SHANGHAI JIAO TONG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

ADPretrain: Advancing Industrial Anomaly Detection via Anomaly Representation Pretraining

NeurIPS 2025

Xincheng Yao
Shanghai Jiao Tong University

Coauthors:
Yan Luo*, Zefeng Qian, Chongyang Zhang*

Preview

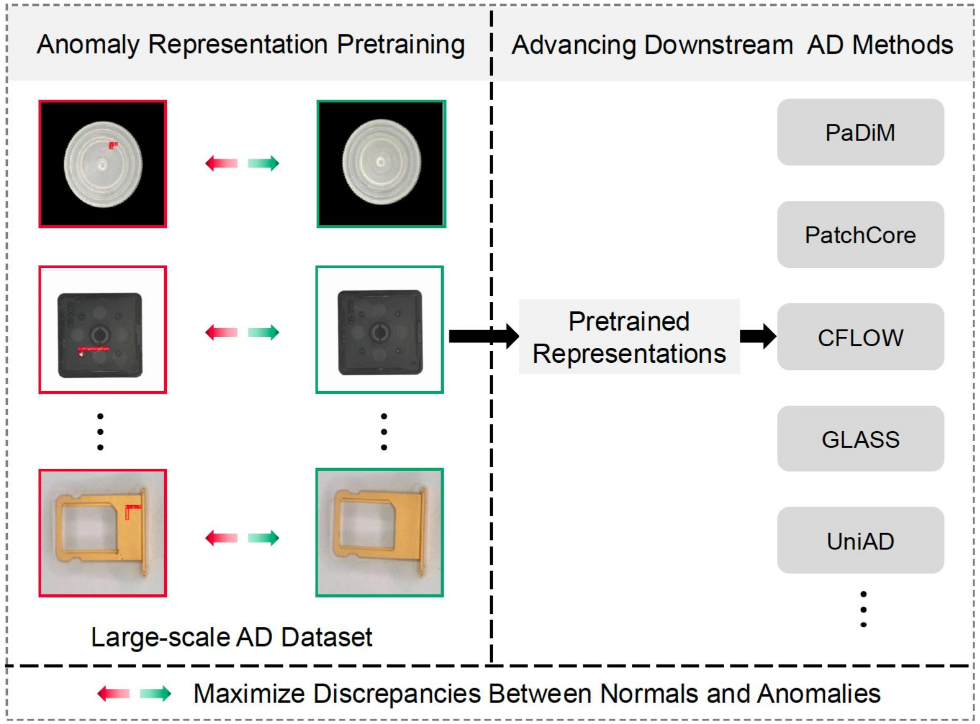
NEURAL INFORMATION PROCESSING SYSTEMS

SHANGHAI JIAO TONG UNIVERSITY

Current Industrial Anomaly Detection

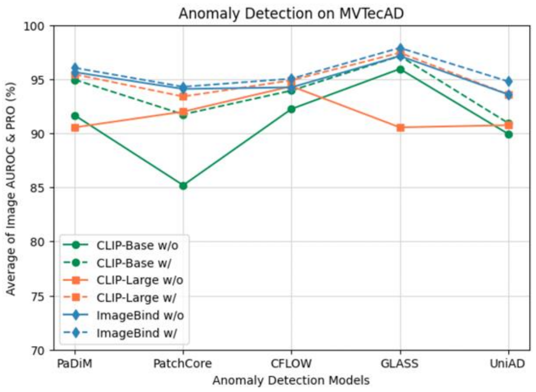
The current **mainstream and state-of-the-art** AD methods are substantially established on **pretrained feature networks**.

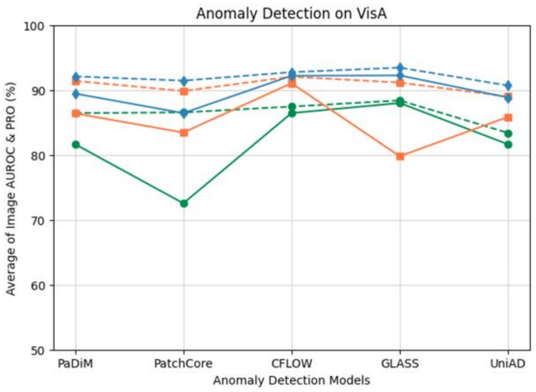
What has been overlooked in anomaly detection research?



Specialized pretrained representations for anomaly detection!

