Anomaly Detection and Localization Using Attention-Guided Synthetic Anomaly and Test-Time Adaptation

Behzad Bozorgtabar[1,2], Dwarikanath Mahapatra[3], Jean-Philippe Thiran[1,2]

[1] LTS5, EPFL, Switzerland [2] Radiology Department, CHUV, Switzerland [3] Inception Institute of AI (IIAI), UAE

Motivation

- Vision transformers (ViTs) are largely overlooked for anomaly detection and segmentation tasks.
- Compared with fully convolutional networks, ViTs offer higher representation power due to their global receptive field.
- Recent self-supervised anomaly detection methods still lag behind methods using pre-trained ImageNet with knowledge transfer/distillation

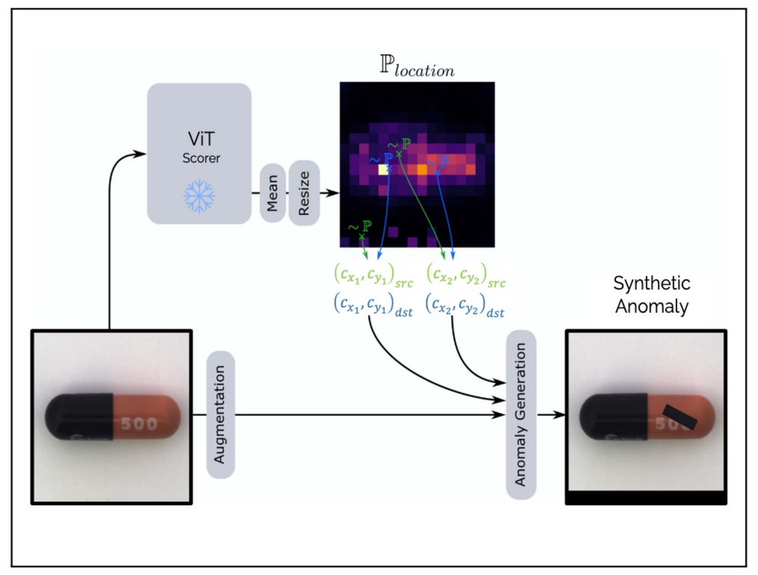
Our contributions

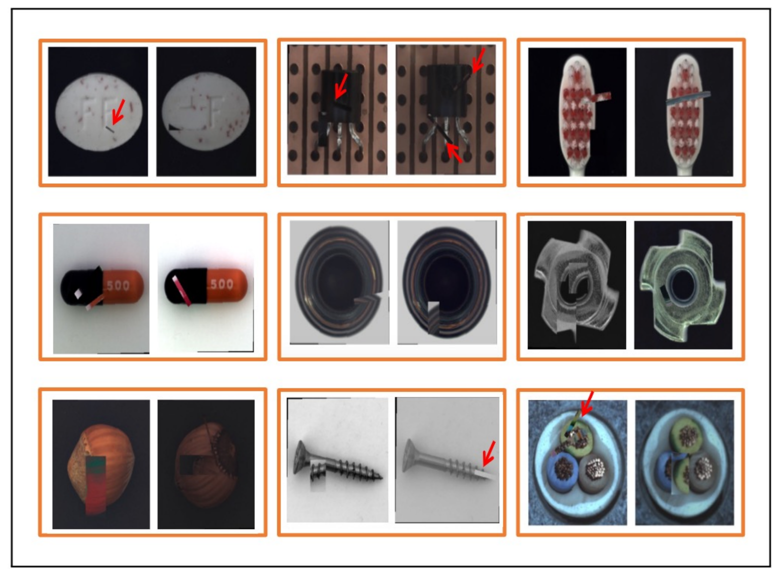
- We focus on the attention mechanism in the transformer and propose a simple yet effective **attention-guided cut-and-paste data augmentation** for creating synthetic anomalies using only nominal training data.
- To alleviate the mismatch between test and training data (e.g., real and synthetic anomalies), we adopt a **test-time adaptation scheme** to match their **class-aware distributional statistics** associated with the **transformer's attention entropy.**
- We show consistent performance improvements over current synthetic anomaly-based methods for anomaly detection and localization on the challenging benchmarks of the MVTec AD and the NIH Chest X-rays.

