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

Behzad Bozorgtabar\textsuperscript{1,2} \newline behzad.bozorgtabar@epfl.ch
Dwarikanath Mahapatra\textsuperscript{3} \newline dwarikanath.mahapatra@inceptioniai.org
Jean-Philippe Thiran\textsuperscript{1,2} \newline jean-philippe.thiran@epfl.ch
\hline
\textsuperscript{1} École Polytechnique Fédérale de Lausanne (EPFL) \newline Lausanne, Switzerland
\textsuperscript{2} Lausanne University Hospital (CHUV) \newline Lausanne, Switzerland
\textsuperscript{3} Inception Institute of Al (IIAI) \newline Abu Dhabi, UAE

Abstract

Despite the impressive success of vision transformers in various vision tasks, they are largely overlooked for anomaly detection and segmentation tasks. In this paper, we focus on the attention mechanism in the transformer and propose a new proxy task for model training followed by a test-time adaptation. In particular, we present a simple yet effective attention-guided cut-and-paste data augmentation for creating synthetic anomalies from nominal training images by intermixing scaled patches of various sizes guided by the transformer’s attention map. Subsequently, we solve a proxy task by discriminating between nominal examples and synthetic anomalies. Furthermore, to alleviate the distribution discrepancy between training and test data, we adopt a test-time adaptation scheme based on the transformer’s attention entropy. Extensive experimental results for anomaly detection and localization task on a popular MVTec AD benchmark and NIH Chest X-ray dataset demonstrate the superiority of our method over competitive baselines and its generalization capabilities to detect and localize test-time anomalies.

1 Introduction

The ability to detect and localize anomalies is a mainstay of many safety-critical applications, ranging from industrial defect detection [3] to out-of-distribution detection on medical images [49]. Due to tedious and costly annotation and also the absence of prior knowledge about types of anomalies, many deep methods, including unsupervised and self-supervised anomaly detection and segmentation algorithms [1, 2, 33], are formulated as one class classification setup, where the objective is to learn a distribution of the nominal samples and define an anomaly score to label those outside the learned distribution as anomalies.

Existing top-performing anomaly detection and localization methods [4, 5, 6] are owing to the use of deep discriminative multi-scale features from the pre-trained convolutional networks on ImageNet [8] with adaptation. Compared with fully convolutional networks,
vision transformers (ViTs) [10, 39] offer higher representation power due to their global receptive field. In particular, ViT models trained in self-supervised setup [5, 19] achieve better generalization and performance gain over convolutional networks. However, transformers largely remain unexplored for anomaly detection and segmentation task.

This paper presents a transformer-based network using a simple yet effective self-supervision task: detecting and localizing attention-guided synthetic anomalies for model training. In addition, a test-time adaptation scheme has been adapted to further improve the model’s generalization.

Contributions. Our contributions are summarized as follows:

- We propose a simple yet effective attention-guided cut-and-paste data augmentation for creating synthetic anomalies using only nominal training data. In particular, unlike previous augmentation strategies, which assume uniform relevance of image pixels for applying synthetic manipulations, we are intrigued by the self-attention mechanisms of the vision transformer and present attention-aware augmentation by intermixing patches only within a salient image region. The proposed scheme simulates spatial irregularities in the real anomaly and creates a more challenging proxy task;

- To further improve the model generalization and alleviates the mismatch between test samples and training data (e.g., the distributional gap between 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 empirically show consistent performance improvements over current synthetic anomaly augmentation methods [18, 35, 45] for anomaly detection and localization on the challenging benchmarks of the MVTec AD [3] and the NIH Chest X-rays [41]. We also demonstrate that test-time adaptation further improves model generalization.

