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

Behzad Bozorgtabar[1,2]
behzad.bozorgtabar@epfl.ch
Dwarikanath Mahapatra[3]
dwarikanath.mahapatra@inceptioniai.org
Jean-Philippe Thiran[1,2]
jean-philippe.thiran@epfl.ch

[1] École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
[2] Lausanne University Hospital (CHUV)
Lausanne, Switzerland
[3] Inception Institute of AI (IIAI)
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 [17, 31, 44] are owing to the use of deep discriminative multi-scale features from the pre-trained convolutional networks on ImageNet [9] 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 [6, 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 [6, 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 manipulation. 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 [6, 24, 36, 40] have been proposed for online adapting of the leaned classifier to mitigate the distributional shift. Test-time training (TTT) [36] proposes test-time model finetuning using the auxiliary task of the rotation prediction. Source hypothesis transfer (SHOT) [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 [36]. Some recent methods [24, 40, 47] propose source-free test-time adaptation setup using only a trained model. For example, test entropy minimization (TENT) [40] adapts the pre-trained model to the target data by continually updating the Batchnorm layer's parameters and statistics [42] 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., [14] incorporate the statistics of source data rather than source data to minimize the distributional shift between data domains. More closely related to ours, [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 [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 [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 $\mathbf{x}_{\texttt{patches}} \in \mathbb{R}^{N \times d}$, and aggregates the global information by a *class* (`[CLS]`) token $\mathbf{x}_{\texttt{[CLS]}} \in \mathbb{R}^{d}$, which is then prepended to form patch embedding $\mathbf{z} = [\mathbf{x}_{\texttt{[CLS]}}; \mathbf{x}_{\texttt{patches}}] \in \mathbb{R}^{(N+1) \times d}$. 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 $j = 1, \cdots, L$, 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

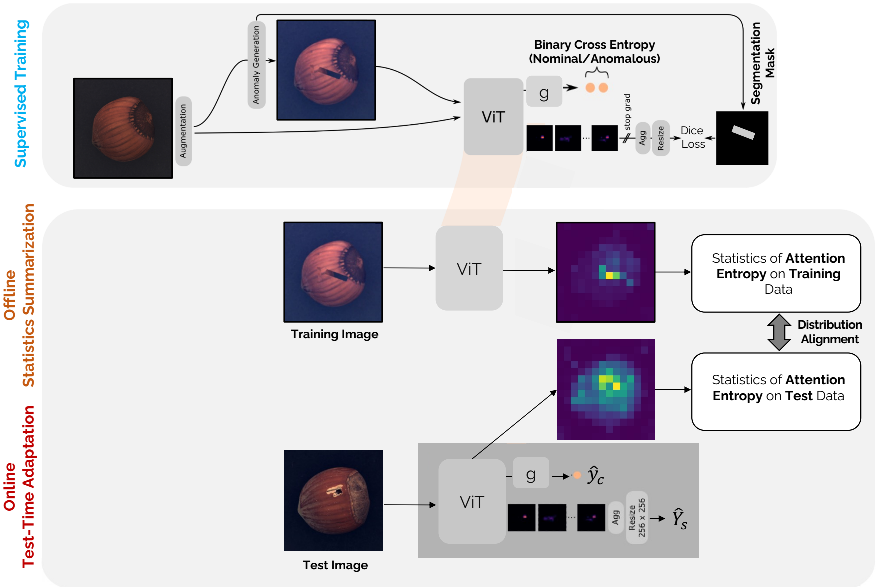


Figure 1: **Schematic diagram of the model training and test-time adaptation. Top**: The model is trained using nominal training images and generated attention-guided synthetic anomalies with corresponding segmentation masks. The attention maps are aggregated using the learnable weights (followed by 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:

$$\mathbf{A}_j = \text{softmax}\left(\mathbf{Q}_j \mathbf{K}_j^\top / \sqrt{d'}\right) \tag{1}$$

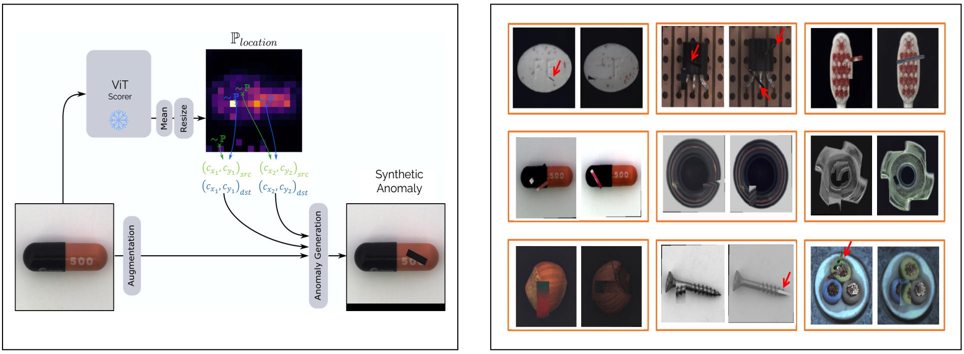
where it forms the *attention matrix* $\mathbf{A}_j$ for each head $j \in [L]$ using a row-wise `softmax`. We average across all L attention heads to obtain the mean attention map $\bar{\mathbf{A}}$. Then, we focus on the image patches that the `[CLS]` token is attending denoted by $\bar{\mathbf{A}}^{\texttt{[CLS]}}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

$$\bar{\mathbf{A}} = \frac{1}{L}\sum_{j=1}^{L} \mathbf{A}_j \tag{2}$$

$$\bar{\mathbf{A}}^{\texttt{[CLS]}} = \left\{\bar{\mathbf{A}}_{1,i} \mid i \in [2, N+1]\right\} \tag{3}$$

The vector $\bar{\mathbf{A}}^{\texttt{[CLS]}}$ 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



(a) (b)

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.

