KAIST USC CMU

ECML PKDD 2023

DSV: An Alignment Validation Loss for Self-supervised Outlier Model Selection



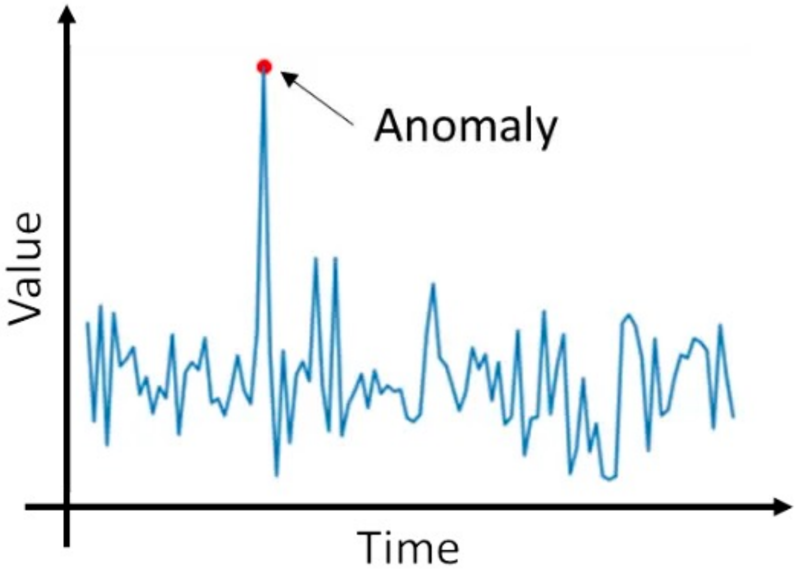

Jaemin Yoo[1] **Yue Zhao[2]** **Lingxiao Zhao[3]** **Leman Akoglu[3]**

Outline

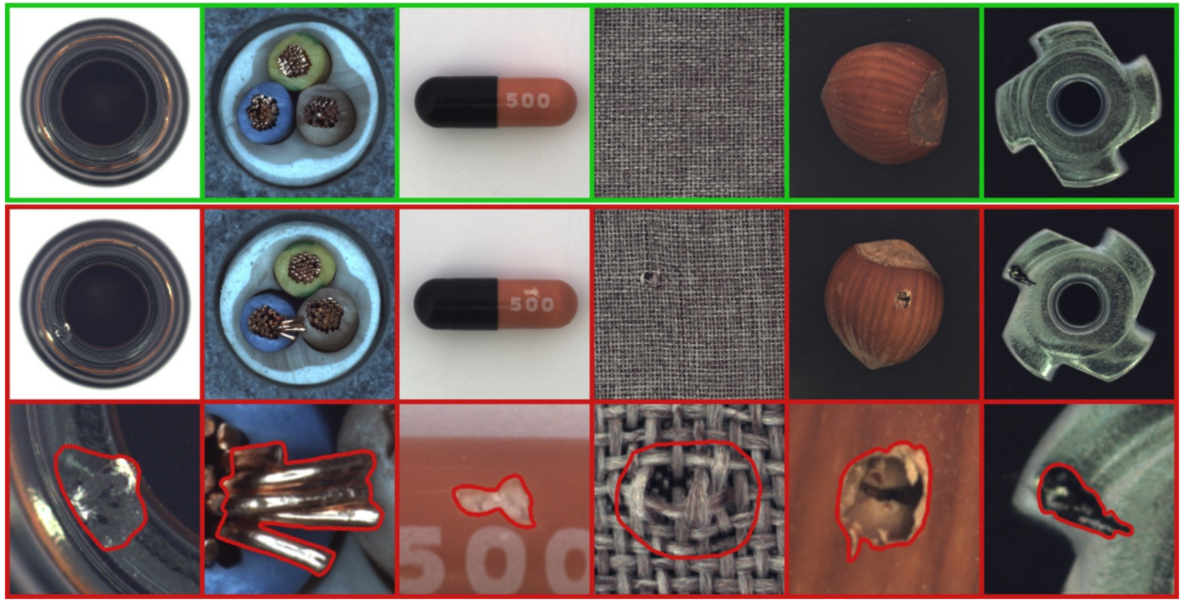
- <u>**Introduction**</u>
- Proposed Method
- Experiments
- Conclusion

Anomaly Detection

- **Anomaly detection (AD)** is to find **anomalies** from a set of data
  - **Unsupervised:** No information about actual anomalies