2 Related Work

Anomaly Detection and Localization Methods. Existing deep learning methods for anomaly detection and localization often rely on the learning of distribution of normal images using lower-dimensional embedding [21, 33] or reconstructions derived from this embedding, e.g., autoencoder architectures [2, 23], generative adversarial networks (GANs) [1, 34], or normalizing flows [12, 32, 44]. Recently, vision transformers [10, 39], particularly those trained in self-supervised setup [5, 19], have shown superior performances compared to fully convolutional networks on various vision tasks. This is mainly due to the global receptive field of transformer architecture, yielding higher representation power. Nevertheless, vision transformers are largely overlooked in recent anomaly detection and localization algorithms, except for a few methods [28, 44]. Another category of anomaly detection and segmentation approaches utilize self-supervised learning by solving various auxiliary tasks such as matching different transformations of the same image [4, 5, 7], predicting geometric transformations [11, 15], colorization [48], or context prediction [25]. More recent self-supervised methods [18, 35, 37, 45] utilize data augmentation to create synthetic anomalies as proxy tasks. These methods either apply synthetic manipulations in the image at random locations [18] or randomly blend image patches from other images [35]. However, these techniques assume all image pixels are equally important for patch blending and manipu-
lation. Different from these approaches, our synthetic anomaly augmentation leverages the non-uniformity of image regions to generate useful anomalies for model training.

**Test-Time Adaptation Methods.** Test-time adaptation aims to enhance the robustness of the model against data shift by finetuning the trained source model using the unlabeled target samples during test time. Several recent works \cite{6, 24, 36, 40} have been proposed for online adapting of the leaned classifier to mitigate the distributional shift. Test-time training (TTT) \cite{36} proposes test-time model finetuning using the auxiliary task of the rotation prediction. Source hypothesis transfer (SHOT) \cite{20} exploits information maximization by optimizing entropy minimization and a diversity regularizer to tackle the domain shift between source and target data. Nevertheless, these approaches require additional architectural modifications \cite{36}. Some recent methods \cite{24, 40, 47} propose source-free test-time adaptation setup using only a trained model. For example, test entropy minimization (TENT) \cite{40} adapts the pre-trained model to the target data by continually updating the Batchnorm layer’s parameters and statistics \cite{40} via entropy minimization. Nonetheless, TENT may yield catastrophic failure due to miscalibrated predictions under a large domain shift. Unlike TENT, which modulates the model parameters only using the target data, similar to our test-time adaptation scheme, some recent methods, e.g., \cite{14} incorporate the statistics of source data rather than source data to minimize the distributional shift between data domains. More closely related to ours, \cite{13} extends a test-time adaptation to a ViT model, which minimizes the mismatch between source and target data distributions. While we adopt this scheme, our test-time adaptation differs from \cite{13} since we propose to match the selective distributional statistics by leveraging the statistics for each predicted class.

# 3 Methodology

This section presents a new proxy task for the transformer-based anomaly detection and localization method. As shown in Figure 1, the proposed method is a three-stage framework. It consists of the proxy task formulated as supervised training (Figure 1 (top)) to detect and localize attention-guided synthetic anomalies generated from only nominal training data, offline statistics summarization for the source training data (Figure 1 (middle)), and a test-time adaptation scheme (Figure 1 (bottom)).

We first discuss our proposed attention-guided synthetic anomalies in Section 3.1. In Section 3.2, we detail our model training using synthetic anomalies. Then, we describe the test-time adaptation scheme in Section 3.3. This is followed by the model evaluation and experimental results in Section 4. Finally, we conclude our work in Section 5.