Attention-guided synthetic anomalies

- **Motivation**: Most anomalies, e.g., defect categories, are around the salient object.
- **Idea**: We create synthetic anomalies that are **more relevant to the task** by focusing on salient object regions derived from the **self-attention mechanism** of transformer.
- The ViT's attention map guides sampling of the **informative locations** to cut and paste patches, yielding a **more realistic approximation of real anomalies**.
- By varying the size, aspect ratio, and color of the local patch, our augmentation creates a **more diverse** compared to SOTA synthetic anomaly-based methods.
- **Proxy Task**: The model is trained using the proxy task of detecting and localizing synthetic anomaly constructed via attention-guided augmentation $Att(\cdot)$ and formulated as a binary classification and segmentation setup.
- The model consists of a ViT encoder f, which is initialized from a self-supervised method, DINO weights, and multi-layer perceptron (MLP) projection head g for image-level classification, and a set of learnable weights for multi-head attention maps.
- We define the training objective using cross-entropy loss $\mathbb{CE}$ for the image-level binary classifier in detecting synthetic anomalies from a set of anomaly-free training images X^u as follows:

$$\mathcal{L}_{\text{Att}} = \frac{1}{2|X^u|}\sum_{\mathbf{x}\in X^u}[\mathbb{CE}(g(f(\mathbf{x})),0)+\mathbb{CE}(g(f(\mathbf{Att}(\mathbf{x}))),1)]$$