Source: Analytics Vidhya



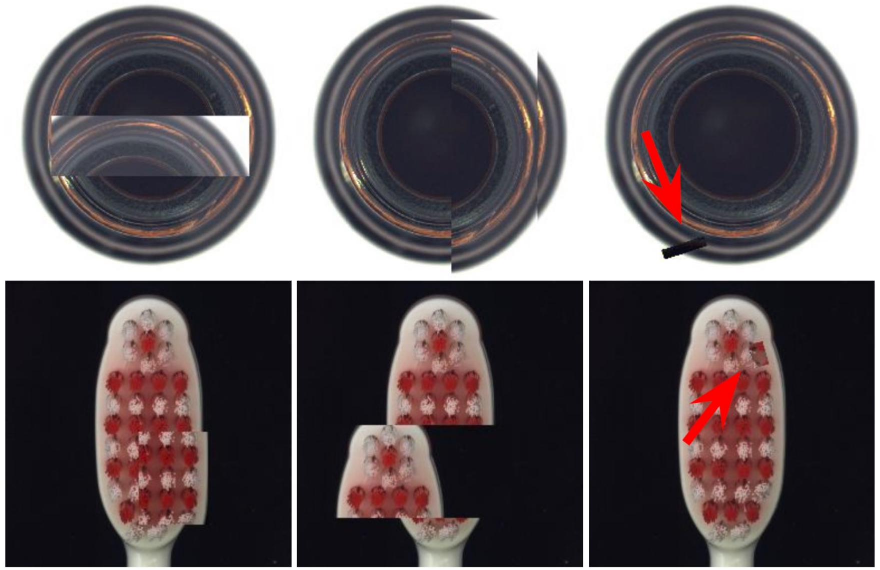
Source: https://www.mvtec.com/company/research/datasets/mvtec-ad

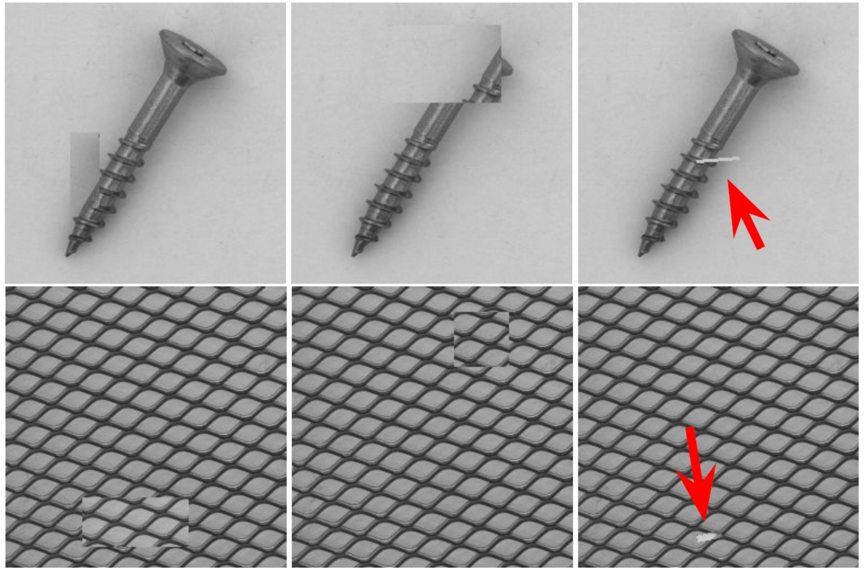
Self-supervised Anomaly Detection

- **Q:** How can we training an accurate detector without labels?
- **Self-supervised anomaly detection (SSAD)** is a promising direction
  - **Idea:** Generate pseudo anomalies with an augmentation function f_{aug}
- **How SSAD works:**
  - Create $\mathcal{D}_{\mathrm{aug}}$ by applying f_{aug} to normal data $\mathcal{D}_{\mathrm{trn}}$
  - Train a supervised classifier ϕ to classify between $\mathcal{D}_{\mathrm{trn}}$ and $\mathcal{D}_{\mathrm{aug}}$

SSAD: Example

- **CutPaste** (Li et al., 2021) is an example of f_{aug}
  - Cuts a random patch from an image and pastes into a different location
  - Generated images look like (local) defects in industrial object images





Li et al. "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization." CVPR 2021

Unsupervised Outlier Model Selection

- For anomaly detection, **model selection** is a crucial problem
- **Why?** No validation (or hold-out) data are given at training
- For SSAD, hyperparameters of f_{aug} are especially important
  - Since they determine the success and the failure of training

***Q:** How can we effectively perform augmentation HP search on SSAD?*

Problem Definition

- **Given**
  - Data augmentation function f_{aug} (e.g., CutOut or CutPaste)
  - Normal-only training data $\mathcal{D}_{\text{trn}}$
  - Unlabeled test data $\mathcal{D}_{\text{test}}$ containing both normal data and anomalies
  - Set $\{\phi_i\}_i$ of detector models trained by f_{aug} with different HPs
- **Goal:** Find the detector ϕ^* showing the highest accuracy on $\mathcal{D}_{\text{test}}$
- **Without** having any labels at the training time