## 3.1 Attention-Guided Synthetic Anomalies

**Self-Attention of the Class Token.** The ViT architecture \cite{10, 39} processes input image \( \mathbf{x} \in \mathbb{R}^{h \times w \times 3} \) with the \( h \times w \) spatial resolution by converting and embedding it to \( N \) patch tokens \( \mathbf{x}_{\text{patches}} \in \mathbb{R}^{N \times d} \), and aggregates the global information by a class ([CLS]) token \( \mathbf{x}_{[\text{CLS}]} \in \mathbb{R}^d \), which is then prepended to form patch embedding \( \mathbf{z} = [\mathbf{x}_{[\text{CLS}]}, \mathbf{x}_{\text{patches}}] \in \mathbb{R}^{(N+1) \times d} \). Given an input patch embedding \( \mathbf{z} \in \mathbb{R}^{(N+1) \times d} \), a multi-head self-attention (MSA) layer projects \( \mathbf{z} \) to the query \( \mathbf{Q}_j \), key \( \mathbf{K}_j \), and value \( \mathbf{V}_j \) sequences for \( j = 1, \ldots, L \), where \( L \) is the number of heads, \( \mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j \in \mathbb{R}^{N+1 \times d'} \) and \( d' = d/L \). Then, the self-attention can be
Figure 1: Schematic diagram of the model training and test-time adaptation. Top: The model is trained using nominal training images and generated attention-guided synthetic anomalies with corresponding segmentation masks. The attention maps are aggregated using the learnable weights (followed by resizing) into a single anomaly segmentation prediction. Bottom: After training, we compute and store the class-aware mean and second central moment associated with transformer attention entropy on training data. During test-time adaptation, we minimize the discrepancy of the distributional statistics for attention entropy between training and test data. For simplicity, we only show one global view for normal and anomalous images.

formulated as:

$$A_j = \text{softmax} \left( Q_j K_j^\top / \sqrt{d_j} \right)$$

where it forms the attention matrix $A_j$ for each head $j \in [L]$ using a row-wise softmax. We average across all $L$ attention heads to obtain the mean attention map $\bar{A}$. Then, we focus on the image patches that the $[\text{CLS}]$ token is attending denoted by $\bar{A}^{[\text{CLS}]} \in [0, 1]^N$, which is the first row of $\bar{A}$:

$$\bar{A} = \frac{1}{L} \sum_{j=1}^{L} A_j$$

$$\bar{A}^{[\text{CLS}]} = \{ \bar{A}_{1,i} | i \in [2, N + 1] \}$$

The vector $\bar{A}^{[\text{CLS}]}$ is then reshaped to $(h/s) \times (w/s)$ 2D attention map (using a patch size of $s \times s$ pixels).

Proposed Synthetic Anomalies. The rationale behind the proposed synthetic anomaly strategy is to create synthetic anomalies that are more relevant to the task by focusing on salient object regions derived from the self-attention of the transformer rather than irrelevant regions from the background. For instance, most defect categories of MVTec AD benchmark, e.g., transistor’s damaged leg, are around the salient object rather than the background.
Figure 2: The illustration of the proposed attention-guided synthetic anomaly generation: (a) the proposed scheme leverages ViT’s attention map to guide sampling of the informative locations to cut and paste patches; (b) pair examples of the synthetic anomalies produced using only nominal images on the MVTec AD dataset. Red arrows highlight the rotated pasted patches of scar shape.

The proposed synthetic anomaly generation is based on a cutting and pasting augmentation strategy. We harness the self-attention map corresponding to the [CLS] token of the last layer of the pre-trained ViT [5] (ViT Scorer) to learn the distribution of salient image regions. More precisely, we generate the mean self-attention map $\bar{A}$ for an input image by averaging across all attention heads (Equation 2). To compute the distribution of the salient image regions, we obtain the softmax of the mean attention map of $\bar{A}_{[CLS]} \in [0, 1]^N$ (Equation 3), which is then resized to the original input image dimensions. Afterward, we re-normalize the distribution, which can be seen as the distribution $P_{\text{location}}$ to guide sampling of the source locations $(c_x, c_y)_{\text{src}} \sim P_{\text{location}}$ and destination locations $(c_x, c_y)_{\text{dst}} \sim P_{\text{location}}$ for cutting and pasting operations (Figure 2 (left)). The proposed proxy task is then to detect and localize synthetic anomaly generated from anomaly-free training images as follows:

1. Compute the distribution of the salient regions from anomaly-free training image using the fixed ViT Scorer. To do so, estimate the softmax distribution of the $\bar{A}_{[CLS]}$.
2. Sample the source location $(c_x, c_y)_{\text{src}} \sim P_{\text{location}}$ for cutting a rectangle patch.
3. Select a rectangle patch of variable sizes and aspect ratios\(^1\) at the source location $(c_x, c_y)_{\text{src}}$ as the center of a patch. Optionally sample a scar shape (thin rectangle) of the image patch (Figure 2 (right)).
4. Optionally apply random rotation (from $-90^\circ$ to $90^\circ$), and color jittering to the patch.
5. Sample the destination location $(c_x, c_y)_{\text{dst}} \sim P_{\text{location}}$ as the patch’s center for pasting.
6. Paste a patch back to an anomaly-free image at the destination location $(c_x, c_y)_{\text{dst}}$.