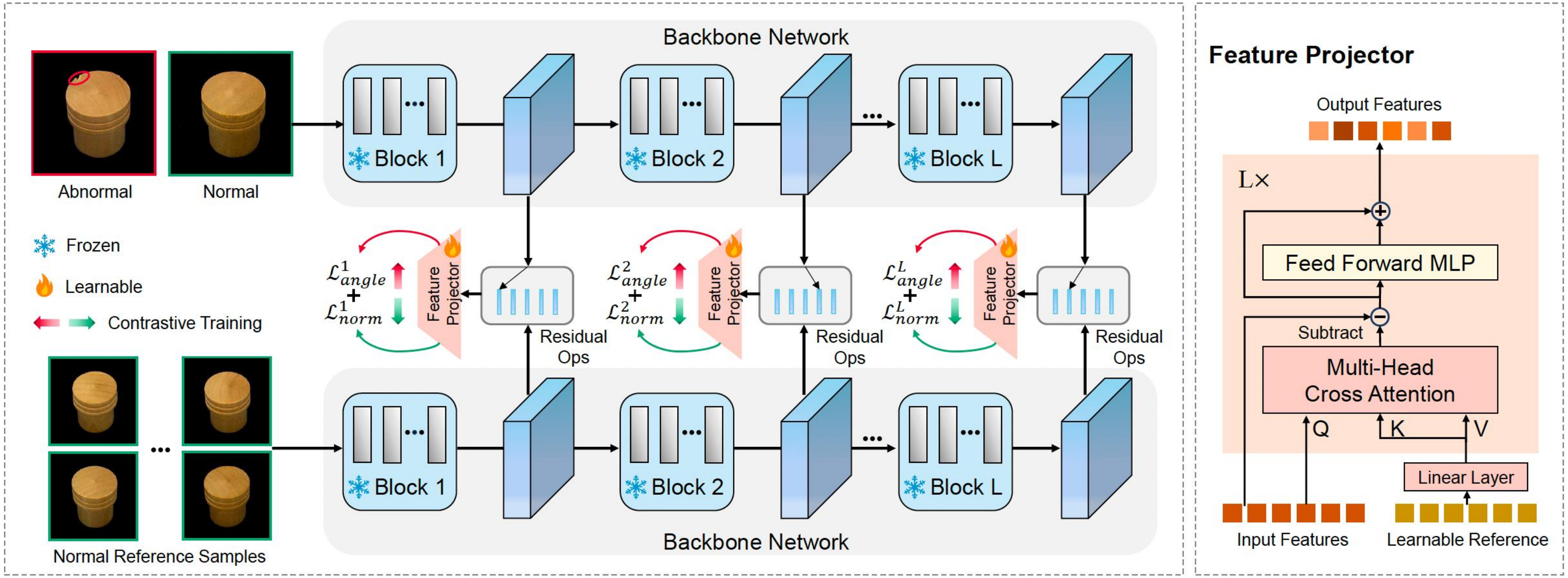
Anomaly Representation Pretraining!





Preview

ADPretrain: Advancing Industrial Anomaly Detection via Anomaly Representation Pretraining



The framework consists of: Residual Features, Angle- and Norm-Oriented Contrastive Losses, Feature Projector.

Outline

Motivation

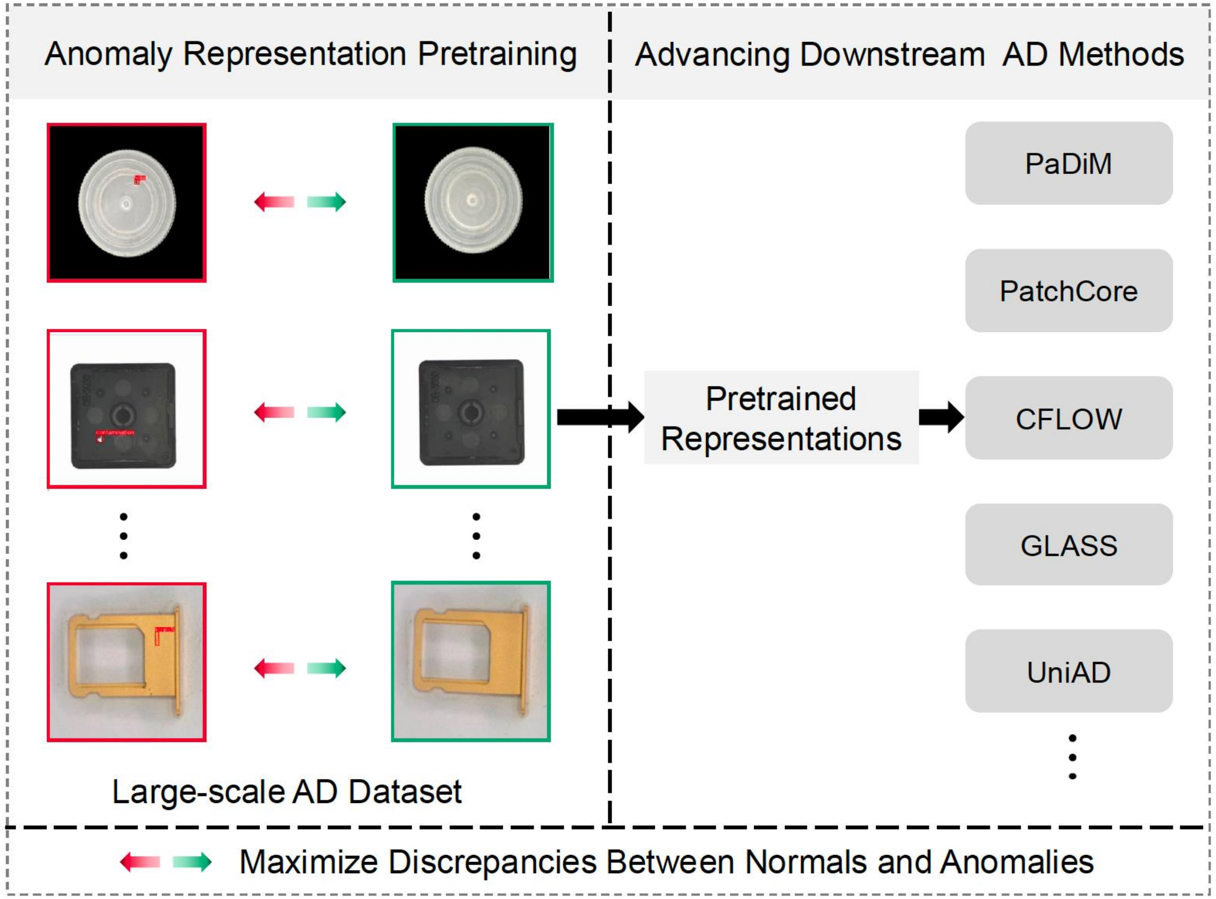
Our core insight: Anomaly detection needs specialized pretrained features!