Outline

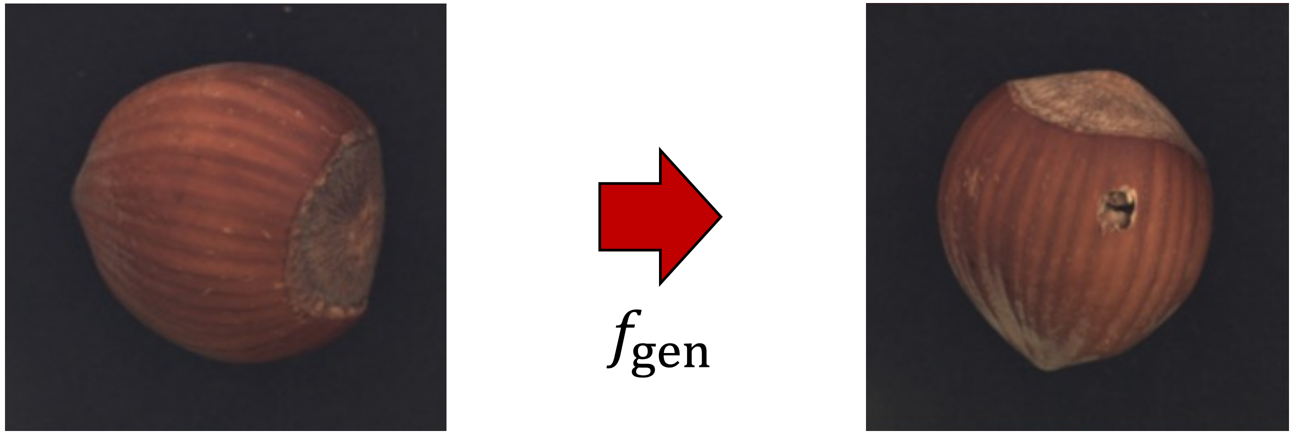
- Introduction
- **<u>Proposed Method</u>**
- Experiments
- Conclusion

DSV: Overview

- We propose **DSV (Discordance and Separability Validation)**
  - Unsupervised validation loss for HP search on SSAD
  - Measures the quality of f_{aug} without requiring any labels
- DSV consists of three main ideas:
  - **Main Idea 1:** Alignment as an embedding distance
  - **Main Idea 2:** Decomposition of the alignment
  - **Main Idea 3:** Surrogate losses without labeled data

Anomaly-Generating Function

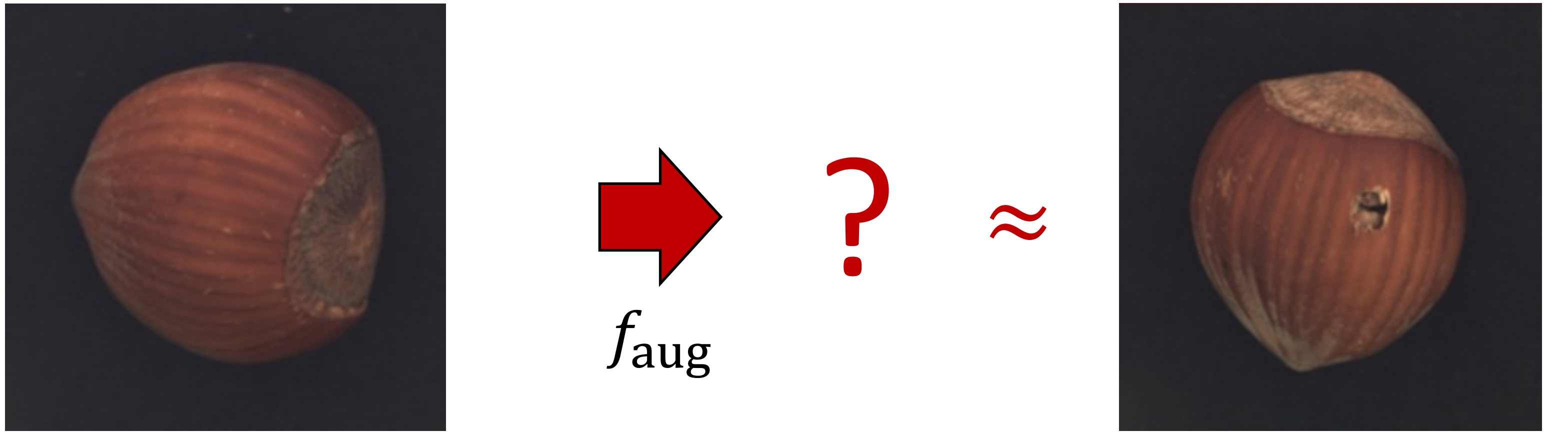
- Let f_{gen} be the anomaly-generating function underlying in $\mathcal{D}_{\mathrm{test}}$
  - Transforms a normal sample $\mathbf{x}$ into an anomaly $f_{\mathrm{gen}}(\mathbf{x})$
  - Hard to formally define in real data



Normal data   Anomaly

DSV: Goal

- **Goal:** Find HPs that make f_{aug} aligned with f_{gen} the most
- **Why?** Detector ϕ is trained to classify between $\mathbf{x}$ and $f_{\mathrm{aug}}(\mathbf{x})$
  - If f_{aug} and f_{gen} are similar, ϕ can detect $f_{\mathrm{gen}}(x)$ as well