3.2 Supervised Training Using Attention-Guided Synthetic Anomalies

We formulate supervised model training with the proxy task to simultaneously detect and localize synthetic anomalies. Given a set of anomaly-free training images $X^u$, for an input

\(^1\)By default, we sample the sizes for the width $r_w$ and height $r_h$ of the patches from a uniform distribution $\sim U(0.1W, 0.4W)$ for the image size of $W \times W$. 
training image $x \in X^u$, we create a synthetic anomaly using the proposed attention-guided approach (Section 3.1), denoted as $\text{Att}(x)$, where $\text{Att}(\cdot)$ is the attention-guided synthetic anomaly generation. Following the multi-crop scheme [4], we increase augmented images and randomly crop each input augmented image (normal or synthetic anomaly) into two large crops (global views) and eight small crops (local views). Subsequently, this creates a larger set of source training images $X^s$. We aim to identify the image-level class $y_c$ (normal/abnormal) for a given image and also predict the segmentation mask $Y_s$ corresponding to the anomaly pixels. For the $i^{th}$ training image, we assign the image-level label $y_{ci} = \{0, 1\}$. We set the label $y_{ci}$ for the anomaly-free image to 0 and 1 otherwise (synthetic anomaly). In addition, we obtain the segmentation masks $Y_s$ for anomalies by tracking where anomalies were pasted (pasted rectangle patch) during synthetic anomaly creation.

In this formulated binary classification and segmentation setup, a learner $\Phi$ estimates both classwise prediction and the corresponding anomaly segmentation maps $\{\hat{y}_c, \hat{Y}_s\}$. The architecture $\Phi$ (Figure 1 (top)) consists of a ViT encoder $f_\phi$, parameterized by $\phi$, which is initialized from DINO weights [4], multi-layer perceptron (MLP) projection head $g_\omega$, parameterized by $\omega$ for image-level classification, and a set of learnable weights for multi-head attention maps from the last layer of encoder $f_\phi$. The projection heads $g_\omega$ takes the $[\text{CLS}]$ token output of the ViT encoder and outputs two neurons. We define the training objective for the image-level binary classifier in detecting attention-guided synthetic anomalies as follows:

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

(4)

where $\text{CE}$ is the cross-entropy loss. We omit the augmented images in Equation 4 for simplicity, but we apply $\text{CE}$ loss for all augmented images.

For the anomaly localization, the attention maps of the ViT’s last layer are aggregated with the learned weights (followed by resizing), yielding a single anomaly segmentation map $\hat{Y}_s$. More precisely, the weight of each attention map is learned to maximize its anomaly localization ability by the Dice loss $L_{\text{Dice}}$, given the ground-truth masks of the synthetic anomalies. The attention maps are detached from the gradient graph, so the model only learns the best to average attention maps without influencing the ViT encoder. Combining the image-level synthetic anomaly detection loss $L_{\text{Att}}$ and the Dice loss $L_{\text{Dice}}$ for the pixel-wise anomaly localization, we derive the optimization problem using the proxy loss $L_{\text{Proxy}}$:

$$L_{\text{Proxy}} = (1 - \lambda) L_{\text{Att}} + \lambda L_{\text{Dice}}$$

(5)

where $\lambda \in [0, 1]$ is used to balance the loss terms.

3.3 Test-Time Adaptation

Due to discrepancies between source training data and target test data, the trained model may suffer from performance degradation. To prevent this, we adopt attention entropy-based test-time adaptation. The proposed model adaptation follows per category learning protocol.