What about current industrial anomaly detection?

- The current **mainstream and SOTA** AD methods are substantially established on **pretrained feature networks** (e.g., ImageNet-pretrained).
- **Two issues**: 1. The pretraining process in natural images **doesn`t match the goal of anomaly detection**.
- 2. Natural images and industrial image data typically have the **distribution shift**.

Anomaly Representation Pretraining:

- Learning **specialized AD representations for AD tasks** that are **better than basic pretrained features** when applied to downstream AD methods.

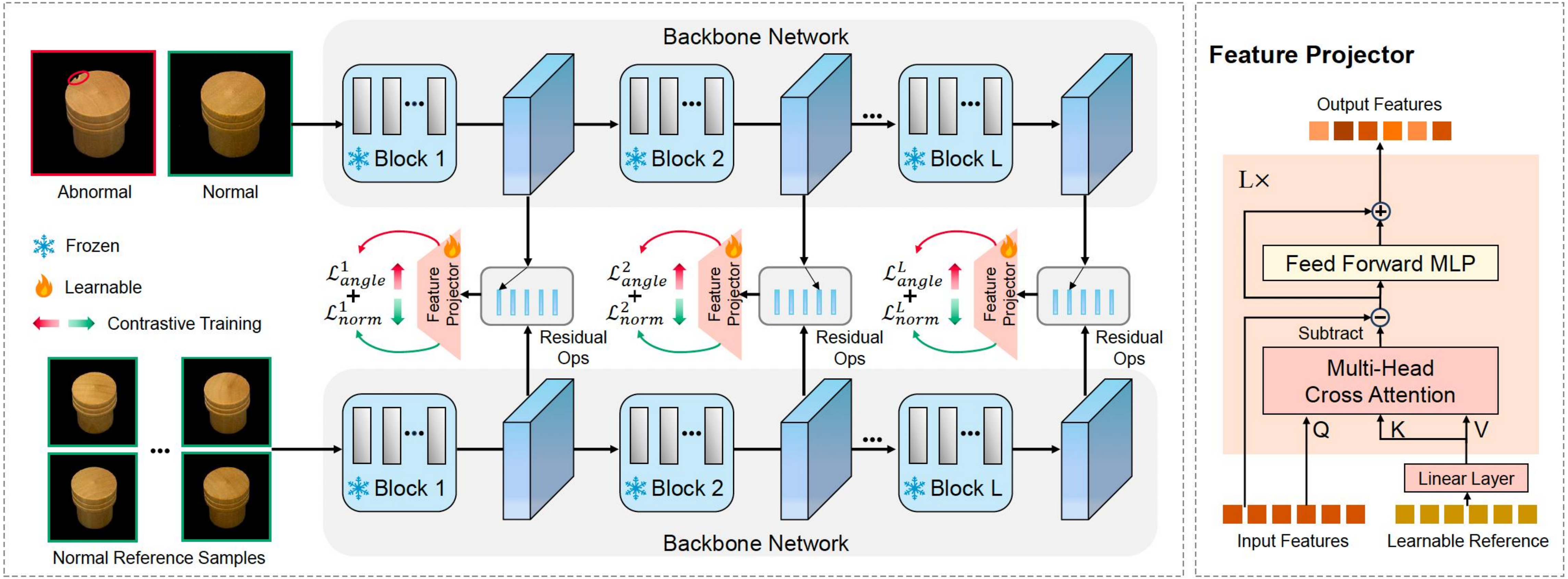


Our Approach: ADPretrain

NEURAL INFORMATION PROCESSING SYSTEMS

SHANGHAI JIAO TONG UNIVERSITY

- **ADPretrain, Framework Overview:**



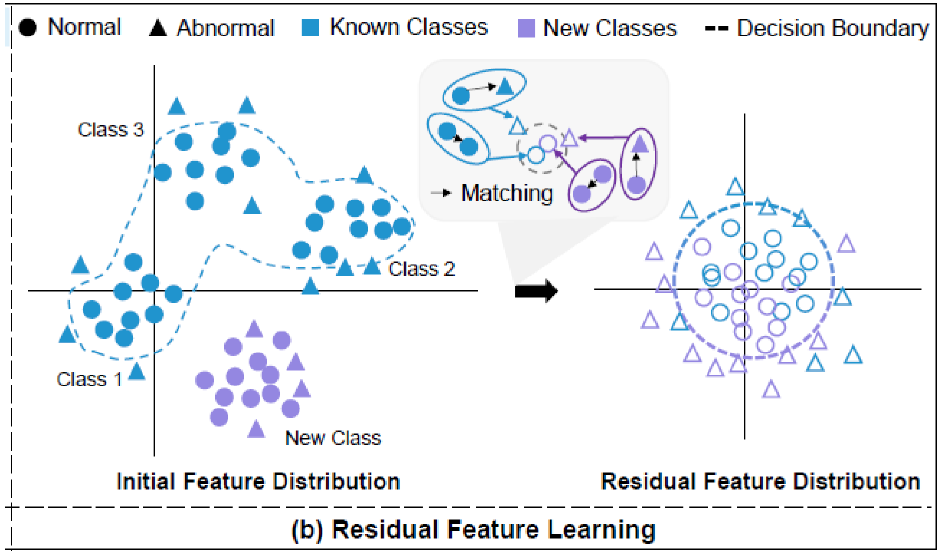
The framework consists of: Residual Features, Angle- and Norm-Oriented Contrastive Losses, Feature Projector.

Our Approach: ADPretrain

- **Construction of Fundamental Anomaly Detection Representations**

- We expect that pretrained features can serve as fundamental features in anomaly detection (i.e., can perform well on various AD datasets).
- A valuable question: **what kind of representations can serve as fundamental (general) representation for anomaly detection?**
- We think that it`s best for **pretrained representations to be domain-invariant**.
- To this end, we choose the recently proposed class-generalizable representation in AD: **Residual Features.**
- **Residual Features:** the residual representation of x is defined as:

$$x_r = x - x_n^*;\ \ x_n^* = \mathrm{argmin}_{x' \in \mathcal{P}} \|x' - x\|_2$$



*source from ResAD paper

Our Approach: ADPretrain

- **Contrastive Losses for Anomaly Representation Pretraining**

- According to the characteristic of anomaly detection (i.e., focus on discrepancies between normals and anomalies), contrastive learning should be the most suitable pretraining paradigm.
- From the feature similarity perspective, the discrepancies between two features are embodied in two aspects: the **angle size** and the **feature norm**.
- Therefore, we propose **angle- and norm-oriented contrastive losses** to maximize the angle size and norm difference between normal and abnormal features simultaneously.
- **Angle-Oriented Contrastive Loss:**

$$\mathcal{L}_{angle}(x_i, x_{i'}) = -log\left(\frac{\exp(sim(\overline{x_i}, \overline{x_{i'}})/\tau)}{\sum_{k=1}^{2N}\mathbb{I}_{[k\neq i]}\cdot\mathbb{I}_{[m_k\neq m_i]}\cdot\exp(sim(\overline{x_i}, \overline{x_k})/\tau)}\right)$$

- Each image is randomly augmented to an augmented image, $x_{i'}, i' = i + N$ is the feature from the same position in the augmented image.
- Only $x_{i'}$ is used as the positive pair, features with different labels from x_i are used as negative pairs.

Our Approach: ADPretrain

- **Contrastive Losses for Anomaly Representation Pretraining**

- **Norm-Oriented Contrastive Loss:**
- This loss aims to **enlarge the feature norm difference** between normal and abnormal features.
- The basic idea follows OCC learning, where normal features are optimized inside the origin-centered hypersphere.

$$\mathcal{L}_{con}(x_i) = -logsig(-(n_i - r)) \cdot \exp(n_i - r)$$

- The above loss is only for normal features, n_i is the feature norm. The normal features are optimized to contract **inside a hypersphere with radius $\boldsymbol{r}$**.