Normal data

Anomaly

Main Idea 1

- **Q:** How can we measure the alignment between f_{aug} and f_{gen}?
- **Idea 1:** Measure the distance between embeddings generated by ϕ

$$\boxed{\mathcal{L}_{\mathrm{ali}} = d\left(\mathcal{Z}_{\mathrm{aug}}, \mathcal{Z}_{\mathrm{test}}^{(a)}\right)}$$

  - d: Distance function between sets of vectors
  - $\mathcal{Z}_{\mathrm{aug}}$: Set of embeddings for augmented training data
  - $\mathcal{Z}_{\mathrm{test}}^{(a)}$: Set of embeddings for test anomalies

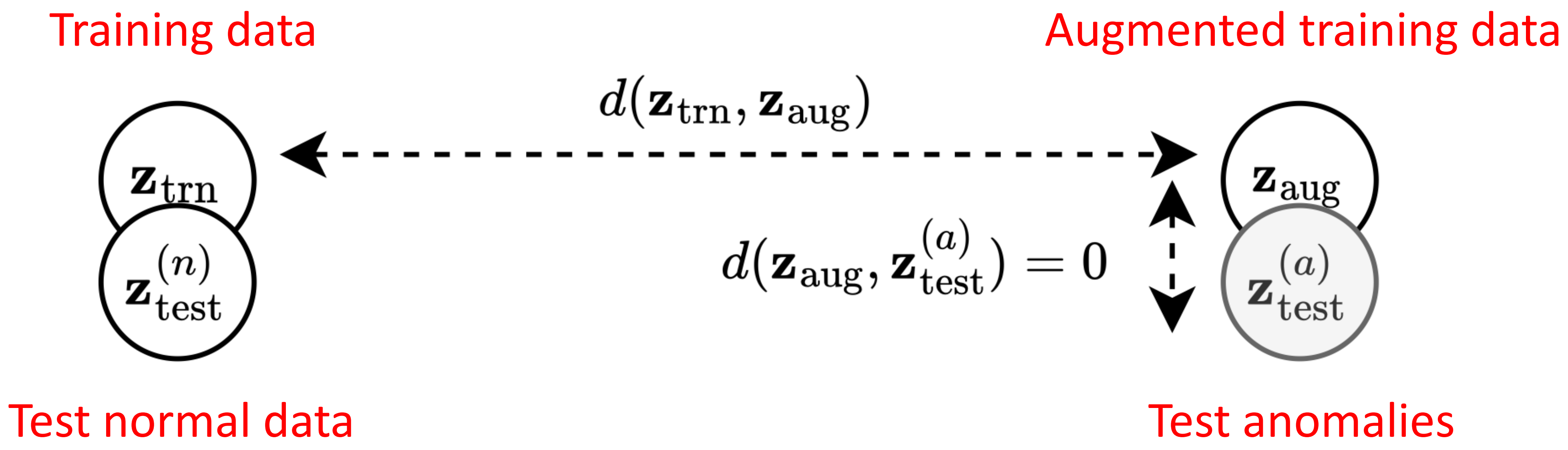
Main Ideas 2 & 3

$$\mathcal{L}_{\mathrm{ali}} = d\left(\mathcal{Z}_{\mathrm{aug}}, \mathcal{Z}_{\mathrm{test}}^{(a)}\right)$$

- **Q:** How can we approximate $\mathcal{L}_{\mathrm{ali}}$ without accessing $\mathcal{Z}_{\mathrm{test}}^{(a)}$?
- **Idea 2:** Decompose it into **discordance** and **separability**
  - They consider two different aspects of the alignment
- **Idea 3:** Design **surrogate losses** to estimate the two terms
  - Design surrogate losses to avoid using any labeled data