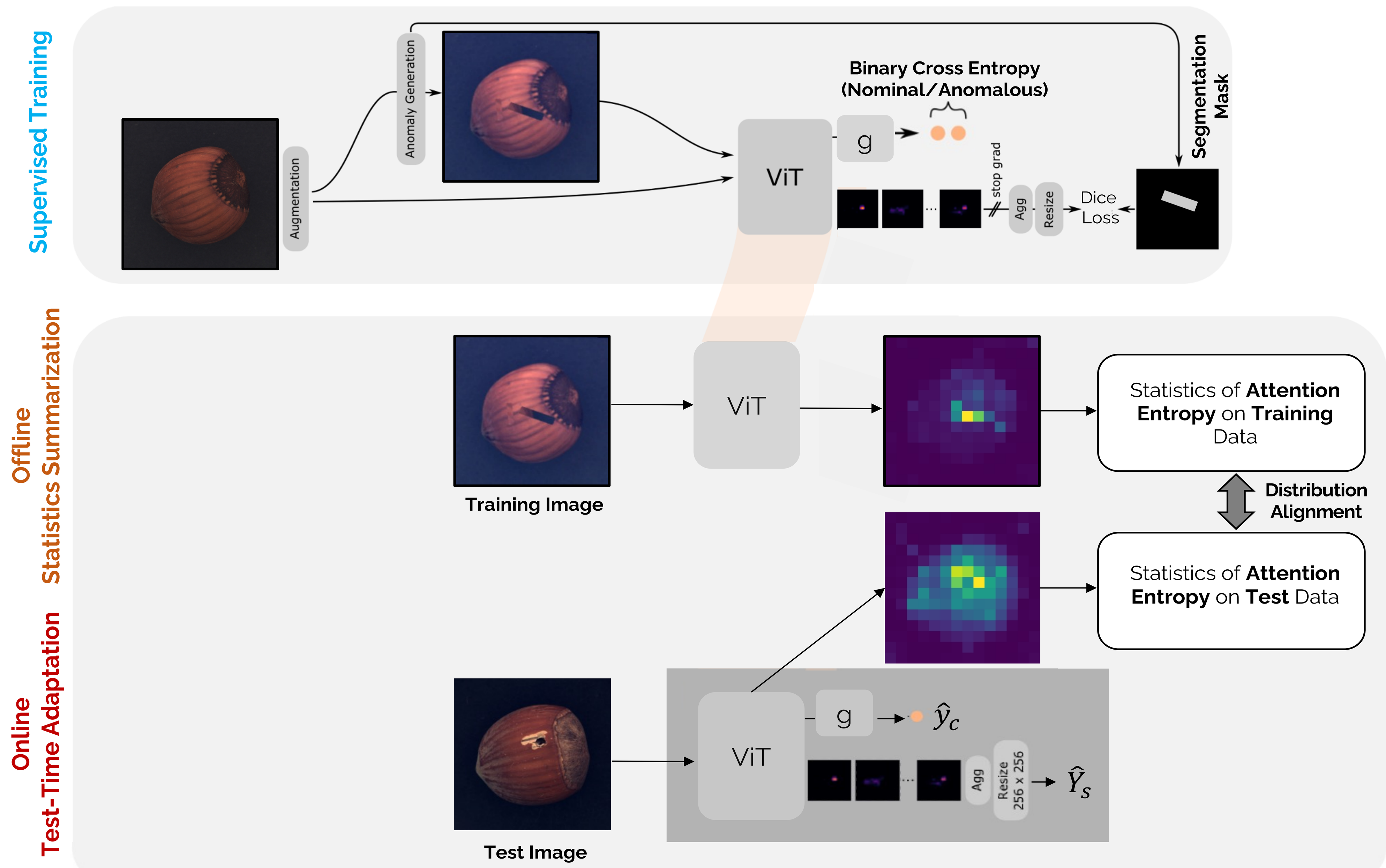




Proposed method

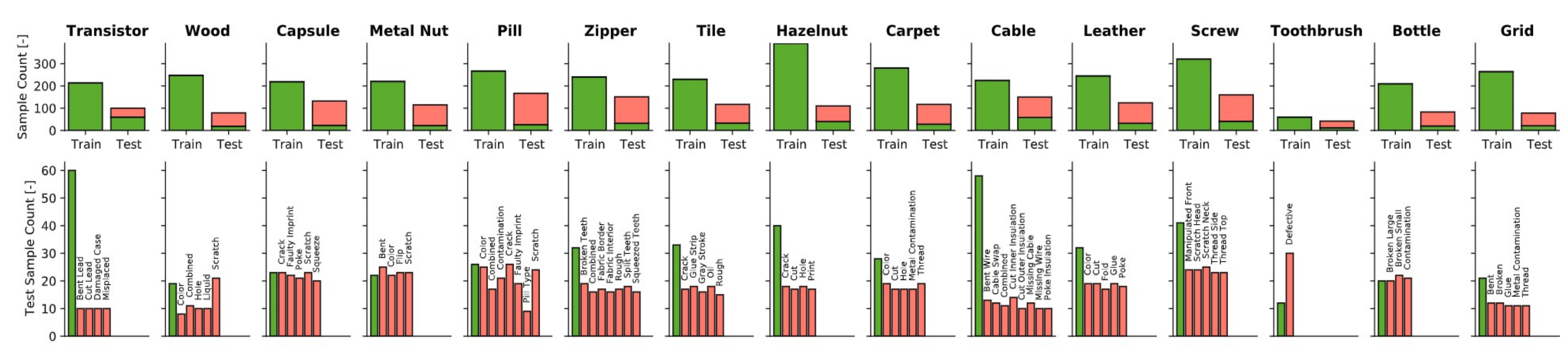
The proposed method is a three-stage framework:

I. The proxy task formulated as supervised training to detect and localize attention-guided synthetic anomalies generated from only nominal training data;
II. Offline statistics summarization (the **class-aware mean** and **second central moment** associated with transformer **attention entropy**) for the source training data;
III. Test-time adaptation, where we minimize the discrepancy of the distributional statistics for attention entropy between training and test data.