Offline Statistics Summarization. We first perform offline source data statistics summarization step as auxiliary information regarding the distribution of source training data. Once training completes, given a sample image from source training data $x^s_i \in X^s$, we store
class-aware statistics, including the mean and second central moment associated with the transformer attention entropy. Let \( \hat{A}(x_i^t; \theta) \in \mathbb{R}^{N \times N} \) denote the learned aggregated attention weight matrix\(^2\) of the ViT encoder’s last layer after model training parameterized by \( \theta \). The attention entropy on the source sample \( x_i^t \) can be calculated for tokens \( j \in N \) as follows:

\[
\mathcal{H}_{i,j} = - \sum_{k=1}^{N} \hat{A}_{j,k}(x_i^t; \theta) \log \hat{A}_{j,k}(x_i^t; \theta) \tag{6}
\]

Then we calculate and store in memory the class-aware mean \( \mu_c^x \) and second central moment \( M_c^x \) associated with attention entropy for source training data as follows:

\[
\mu_c^x = \text{Concat}_{j \in N} \left( \frac{1}{|X_c^s|} \sum_{x_i^t \in X_c^s} \mathcal{H}_{i,j} \right), \quad c = (1, 2) \tag{7}
\]

\[
M_c^x = \text{Concat}_{j \in N} \left( \frac{1}{|X_c^s|} \sum_{x_i^t \in X_c^s} \left( \mathcal{H}_{i,j} - \mu_c^x \right)^2 \right) \tag{8}
\]

where \( X_c^s \subset X^s \) contains all the source training images whose labels are \( c \) (either of two classes of nominal or synthetic anomaly). \( \mu_c^{x,j} = \frac{1}{|X_c^s|} \sum_{x_i^t \in X_c^s} \mathcal{H}_{i,j} \), and Concat denotes the concatenation operation along the token dimension.

**Test-Time Adaptation Using Class-Aware Statistics Alignment.** At test time, our model sequentially processes a mini-batch of test images from the target dataset \( X^t = \{x_i^t\}_{i=1}^{N_t} \) and is adapted to minimize the distance between class-aware statistics estimated from the mini-batch of test images and the stored fixed values obtained from source training data. Let \( X_m^t \subset X^t \), \((m = 1, \ldots, M)\) denote the \( m^{th} \) mini-batch of the target test data. We first obtain a subset of \( X_c^t, m \subset X_m^t \), which includes all unlabeled test images in the current mini-batch \( X_m^t \), which are assigned to class \( c \) by pseudo labeling: \( \arg\max_c g_\omega(f_\phi(x_i^t)) \). Then, for each test image \( x_i^t \in X_c^t, m \), similar to training data, we compute the attention entropy as in Equation 9:

\[
\mathcal{H}_{i,j}^{t,m} = - \sum_{k=1}^{N} \hat{A}_{j,k}(x_i^t; \tilde{\theta}) \log \hat{A}_{j,k}(x_i^t; \tilde{\theta}) \, , \quad x_i^t \in X_c^{t,m} \tag{9}
\]

where \( \hat{A}(x_i^t; \tilde{\theta}) \) denotes the aggregated attention weight matrix parameterized by \( \tilde{\theta} \) during adaptation. Subsequently, the class-aware mean \( \mu_c^x \) and second central moment \( M_c^x \) of the attention entropy are computed for \( m^{th} \) mini-batch of target test data as follows:

\[
\mu_c^{t,m} = \text{Concat}_{j \in N} \left( \frac{1}{|X_c^{t,m}|} \sum_{x_i^t \in X_c^{t,m}} \mathcal{H}_{i,j}^{t,m} \right), \quad c = (1, 2) \tag{10}
\]

\[
M_c^{t,m} = \text{Concat}_{j \in N} \left( \frac{1}{|X_c^{t,m}|} \sum_{x_i^t \in X_c^{t,m}} \left( \mathcal{H}_{i,j}^{t,m} - \mu_c^{t,m} \right)^2 \right) \tag{11}
\]