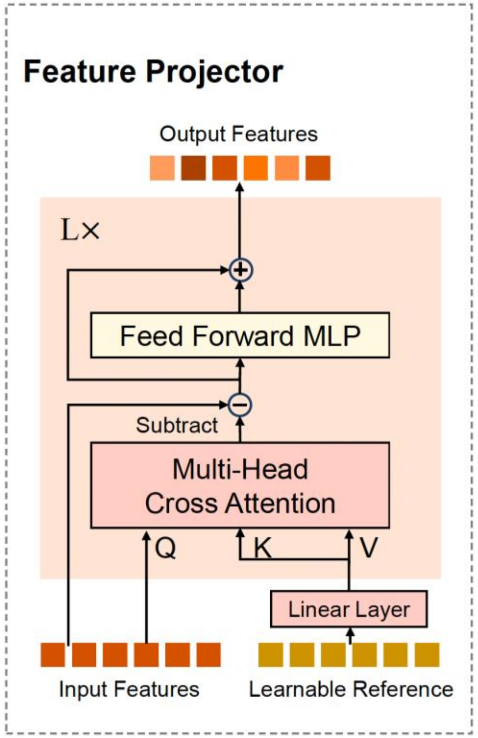
$$\mathcal{L}'_{con}(x_j) = \begin{cases} -logsig\left(-(r' - n_j)\right) \cdot \exp^{r'-n_j}, & n_j \le r' \\ 0, & n_j > r' \end{cases}$$

- The above loss is only for abnormal features. The abnormal features are pushed **outside a hypersphere with radius $\boldsymbol{r'}$**, $r' = r + \Delta r$.

$$\mathcal{L}_{norm}(x_i) = \mathbb{I}_{[m_i=0]} \cdot \mathcal{L}_{con}(x_i) + \mathbb{I}_{[m_i=1]} \cdot \mathcal{L}'_{con}(x_i).$$

Our Approach: ADPretrain

- **Feature Projector**



- Features yielded by the Feature Projector are optimized by the angle- and norm-oriented contrastive losses.
- The Feature Projector is based on Transformer architecture, but we alter self-attention to our proposed learnable key/value attention.

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

SHANGHAI JIAO TONG UNIVERSITY

- **Datasets:**
  - Pretraining dataset: RealIAD.
  - Downstream AD datasets: MVTecAD, VisA, BTAD, MVTed3D, MPDD.
- **Backbones:**
  - DINOv2-Base, DINOv2-Large, CLIP-Base, CLIP-Large, ImageBind.
- **Downstream AD Methods:**
  - PaDiM, PatchCore, CFLOW, GLASS, UniAD, and our proposed simple FeatureNorm.
- **Metrics:**
  - image-level AUROC, pixel-level PRO.