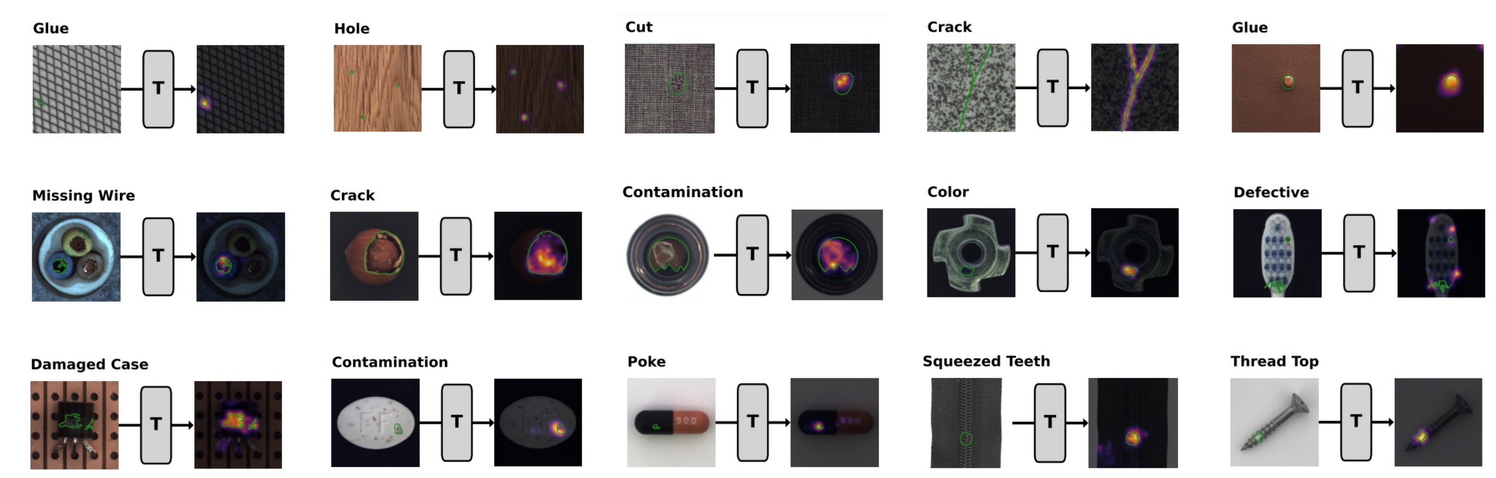
Datasets

- **NIH Chest X-ray dataset** comprises frontal-view X-ray images (1024 × 1024 pixels) labeled either as normal or with one or more of the 14 classes of thoracic diseases. The training set contains 50,500 anomaly-free X-ray images. The test set contains rough bounding box annotations of anomalies for 880 X-ray images (503 for male and 377 for female patients).
- **MVTec AD dataset** is composed of 15 categories (5 textures and 10 object categories) of industrial images with a total of 3629 anomaly-free training images and 1725 test images (700 × 700 ~ 1024 × 1024 pixels), including a mixture of anomaly-free images and various anomaly types. This dataset also contains pixel-level annotations for all defective areas.

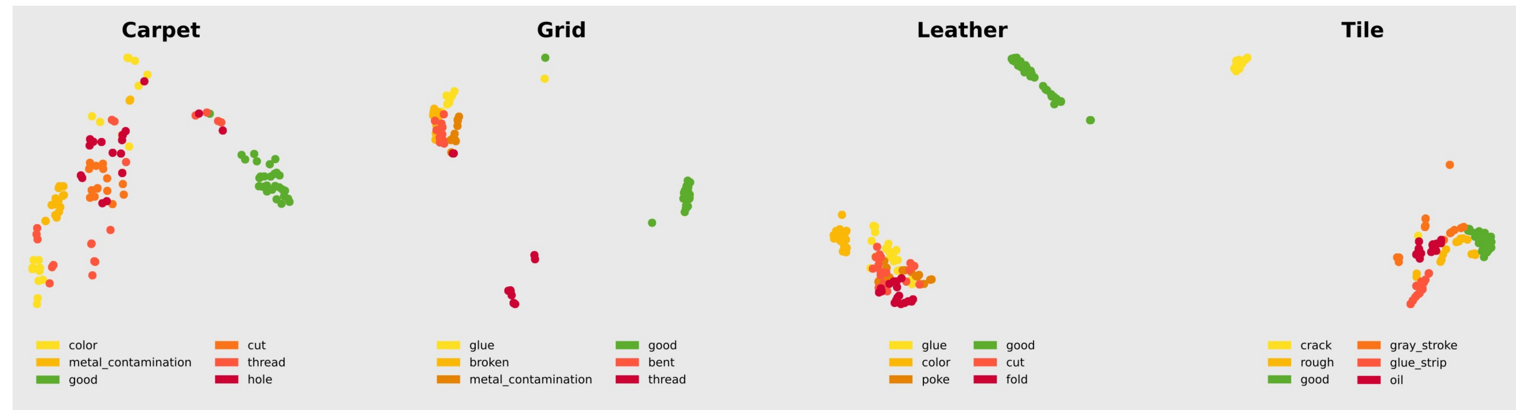


Qualitative results

Anomaly localization results from our method superimposed on the test images on the MVTec AD dataset. The green boundary denotes the ground-truth anomalies.