\(^2\hat{A}\) denotes the learned attention weight aggregated over all heads without the first entry. Alternatively, the most representative head can be used.
where $\mu^{t,m,j}_{c} = \frac{1}{|X^{t}|} \sum_{x^{t}_i \in X^{t}} \mathcal{H}^{t,m}_{i,j}$. Given statistics computed from source training data and mini-batch of test samples, a test-time adaptation loss $L_{TTA}$ is defined as follows:

$$L_{TTA} = \frac{1}{|C'|} \left( \frac{1}{\log N} \sum_{c \in C'} \| \mu^{s}_{c} - \mu^{t,m}_{c} \|_2 + \frac{1}{(\log N)^2} \sum_{c \in C'} \| M^{s}_{c} - M^{t,m}_{c} \|_2 \right)$$

(12)

where $\log N$ is the maximum value of the attention entropy, and $C'$ denotes a set of the pseudo-labeled classes in the current mini-batch of test images.

### 4 Experiments

**Training Setup and Metrics.** We use PyTorch 1.9 [26] and train each model on a single GeForce RTX 2080 Ti GPU. For the transformer encoder $f$, we use a ViT-small (ViT-S/16) initialized from DINO weights [5]. The optimization is performed using the stochastic gradient descent (SGD) with a momentum of 0.9 and gradient clipping at global norm 1.0 for model training and test-time adaptation. We use a batch size of 16 and a learning rate of $5 \times 10^{-4}$. The learning rate is linearly warmed-up during the first ten epochs and then follows a cosine schedule [22], and the total number of iterations is 1800 during adaptation. Two global views of $224 \times 224$ pixels and eight local views of $96 \times 96$ pixels are constructed. For the evaluation metrics, we use the area under the receiver operating characteristic curve (AUROC) for image-level anomaly detection and pixel-wise AUROC for anomaly localization. We set $\lambda = 0.05$ using a hyper-parameter search $\lambda \in \{0.01, 0.05, 1.0, 5.0, 1\}$.

### 4.1 Datasets and Experimental Results

**MVTec AD dataset** [3] contains 15 categories (10 object categories and 5 texture categories) of industrial images with a total of 3629 anomaly-free training images and 1725 test images ($700 \times 700 \sim 1024 \times 1024$ pixels), including a mixture of anomaly-free images and various anomaly types. This dataset also contains pixel-level annotations for all defective areas. In Table 1, we conduct performance comparisons of our method after test-time adaptation against prior art anomaly detection and localization methods on the MVTec AD dataset. The competitive baselines include state-of-the-art synthetic anomaly-based [18, 35, 37, 38, 45], self-supervised, e.g., [28, 43, 46] or transformer-based method [28], and methods transferring pre-trained representations [8, 29, 32] from ImageNet. The detailed comparison results on the MVTec AD show that our method surpasses prior art methods and achieves the highest average AUROC (98.4% AUROC on image level and 98.2% AUROC on pixel-level) among all categories for both anomaly detection and localization tasks. It has been shown in [28] that recent self-supervised anomaly detection methods still lag behind methods using pre-trained ImageNet with knowledge transfer/distillation. Nevertheless, even though we initialize our model using DINO weights [3] pre-trained only on unlabeled images, our method outperforms recent transfer learning methods [29, 32] that benefit from supervised pre-trained networks on ImageNet. These quantitative results are supported by the qualitative results of precise anomaly localization in Figure 3 on the MVTec AD test set.

**NIH dataset** [41] comprises frontal-view X-ray images ($1024 \times 1024$ pixels) labeled either as normal or with one or more of the 14 classes of thoracic diseases from 30,805 patients.
Table 1: Performance comparison with the prior art for anomaly detection (image-level AUROC %) and localization (pixel-level AUROC %) on the MVTeC AD dataset. SA. denotes synthetic anomaly-based methods. SSL & IN. denote self-supervised methods and models pretrained on ImageNet (highlighted by $\xi$). The best average results are in **bold**.

