and **CutPaste**
- SPD with different **backbones**

Table 5. Ablation study

	VisA (1-class)				MVTec-AD (1-class)			
	Classification		Segmentation		Classification		Segmentation	
	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
SimSiam w/ Res50	80.2	78.1	9.1	93.1	92.6	83.9	29.7	92.1
+SPD ($\eta = 0.1$)	82.8	81.2	9.4	92.7	94.1	88.0	32.0	92.2
+SPD ($\eta = 0.5$)	80.5	79.3	8.7	93.0	93.3	84.9	30.1	91.9
+SPD ($\eta = 1.0$)	81.5	79.8	9.4	92.8	93.4	85.8	30.0	92.0
+SPD w/ CutPaste	78.8	77.0	9.7	93.1	93.5	85.2	28.2	91.3
+SPD w/ Xent	71.4	66.6	2.7	84.8	86.3	71.0	15.2	82.6
SimSiam w/ WideRes50	80.3	77.7	9.9	93.6	93.0	84.7	31.3	92.2
+SPD	81.9	80.4	10.5	93.7	93.4	85.4	32.5	92.8

4.4. Ablation Study

- Results with PatchCore[33]

Table 6. 1-class performance evaluation on VisA and MVTec-AD with PatchCore.

Backbone: Wide ResNet50	VisA (1-class)				MVTec-AD (1-class)			
	Classification		Segmentation		Classification		Segmentation	
	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	93.3	92.4	38.4	98.4	99.2	99.8	48.8	97.6
Sup. pre-train+SPD	93.8 (+0.5)	92.5 (+0.1)	39.3 (+0.9)	98.1 (-0.3)	99.0 (-0.2)	99.7 (-0.1)	49.3 (+0.5)	97.5 (-0.1)

- Extending SPD to other tasks
  - SPD also improves ImageNet supervised classification accuracy
    - 69.8% → 70.2% for ResNet-18
    - 76.1% → 76.4% for ResNet-50

4.4. Ablation Study

- Qualitative results
  - Attention maps of anomaly segmentation results

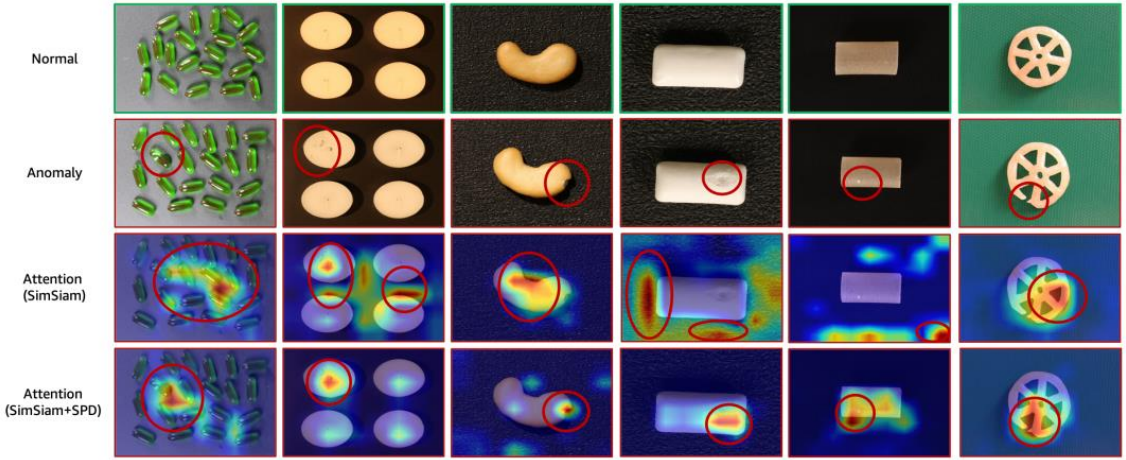


Fig. 11. Attention maps generated by GradCAM. 1st row: normal images; 2nd row: anomalous images; 3rd row: GradCAM based on SimSiam; 4th row: GradCAM based on SimSiam+SPD. Defects and high energy (red) parts in attentions are highlighted. Best viewed by zooming in.

5. Conclusions

- Spot-the-difference **(SPD)** training
  - Regularize pretrained models' **local sensitivity** to anomalous patterns.
  - **SimSiam+SPD** obtains superior or competitive performances in **low-shot** regime.
  - **Supervised learning+SPD** presents **better** performances in various setups.
- Visual Anomaly (**VisA**) dataset
  - The largest industrial anomaly detection dataset.

[33] *Roth, K., Pemula, L., Zepeda, J., Sch¨olkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2022)*

Thanks For Listening !