Experiments

- **Comparison on five AD methods, five AD datasets, and five backbones.**

Model	Datasets	PaDiM [8]	PaDiM†	PatchCore [33]	PatchCore†	CFLOW [12]	CFLOW†	GLASS [6]	GLASS†	UniAD [50]	UniAD†	FeatureNorm	FeatureNorm†
DINOv2-Base [26]	MVTecAD	95.6/93.1	95.9+0.3/92.5-0.6	95.5/82.7	99.0+3.5/87.4+4.7	97.7/92.3	98.3+0.6/92.9+0.6	98.3/93.5	99.0+0.7/95.2+1.7	71.1/81.5	97.1+26.0/91.2+9.7	48.4/28.9	98.2/92.8
	VisA	91.7/84.4	93.1+1.4/85.7+1.3	82.8/69.9	92.9+10.1/81.3+11.4	94.3/89.4	95.2+0.9/88.6-0.8	93.3/90.1	93.5+0.2/87.7-2.4	90.6/84.4	94.4+3.8/87.3+2.9	52.2/30.1	94.8/87.2
	BTAD	96.6/74.4	95.2-1.4/74.7+0.3	90.2/61.7	94.3+4.1/62.2+0.5	93.2/70.4	95.2+2.0/72.3+1.9	92.3/78.2	95.2+2.9/81.6+3.4	78.0/67.9	93.4+15.4/71.1+3.2	54.9/15.3	95.1/71.7
	MVTec3D	82.1/92.0	81.5-0.6/92.4+0.4	72.9/78.9	82.8+9.9/87.5+8.6	88.6/91.9	88.5-0.1/92.5+0.6	84.5/90.2	87.3+2.8/90.7+0.5	66.6/78.3	85.3+18.7/91.8+13.5	49.0/54.6	84.8/91.2
	MPDD	80.9/87.8	88.3+7.4/92.1+4.3	89.4/72.2	92.4+3.0/88.7+16.5	91.3/93.2	91.7+0.4/93.7+0.5	91.1/90.6	95.7+4.6/92.3+1.7	76.9/53.4	88.6+11.7/91.9+38.5	44.0/36.3	93.6/93.1
DINOv2-Large [26]	MVTecAD	98.7/91.0	98.6-0.1/92.4+1.4	97.6/83.8	99.0+1.4/88.0+4.2	98.8/92.7	98.9+0.1/93.2+0.5	98.4/95.3	99.1+0.7/96.2+0.9	79.6/83.0	96.9+17.3/91.6+8.6	48.4/31.4	98.7/93.1
	VisA	92.6/85.6	95.1+2.5/86.7+1.1	85.1/71.1	91.5+6.4/82.5+11.4	96.2/90.0	96.9+0.7/90.6+0.6	93.3/90.4	94.0+0.7/91.8+1.4	90.9/84.0	94.8+3.9/88.7+4.7	45.7/33.4	96.2/88.3
	BTAD	94.0/75.2	94.6+0.6/75.6+0.4	93.6/63.2	93.5-0.1/61.8-1.4	94.8/74.7	95.8+1.0/76.4+2.0	93.8/80.9	95.6+1.8/82.3+1.4	85.1/71.9	92.3+7.2/72.5+0.6	46.1/18.7	94.3/73.7
	MVTec3D	83.0/87.7	86.6+3.6/88.8+1.1	75.7/74.7	82.5+6.8/85.3+10.6	91.8/93.0	91.1-0.7/93.3+0.3	83.4/92.0	87.8+4.4/93.3+1.3	79.2/87.9	83.0+3.8/91.8+3.9	47.4/53.5	86.0/91.9
	MPDD	87.5/86.8	94.3+6.8/90.1+3.3	93.6/79.5	90.7-2.9/87.0+7.5	95.3/94.0	93.4-1.9/94.1+0.1	91.6/95.7	95.3+3.7/98.0+2.3	76.0/62.9	90.6+14.6/92.6+29.7	39.1/45.8	94.6/95.0
CLIP-Base [29]	MVTecAD	93.4/89.9	98.1+4.7/91.8+1.9	92.9/76.0	98.3+5.4/85.7+9.7	94.2/90.3	96.7+2.5/91.2+0.9	97.1/94.8	98.0+0.9/93.8-1.0	92.2/87.7	93.1+0.9/88.8+1.1	48.7/36.3	96.9/91.4
	VisA	87.7/75.6	92.6+4.9/80.3+4.7	81.2/63.3	92.5+11.3/80.6+17.3	89.5/83.4	91.5+2.0/84.7+1.3	92.1/85.0	91.4-0.7/85.4+0.4	83.5/79.8	87.0+3.5/79.8+0.0	48.9/31.9	92.5/82.6