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Textures</th>
<th>Image-Level AUROC (in %)</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>CutPaste (3-way)</td>
<td>RoResNet16-Upcpling</td>
<td>91.1</td>
<td>99.9 100.0 95.4 98.6 98.1 96.6 98.2</td>
<td>97.3 99.3 92.4 86.5 88.5</td>
</tr>
<tr>
<td>FPI [21]</td>
<td>Residual Encoder-Decoder</td>
<td>56.0 99.3</td>
<td>91.7 92.2 74.4 70.2 68.0 67.5</td>
<td>86.0 84.8 71.4 61.2 58.5</td>
</tr>
<tr>
<td>IBM [24]</td>
<td>Residual Encoder-Decoder</td>
<td>65.4 100.0</td>
<td>100.0 99.4 91.9 97.6 68.0 84.9</td>
<td>82.7 92.9 69.3 67.1 97.1</td>
</tr>
<tr>
<td>NMA [32]</td>
<td>Residual-Encoder-ResNet18</td>
<td>93.6 99.9</td>
<td>99.0 100.0 97.5 97.7 94.5 95.2</td>
<td>95.4 92.7 99.2 96.3 97.8</td>
</tr>
<tr>
<td>DRADDM [35]</td>
<td>Residual Encoder-Decoder</td>
<td>97.0 99.9</td>
<td>100.0 99.6 99.1 99.2 91.8 98.5</td>
<td>100.0 98.7 88.9 97.9 100.0</td>
</tr>
</tbody>
</table>

**Overall Average**

| FGMD $\ddagger$ | EffiResNet-B5 | 97.9 |
| Patch2VID [13] | Convolutional Backbone | 92.9 94.6 | 99.9 97.8 98.6 96.8 98.0 95.7 | 92.0 94.0 89.1 81.3 100.0 | 91.3 97.9 94.4 |
| DifferNet $\ddagger$ | AlexNet | 92.9 84.0 | 97.1 99.4 99.8 97.5 99.0 88.9 | 98.5 95.9 95.9 98.9 | 91.8 98.0 94.8 |
| SPADD [16]      | RoResNet-152 | 98.6 90.0 | 99.9 98.8 95.8 98.1 93.2 98.6 | 98.9 89.9 96.5 99.9 98.9 | 91.0 98.8 96.2 |
| Info [22]       | Modified ViT-B/16 | 98.8 100.0 | 100.0 99.2 97.5 100.0 78.5 94.3 | 99.7 90.6 96.2 95.7 100.0 | 98.8 99.4 98.0 |

| Ours            | ViT-B/16       | 100.0 | 99.7 99.8 99.7 99.1 97.4 97.6 97.6 99.7 99.0 96.1 98.5 | 93.9 99.6 98.4 |

**Overall Average**

| FGMD $\ddagger$ | EffiResNet-B5 | 97.9 |
| Patch2VID [13] | Convolutional Backbone | 92.9 94.6 | 99.9 97.8 98.6 96.8 98.0 95.7 | 92.0 94.0 89.1 81.3 100.0 | 91.3 97.9 94.4 |
| DifferNet $\ddagger$ | AlexNet | 92.9 84.0 | 97.1 99.4 99.8 97.5 99.0 88.9 | 98.5 95.9 95.9 98.9 | 91.8 98.0 94.8 |
| SPADD [16]      | RoResNet-152 | 98.6 90.0 | 99.9 98.8 95.8 98.1 93.2 98.6 | 98.9 89.9 96.5 99.9 98.9 | 91.0 98.8 96.2 |
| Info [22]       | Modified ViT-B/16 | 98.8 100.0 | 100.0 99.2 97.5 100.0 78.5 94.3 | 99.7 90.6 96.2 95.7 100.0 | 98.8 99.4 98.0 |

| Ours            | ViT-B/16       | 100.0 | 99.7 99.8 99.7 99.1 97.4 97.6 97.6 97.6 99.7 99.0 96.1 98.5 | 93.9 99.6 98.4 |

Table 2: Performance comparison with recent synthetic anomaly augmentation-based methods for anomaly localization (pixel-level AUROC %), and standard error across five different random seeds on the NIH Chest X-ray dataset.