The t-SNE visualization of the learned features (before the projection head) on the MVTec AD dataset. The green dots represent nominal features for four categories. Results demonstrate well-separated feature distribution (normal vs. anomaly).



Quantitative results

- Our method achieves the highest average AUROC on the MVTec AD (**98.4%** AUROC on the image level and **98.2%** AUROC on the pixel level) compared to other baselines.
- Our method outperforms the CutPaste method and second-best method by a gain of **~20%** and **+8.7%** pixel-level AUROC on the NIH dataset.

Method	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	Hazelnut	Metal Nut	Pill	Screw	Toothbrush	Transistor	Zipper	Overall Average
	Textures					Objects										
Image-Level AUROC (in %)																
CutPaste (3-way) [18]	93.1	99.9	100.0	93.4	98.6	98.3	80.6	96.2	97.3	99.3	92.4	86.3	98.3	95.5	99.4	95.2
FPI [37]	56.0	99.5	91.7	90.2	74.4	90.2	68.0	87.5	86.0	88.4	71.8	61.2	85.8	79.6	97.7	81.9
PII [38]	65.6	100.0	100.0	98.4	91.9	97.6	68.9	84.9	82.7	98.9	86.3	74.7	93.1	90.1	99.8	88.9
NSA [35]	95.6	99.9	99.9	100.0	97.5	97.7	94.5	95.2	94.7	98.7	99.2	90.2	100.0	95.1	99.8	97.2
DRAEM[45]	97.0	99.9	100.0	99.6	99.1	99.2	91.8	98.5	100.0	98.7	98.9	93.9	100.0	93.1	100.0	98.0
Ours	100.0	99.7	99.8	99.7	96.3	99.1	95.8	97.6	99.7	99.8	98.1	96.5	98.5	95.9	99.6	**98.4**
Pixel-Level AUROC (in %)																
CutPaste (3-way) [18]	98.3	97.5	99.5	90.5	95.5	97.6	90.0	97.4	97.3	93.1	95.7	96.7	98.1	93.0	99.3	96.0
FPI [37]	70.8	94.2	88.3	65.0	71.1	91.8	66.5	95.9	89.8	96.2	62.3	90.4	81.8	78.5	91.8	82.3
PII [38]	97.2	98.9	99.2	98.0	91.1	93.1	70.2	90.2	97.0	95.4	95.3	92.8	81.3	86.9	93.8	92.0
NSA [35]	95.5	99.2	99.5	99.3	90.7	98.3	96.0	97.6	97.6	98.4	98.5	96.5	94.9	88.0	94.2	96.3
DRAEM [45]	95.5	99.7	98.6	99.2	96.4	99.1	94.7	94.3	99.7	99.5	97.6	97.6	98.1	90.9	98.8	97.3
Ours	99.2	98.4	99.4	97.6	97.0	97.6	98.2	98.6	98.3	98.6	98.5	99.3	98.1	95.1	99.1	**98.2**

Pixel-Level Anomaly Localization AUROC (in %)	CutPaste [18]	PaDiM [8]	FPI [37]	Ours w/o TTA	Ours
	Methods			Ablation for TTA	
Male ♂	52.6±1.3	54.2±0.8	63.4±0.9	70.8±0.8	**72.4±0.6**
Female ♀	51.8±1.2	53.8±0.9	62.9±1.1	70.4±0.9	**72.7±0.7**

Limitation and future work

- **Limitation**: Sometimes, salient regions generated by the transformer attention map have some randomness. This may deteriorate the distributional statistics alignment of attention entropy used for test-time adaptation.
- As the **future work**, we aim to create more realistic and diverse synthetic anomalies to further improve our method's generalizability.