	BTAD	93.9/70.2	95.4+1.5/73.2+3.0	91.4/58.4	91.6+0.2/63.6+5.2	94.3/71.0	93.9-0.4/72.7+1.7	92.5/80.5	94.3+1.8/81.3+0.8	90.8/66.8	94.2+3.4/72.5+5.7	47.7/14.0	94.1/72.0
	MVTec3D	71.7/80.8	81.3+9.6/87.1+6.3	65.0/62.7	77.9+12.9/84.6+21.9	82.1/90.1	83.8+1.7/90.4+0.3	79.2/90.8	80.7+1.5/89.2-0.6	65.3/81.6	82.6+17.3/90.7+9.1	50.9/10.4	81.4/90.6
	MPDD	85.6/79.3	92.0+6.4/89.5+10.2	80.5/59.9	90.6+10.1/88.6+28.7	85.2/90.4	90.1+4.9/92.7+2.3	92.6/94.2	94.3+1.7/94.3+0.1	82.2/78.4	86.8+4.6/91.5+13.1	50.8/20.6	93.6/91.7
CLIP-Large [29]	MVTecAD	89.4/91.7	98.4+9.0/92.5+0.8	97.0/87.0	98.9+1.9/87.5+0.5	97.3/91.4	98.0+0.7/91.8+0.4	93.8/87.3	98.8+5.0/95.2+7.9	92.6/88.9	96.4+3.8/90.8+1.9	51.6/70.0	98.4/92.9
	VisA	90.1/82.7	95.2+5.1/87.6+4.9	87.7/78.6	94.5+6.8/85.2+6.6	93.6/88.5	94.9+1.3/89.7+1.2	83.0/76.6	93.4+10.4/88.9+12.3	86.6/85.1	91.3+4.7/86.7+1.6	50.3/67.8	94.8/89.8
	BTAD	90.8/73.4	95.4+4.6/74.9+1.5	91.8/64.1	93.9+2.1/65.7+1.6	93.7/74.4	94.8+1.1/72.9-1.5	94.0/76.4	95.5+1.5/82.6+6.2	83.8/71.5	94.8+11.0/74.2+2.7	54.8/15.0	94.2/74.8
	MVTec3D	71.4/87.0	84.7+13.3/92.2+5.2	75.1/82.6	83.3+8.2/87.4+4.8	84.9/91.8	85.1+0.2/92.6+0.8	82.9/91.2	86.2+3.3/89.9-1.3	76.4/90.5	81.4+5.0/92.4+1.9	51.9/18.9	84.4/93.0
	MPDD	82.9/88.8	94.8+11.9/94.2+5.4	87.7/83.4	92.6+4.9/92.5+9.1	92.2/93.5	90.1-2.1/94.7+1.2	91.6/95.7	94.1+2.5/98.3+2.6	73.0/76.2	91.7+18.7/94.5+18.3	50.3/26.4	94.1/95.1
ImageBind [11]	MVTecAD	97.9/92.6	98.8+0.9/92.1-0.5	98.5/88.9	98.9+0.4/88.8-0.1	97.8/90.7	98.6+0.8/91.5+0.8	98.7/95.6	99.4+0.7/96.4+0.8	96.0/91.2	98.1+2.1/91.5+0.3	83.5/80.3	98.6/92.6
	VisA	92.6/86.3	95.6+3.0/88.6+2.3	91.4/81.9	94.8+3.4/86.3+4.4	94.9/89.5	95.3+0.4/90.2+0.7	94.9/88.7	95.9+1.0/91.0+2.3	90.3/87.4	93.2+2.9/88.2+0.8	70.0/73.6	95.3/90.0
	BTAD	94.6/75.9	95.9+1.3/76.8+0.9	94.6/66.7	95.6+1.0/67.3+0.6	94.9/72.6	95.4+0.5/75.6+3.0	94.9/84.8	95.8+0.9/84.3-0.5	67.1/59.2	94.7+27.6/75.8+16.6	37.5/19.3	93.5/76.9
	MVTec3D	79.5/90.3	84.4+4.9/92.0+1.7	78.4/86.3	82.6+4.2/87.0+0.7	85.8/91.8	83.5-2.3/91.8+0.0	83.5/91.8	86.2+2.7/91.8+0.0	80.2/90.8	80.8+0.6/92.0+1.2	52.2/64.5	83.3/92.2
	MPDD	91.0/92.0	94.4+3.4/95.1+3.1	92.6/89.1	94.8+2.2/94.2+5.1	92.6/94.0	91.5-1.1/95.0+1.0	96.4/99.0	95.7-0.7/99.0+0.0	60.7/52.3	93.6+32.9/95.0+42.7	44.9/40.7	94.2/95.6