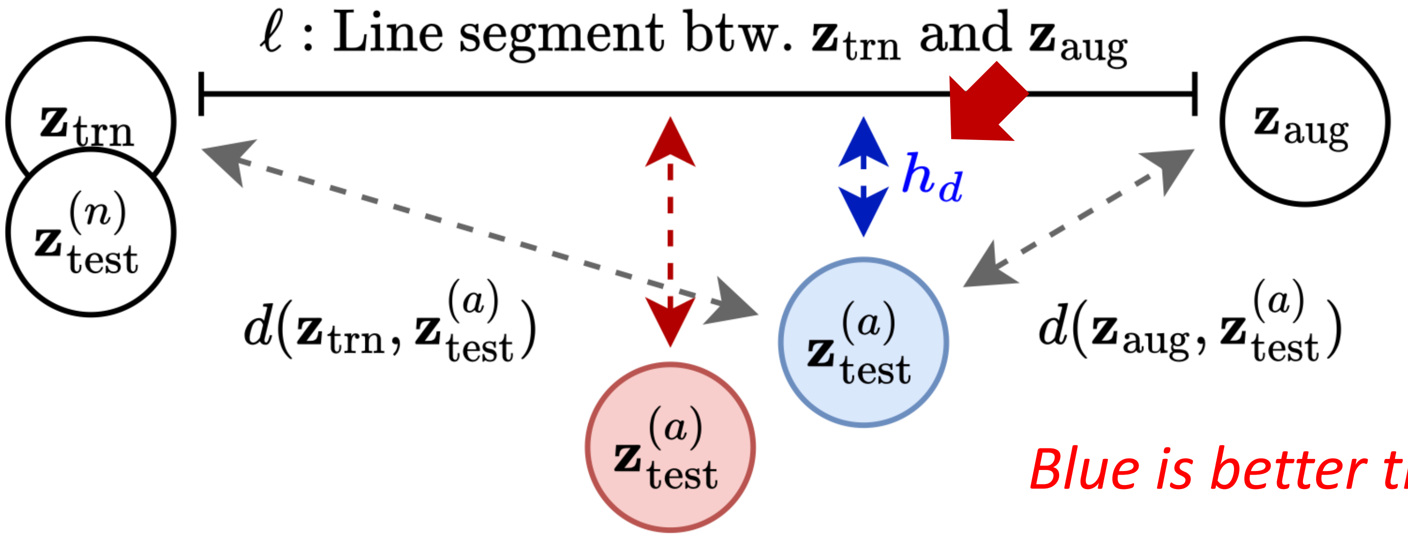
Main Idea 2: Visualization

- **Assumption:** All sets are of size one, e.g., $\mathcal{Z}_{\mathrm{trn}} = \{\mathbf{z}_{\mathrm{trn}}\}$
- We illustrate the case of perfect alignment as follows:



Main Idea 2: Discordance

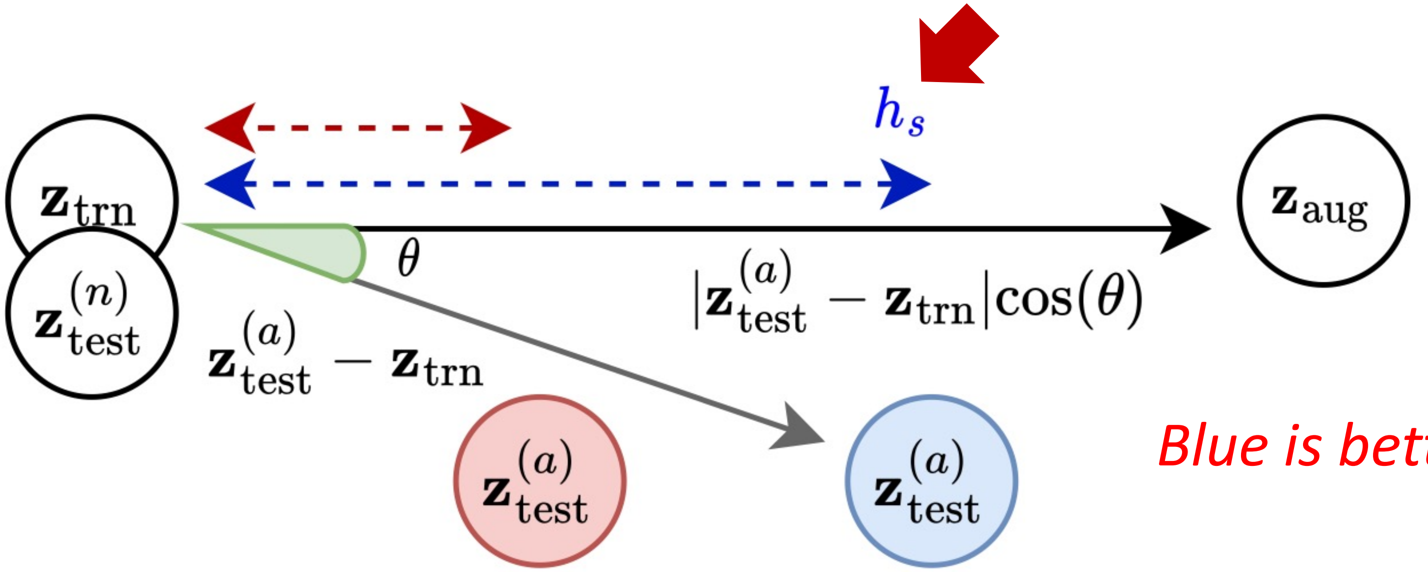
- Let ℓ be the line segment between $\mathcal{Z}_{\text{trn}}$ and $\mathcal{Z}_{\text{aug}}$
- **Discordance h_d** measures the distance between $\mathcal{Z}_{\text{test}}^{(a)}$ and ℓ



Blue is better than red

Main Idea 2: Separability

- Let ℓ be the line segment between $\mathcal{Z}_{\text{trn}}$ and $\mathcal{Z}_{\text{aug}}$
- **Separability h_s** measures the distance between $\mathcal{Z}_{\text{trn}}$ and $\mathcal{Z}^{(a)}_{\text{test}}$ on ℓ