<table>
<thead>
<tr>
<th>Pixel-Level Anomaly Localization AUROC (in %)</th>
<th>CutPaste $\ddagger$</th>
<th>PaDM $\ddagger$</th>
<th>FPI [21]</th>
<th>Ours w/o TTA</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male $\sigma$</td>
<td>52.6±1.3</td>
<td>34.2±0.8</td>
<td>63.4±0.9</td>
<td>70.8±0.8</td>
<td>72.4±0.6</td>
</tr>
<tr>
<td>Female $\varphi$</td>
<td>51.8±1.2</td>
<td>53.8±0.9</td>
<td>62.9±1.1</td>
<td>70.4±0.9</td>
<td>72.7±0.7</td>
</tr>
</tbody>
</table>

The training set contains 50,500 anomaly-free X-ray images, and the test set includes 25,595 X-rays (15,735 normal and 9,860 anomalous images). The test set contains rough bounding box annotations of anomalies for 880 X-ray images (503 for male and 377 for female patients). In Table 2, we show the generalization capability of our method beyond industrial images. We provide additional quantitative results by comparing our method against the recent synthetic anomaly augmentation-based methods [8, 13, 26] for anomaly localization task on the NIH Chest X-rays [33]. The competitive self-supervised method, FPI [21], performs well on this task by using Poisson image editing [26] designed for localizing more subtle anomalies. Nonetheless, our method trained with the proposed attention-guided synthetic anomaly notably outperforms the second-best method, FPI [21], (~ 9% pixel-level AUROC) and creates more task relevant synthetic anomalies.

4.2 Ablation Studies

We investigate the impact of test-time adaptation (TTA) on the final model’s performance by comparing our trained model adapted with various TTA methods: TENT [33], pseudo label (PL) [34], and a variant of our method without any adaptation (Ours w/o TTA). We use the same architecture and hyperparameters for a fair comparison across all the baselines. We only update the parameters of the classification projection head and modulate the layer normalization parameters of the ViT encoder while other architecture parameters remain unaffected. The experimental results on the MVTeC AD (Table 3) demonstrate that our final model using TTA scheme (Ours) outperforms the other baselines, e.g., TENT optimized...
Anomaly localization examples

by the entropy distribution of image classification. In addition, together with the results in Table 2, it is verified that an adapted model using the TTA scheme can yield performance gain solely compared to a model without adaptation (Ours w/o TTA).

Furthermore, we evaluate the effect of the data augmentation for creating synthetic anomalies by using different augmentation methods (Aug-CutPaste [18], Aug-NSA [35], and Aug-FPA [37]) during model training on source data. We use the same TTA setup for all baselines. The superior results in Table 3 over current augmentation methods, e.g., CutPaste [18], verify that focusing on salient image regions and alleviating the unfounded assumption of uniform relevance of patches, we can generate more realistic synthetic anomalies. Moreover, adding a multi-crop scheme can improve the model performance without the noticeable overhead.

5 Conclusion

We propose a transformer-based method based on a new self-supervised proxy task for anomaly detection and localization. We leverage the attention map of a transformer to account for the non-uniform relevance of patches for creating synthetic anomalies, simulating natural spatial irregularities. Furthermore, we adopt test-time adaptation to reduce the distributional differences between training and test data based on the transformer’s attention entropy statistics, yielding better generalization to detect and localize real anomalies.

The rationale behind the proposed synthetic anomaly is that most anomalous categories, e.g., defects’ types, are not random and reside within salient foreground objects. Nonetheless, a limitation that remains is that 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. Our future work focuses on addressing this problem. Furthermore, this paper opens up a few interesting directions for future research. First, we aim to create more realistic and diverse synthetic anomalies to further improve our method’s generalizability. Second, we explore different supervisory signals used for test-time adaptation [33, 40], which can be complementary to ours.
Acknowledgments The authors would like to thank Antoine Spahr for helping with the experiments and implementing the proposed synthetic anomalies.

References