- We reproduce these methods based on their official open-source code and default hyperparameters.
- We **only replace the original features with our pretrained features**.

Ablations

- **Ablation study results:**

(a) Framework ablation studies.

ExpID	Pretrained Representations	Contrastive Losses	Feature Projector	Backbone Network	PaDiM	PatchCore	FeatureNorm
2.1	Non-residual	/	/	Fixed	92.6/86.3	91.6/81.3	49.2/44.5
2.2		Angle&Norm	w/	Fixed	93.5/86.1	93.6/85.1	82.9/83.9
2.3		Angle&Norm	/	Non-fixed	89.3/83.4	91.9/84.8	81.0/83.2
2.4	Residual	/	/	Fixed	93.9/85.6	92.9/86.5	91.3/86.8
2.5		Angle	w/	Fixed	93.9/85.2	93.2/83.8	83.9/83.0
2.6		Norm	w/	Fixed	93.7/85.3	90.9/84.5	92.4/85.1
2.7		Angle&Norm	w/	Fixed	95.4/88.7	94.6/87.0	94.2/89.0
2.8		Angle&Norm	/	Non-fixed	81.3/56.0	78.4/32.1	51.9/20.7

(b) Architecture ablation studies for the Feature Projector.

Architecture	PaDiM	PatchCore	FeatureNorm
Linear Projector	93.8/87.4	94.0/86.8	87.9/82.1
MLP Projector	93.4/88.1	94.2/79.1	94.2/90.2
Self Attention	93.7/88.8	92.9/84.3	92.9/84.9
Cross Attention	94.8/88.9	94.6/82.9	93.1/86.5
Self + Cross Attention	94.8/88.7	94.0/80.9	92.9/84.4
Learnable Key/Value Attention (ours)	95.5/88.8	94.7/87.6	94.5/89.3

- 1. Residual features are better pretrained AD representations.
- 2. The two proposed contrastive losses are effective and are also complementary.
- 3. Currently, the backbone network need to remain fixed. Training the whole feature network is still challenging, needs larger-scale and better-quality datasets to support.
- 4. Our Learnable Key/Value Attention (LKV-Attn) can outperform other network architectures.