Blue is better than red

Main Idea 2: Summary

- **Observation:** $\mathcal{L}_{\text{ali}} = d\left(\mathcal{Z}_{\text{aug}}, \mathcal{Z}_{\text{test}}^{(a)}\right)$ is minimized if and only if
  - The discordance h_d is zero
  - The separability h_s is one
- Problem now is to minimize h_d and to maximize h_s up to one
  - **Benefit:** Easier to design **surrogate losses** for h_d and h_s than for $\mathcal{L}_{\text{ali}}$

Main Idea 3

- **Question:** How can we design label-free **surrogate losses**?
- **Approach:**

Use $\mathcal{Z}_{\text{test}} = \mathcal{Z}_{\text{test}}^{(n)} \cup \mathcal{Z}_{\text{test}}^{(a)}$ instead of each $\mathcal{Z}_{\text{test}}^{(n)}$ or $\mathcal{Z}_{\text{test}}^{(a)}$

$$h_d = \frac{d\left(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{test}}^{(a)}\right) + d\left(\mathcal{Z}_{\text{aug}}, \mathcal{Z}_{\text{test}}^{(a)}\right)}{d(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}})} - 1$$

$$h_s = \frac{\text{proj}\left(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}}, \mathcal{Z}_{\text{test}}^{(a)}\right)}{d(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}})}$$

$$\mathcal{L}_{\text{dis}} = \frac{d(\mathcal{Z}_{\text{trn}} \cup \mathcal{Z}_{\text{aug}}, \mathcal{Z}_{\text{test}})}{d(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}})}$$

$$\mathcal{L}_{\text{sep}} = \frac{\text{std}(\{\text{proj}(\mu_{\text{trn}}, \mathbf{z}_{\text{aug}}, \mathbf{z}_{\text{test}})\})}{d(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}})}$$

Main Idea 3: Theoretical Analysis

- We show theoretically that

  $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{sep}}$ are **good approximations** of h_d and h_s, respectively

$$\mathcal{L}_{\text{dis}}: \quad c_2 h_d + c_2 + c_3 \leq \mathcal{L}_{\text{dis}}(\cdot) \leq c_2 h_d + c_2 + c_3 + \frac{(c_1 + c_3)(\sigma + \epsilon)}{d(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}})},$$

$$\mathcal{L}_{\text{sep}}: \quad \mathcal{L}_{\text{sep}}(\mathcal{Z}_{\text{trn}}, \mathcal{Z}_{\text{aug}}, \mathcal{Z}_{\text{test}}) = \sqrt{\gamma(1-\gamma)}\, h_s + \frac{\sqrt{\gamma}\bar{\sigma}_{\text{test}}}{\|\mathbf{z}_{\text{aug}} - \mathbf{z}_{\text{trn}}\|}.$$

DSV: Summary

- Our DSV loss $\mathcal{L}_{\mathrm{DSV}}$ is the combination of $\mathcal{L}_{\mathrm{dis}}$ and $\mathcal{L}_{\mathrm{sep}}$
  - **Idea** is to minimize $\mathcal{L}_{\mathrm{dis}}$ while maximizing $\mathcal{L}_{\mathrm{sep}}$ to some extent

$$\mathcal{L}_{\mathrm{DSV}}(\mathcal{Z}_{\mathrm{trn}}, \mathcal{Z}_{\mathrm{aug}}, \mathcal{Z}_{\mathrm{test}}) = \mathcal{L}_{\mathrm{dis}}(\cdot) - \frac{\max(\mathcal{L}_{\mathrm{sep}}(\cdot), 1/2)}{\mathcal{L}_{\mathrm{dis}}(\cdot)},$$

- We search for ϕ^* that shows the smallest $\mathcal{L}_{\mathrm{DSV}}$:

$$\phi^* = \mathrm{argmin}_{\phi \in \Phi}\, \mathcal{L}_{\mathrm{DSV}}(\phi; \mathcal{D}_{\mathrm{trn}}, \mathcal{D}_{\mathrm{test}}, f_{\mathrm{aug}})$$

Outline

- Introduction
- Proposed Method
- **Experiments**
- Conclusion

Experimental Questions

- **Q1: Performance**
  - How good are the models selected by DSV?
- **Q2: Ablation study**
  - Are both the *discordance* and *separability* meaningful?
- **Q3: Case studies (and visualization)**
  - How well does DSV work on individual AD tasks?

Experimental Settings

- **Datasets:** MVTec AD and MPDD for image AD
  - 21 different tasks in total
- **Detector model:** ResNet18-based classifier
- **Augmentation functions:** CutOut, CutAvg, CutDiff, and CutPaste
  - CutAvg and CutDiff are variants of CutOut
- **Target hyperparameters:** Patch size in f_{aug}

Q1. Performance

- Average AUC and rank across 21 different tasks in the two datasets
- Our DSV outperforms all competitors in 6 of the 8 cases

AUC:

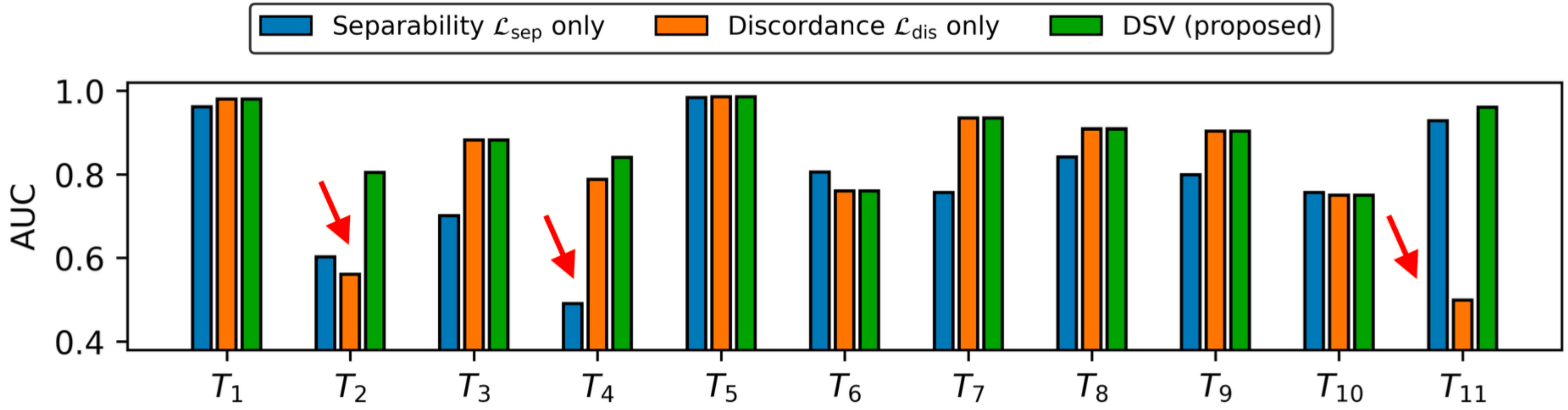
f_{aug}	Avg.	Rand.	Base	MMD	STD	MC	SEL	HITS	**DSV**
CutOut	0.739	<u>0.776</u>	0.741	0.735	0.739	0.749	0.727	0.757	**0.813**
CutAvg	0.739	**0.817**	0.721	0.692	0.745	0.751	0.744	0.742	<u>0.806</u>
CutDiff	0.743	0.711	0.739	0.730	0.744	0.747	0.741	<u>0.777</u>	**0.811**
CutPaste	0.788	0.841	0.694	0.756	0.818	<u>0.862</u>	0.830	0.850	**0.884**

Rank:

f_{aug}	Avg.	Rand.	Base	MMD	STD	MC	SEL	HITS	**DSV**
CutOut	7.33	6.10	6.62	6.93	6.29	6.50	7.10	<u>5.43</u>	**3.79**
CutAvg	7.00	<u>5.02</u>	7.64	8.36	5.52	5.48	5.98	5.60	**4.19**
CutDiff	6.43	7.24	6.45	7.38	6.00	<u>5.64</u>	6.24	6.21	**3.60**
CutPaste	7.67	6.29	8.67	7.21	5.60	**4.33**	5.17	4.64	<u>4.57</u>