Further Analysis

NEURAL INFORMATION PROCESSING SYSTEMS

SHANGHAI JIAO TONG UNIVERSITY

- **Sample Efficiency & Robustness to Noise:**

(a) Sample efficiency experiment results.

Datasets	PaDiM	PaDiM†	PatchCore	PatchCore†
MVTecAD	81.4/88.6	96.8+15.4/90.3+1.7	96.5/86.2	98.2+1.7/87.7+1.5
VisA	82.8/79.3	93.0+10.2/83.3+4.0	88.9/78.7	93.9+5.0/85.7+7.0
BTAD	89.2/75.2	94.8+5.6/76.3+1.1	93.1/65.5	94.0+0.9/68.9+3.4
MVTec3D	66.3/85.8	82.1+15.8/91.9+6.1	74.3/84.7	80.8+6.5/87.1+2.4
MPDD	63.1/75.6	91.4+28.3/94.8+19.2	79.1/84.4	90.8+11.7/93.3+8.9

(b) Robustness experiment results.

Datasets	PaDiM	PaDiM†	PatchCore	PatchCore†
MVTecAD	86.6/82.7	88.2+1.6/84.6+1.9	87.7/80.6	89.2+0.8/81.4+1.5
VisA	82.8/75.6	87.4+4.6/79.9+4.3	83.4/71.5	86.9+3.5/78.7+7.2
BTAD	85.4/70.9	90.5+5.1/73.4+2.5	85.1/62.4	86.3+1.2/65.6+3.2
MVTec3D	72.0/83.1	76.2+4.2/84.9+1.8	71.5/78.6	73.8+2.3/80.5+1.9
MPDD	79.8/83.8	84.0+4.2/86.4+2.6	83.4/79.1	84.5+1.1/85.8+6.7

- 1. **Sample Efficiency**. Only 10% normal samples are used for training. With less training data, our pretrained features can bring more significant performance improvement.
- 2. **Robustness**. We add abnormal data from the test set to the training set. With noise, our pretrained features still bring performance improvement, they are robust to noisy training data.

Further Analysis

- **Few-shot Anomaly Detection**

Setup	Method	Venue	MVTecAD			VisA		
			I-AUROC	P-AUROC	PRO	I-AUROC	P-AUROC	PRO
2-shot	SPADE*	arXiv2020	82.9	92.0	85.7	80.7	96.2	85.7
	PatchCore*	CVPR2022	86.3	93.3	82.3	81.6	96.1	82.6
	WinCLIP*	CVPR2023	94.4	96.0	88.4	84.6	96.8	86.2
	AnomalyGPT#,‡	AAAI2024	95.5	95.6	90.0	88.6	96.4	83.4
	PromptAD#	CVPR2024	95.7	96.2	88.5	88.3	97.1	85.8
	InCTRL	CVPR2024	94.0	/	/	85.8	/	/
	ResAD‡	NeurIPS2024	94.4	95.6	/	84.5	95.1	/
	KAG-Prompt#,‡	AAAI2025	**96.6**	**96.5**	**91.1**	**92.7**	97.4	86.7
	FeatureNorm‡ (ours)	-	95.3	95.6	90.9	92.4	**97.6**	**87.5**
4-shot	SPADE*	arXiv2020	84.8	92.7	87.0	81.7	96.6	87.3
	PatchCore*	CVPR2022	88.8	94.3	84.3	85.3	96.8	84.9
	WinCLIP*	CVPR2023	95.2	96.2	89.0	87.3	97.2	87.6
	AnomalyGPT#,‡	AAAI2024	96.3	96.2	90.7	90.6	96.7	84.6
	PromptAD#	CVPR2024	96.6	96.5	90.5	89.1	97.4	86.2
	InCTRL	CVPR2024	94.5	/	/	87.7	/	/
	ResAD‡	NeurIPS2024	94.2	**96.9**	/	90.8	97.5	/
	KAG-Prompt#,‡	AAAI2025	**97.1**	96.7	**91.4**	93.3	97.7	87.6
	FeatureNorm‡ (ours)	-	96.2	95.9	91.3	**94.5**	**98.1**	**89.3**

- One valuable advantage: the feature norms can be directly used as anomaly scores.
- For few-shot anomaly detection, our **simple FeatureNorm is comparable and even superior** (on VisA).
- With **better AD representations**, we can **simply achieve good FSAD results** without designing elaborate methods.

Conclusions

NEURAL INFORMATION PROCESSING SYSTEMS

SHANGHAI JIAO TONG UNIVERSITY

Anomaly Representation Pretraining!

What kind of representations can serve as the fundamental (general) representation for anomaly detection?

More attention could be paid to anomaly representation pretraining, **rather than constantly focusing on designing more sophisticated AD models**.

SHANGHAI JIAO TONG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks!

Contact Us:
sunny_zhang@sjtu.edu.cn