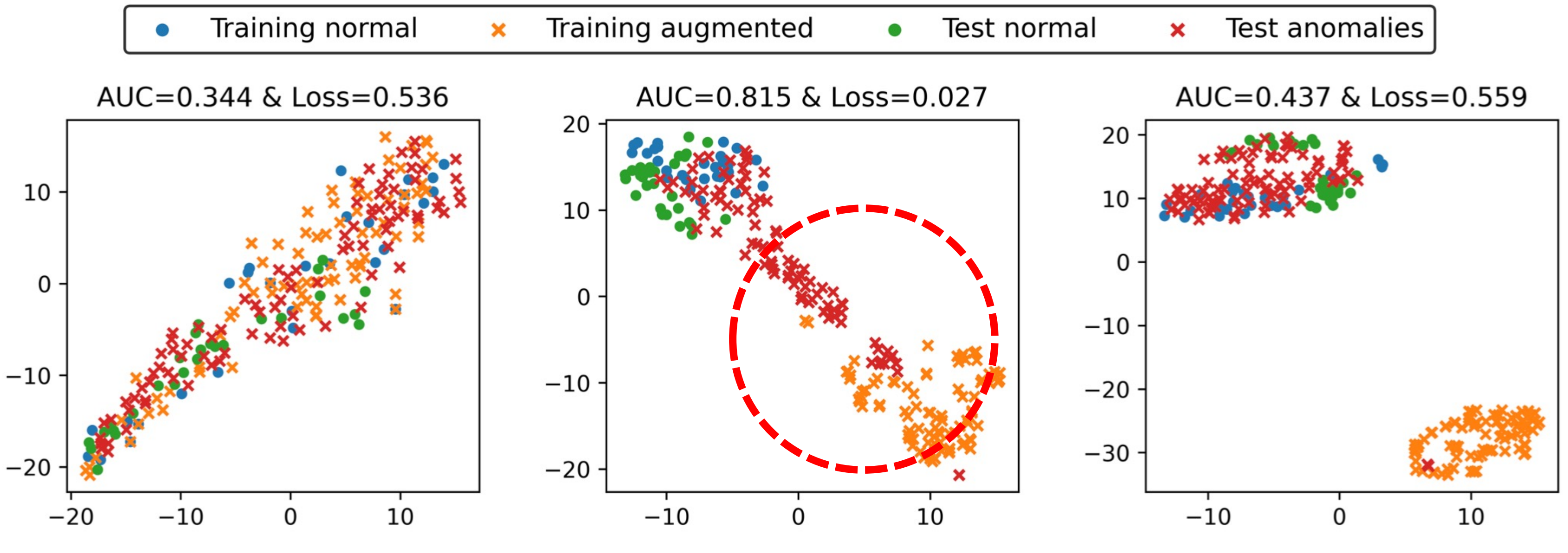
Q2. Ablation Study

- Comparison between $\mathcal{L}_{\mathrm{dis}}$, $\mathcal{L}_{\mathrm{sep}}$, and $\mathcal{L}_{\mathrm{DSV}}$ when $f_{\mathrm{aug}} = \mathrm{CutPaste}$
- DSV shows a dramatic improvement in a few cases
  - E.g., tasks T_2 (both fail), T_4 (L_{sep} fails), T_{11} and T_{14} ($\mathcal{L}_{\mathrm{dis}}$ fails)



Q3. Case Studies (1)

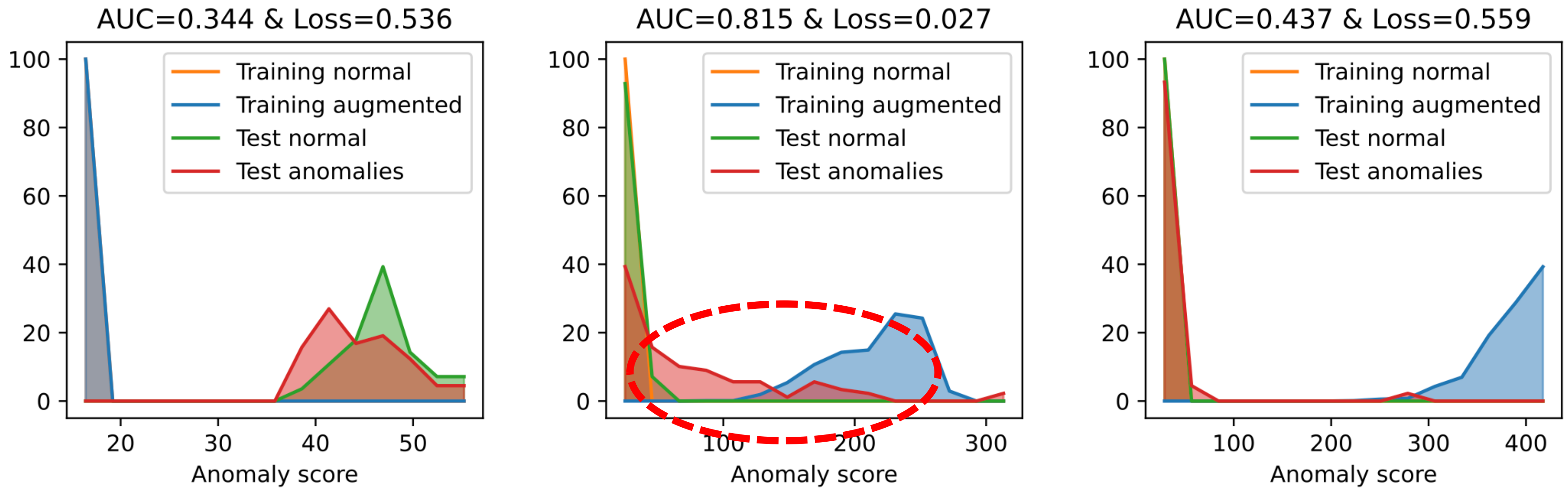
- Embedding distributions with different patch sizes on CutOut
- In (b), augmented data and test anomalies are best aligned with DSV



(a) CutOut (0.00004) (b) CutOut (0.00128) (c) CutOut (0.04)

Q3. Case Studies (2)

- Anomaly score distributions with different patch sizes on CutOut
- In (b), augmented data and test anomalies are best aligned with DSV



(a) CutOut (0.00004) (b) CutOut (0.00128) (c) CutOut (0.04)

Outline

- Introduction
- Proposed Method
- Experiments
- **Conclusion**

Conclusion

- We propose DSV, a validation loss for model selection on SSAD
- DSV consists of three main ideas:
  - **Main Idea 1:** Define alignment as the embedding distance
  - **Main Idea 2:** Decompose the alignment into discordance and separability
  - **Main Idea 3:** Design surrogate losses, which do not require labels
- DSV outperforms the baselines for unsupervised model selection
- **Paper and code:** https://github.com/jaeminyoo/DSV