(cf. Figure 2). Detecting and localizing synthetic anomalies can be a practical proxy task for model training, bypassing the need for labeled data of natural anomalies.

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 $\bar{\mathbf{A}}$ 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 `[CLS]` token $\bar{\mathbf{A}}^{\texttt{[CLS]}} \in [0,1]^N$ (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 $\mathbb{P}_{location}$ to guide sampling of the source locations $(c_x,c_y)_{src} \sim \mathbb{P}_{location}$ and destination locations $(c_x,c_y)_{dst} \sim \mathbb{P}_{location}$ 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 $\bar{\mathbf{A}}^{\texttt{[CLS]}}$.
2. Sample the source location $(c_x,c_y)_{src} \sim \mathbb{P}_{location}$ for cutting a rectangle patch.
3. Select a rectangle patch of variable sizes and aspect ratios[1] at the source location $(c_x,c_y)_{src}$ 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° to 90°), and color jittering to the patch.
5. Sample the destination location $(c_x,c_y)_{dst} \sim \mathbb{P}_{location}$ as the patch's center for pasting.
6. Paste a patch back to an anomaly-free image at the destination location $(c_x,c_y)_{dst}$.

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 $\sim \mathcal{U}(0.1W, 0.4W)$ for the image size of $W \times W$.

training image $\mathbf{x} \in X^u$, we create a synthetic anomaly using the proposed attention-guided approach (Section 3.1), denoted as $\mathbf{Att}(\mathbf{x})$, where $\mathbf{Att}(\cdot)$ 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 $\mathbf{y}_c$ (normal/abnormal) for a given image and also predict the segmentation mask $\mathbf{Y}_s$ corresponding to the anomaly pixels. For the i^{th} training image, we assign the image-level label $\mathbf{y}_{ci} = \{\mathbf{0}, \mathbf{1}\}$. We set the label $\mathbf{y}_{ci}$ for the anomaly-free image to $\mathbf{0}$ and $\mathbf{1}$ otherwise (synthetic anomaly). In addition, we obtain the segmentation masks $\mathbf{Y}_s$ 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 Φ estimates both classwise prediction and the corresponding anomaly segmentation maps $\{\hat{\mathbf{y}}_c, \hat{\mathbf{Y}}_s\}$. The architecture Φ (Figure 1 (**top**)) consists of a ViT encoder f_ϕ, parameterized by ϕ, which is initialized from DINO weights [5], *multi-layer perceptron* (MLP) projection head g_ω, parameterized by ω for image-level classification, and a set of *learnable weights* for multi-head attention maps from the last layer of encoder f_ϕ. The projection heads g_ω takes the `[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:

$$\mathcal{L}_{\text{Att}} = \frac{1}{2|X^u|} \sum_{\mathbf{x} \in X^u} [\mathbb{CE}(g(f(\mathbf{x})), 0) + \mathbb{CE}(g(f(\mathbf{Att}(\mathbf{x}))), 1)] \tag{4}$$

where $\mathbb{CE}$ is the cross-entropy loss. We omit the augmented images in Equation 4 for simplicity, but we apply $\mathbb{CE}$ 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 $\hat{\mathbf{Y}}_s$. More precisely, the weight of each attention map is learned to maximize its anomaly localization ability by the Dice loss $\mathcal{L}_{\text{Dice}}$, 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 $\mathcal{L}_{\text{Att}}$ and the Dice loss $\mathcal{L}_{\text{Dice}}$ for the pixel-wise anomaly localization, we derive the optimization problem using the proxy loss $\mathcal{L}_{\text{Proxy}}$:

$$\mathcal{L}_{\text{Proxy}} = (1-\lambda)\mathcal{L}_{\text{Att}} + \lambda \mathcal{L}_{\text{Dice}} \tag{5}$$

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 $\mathbf{x}_i^s \in X^s$, we store

class-aware statistics, including the *mean* and *second central moment* associated with the transformer *attention entropy*. Let $\hat{\mathbf{A}}\left(\mathbf{x}_i^s;\theta\right) \in \mathbb{R}^{N\times N}$ denote the learned aggregated attention weight matrix[2] of the ViT encoder's last layer after model training parameterized by θ. The *attention entropy* on the source sample $\mathbf{x}_i^s$ can be calculated for tokens $j \in N$ as follows:

$$\mathcal{H}_{i,j}^s = -\sum_{k=1}^{N} \hat{\mathbf{A}}_{j,k}\left(\mathbf{x}_i^s;\theta\right) \log \hat{\mathbf{A}}_{j,k}\left(\mathbf{x}_i^s;\theta\right) \tag{6}$$

Then we calculate and store in memory the class-aware *mean* μ_c^s and *second central moment* $\mathbb{M}_c^s$ associated with *attention entropy* for source training data as follows:

$$\mu_c^s = \underset{j\in N}{\texttt{Concat}}\left(\frac{1}{|X_c^s|}\sum_{\mathbf{x}_i^s\in X_c^s}\mathcal{H}_{i,j}^s\right), \qquad c=(1,2) \tag{7}$$

$$\mathbb{M}_c^s = \underset{j\in N}{\texttt{Concat}}\left(\frac{1}{|X_c^s|}\sum_{\mathbf{x}_i^S\in X_c^s}\left(\mathcal{H}_{i,j}^s-\mu_c^{s,j}\right)^2\right) \tag{8}$$

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). $\mu_c^{s,j} = \frac{1}{|X_c^s|}\sum_{\mathbf{x}_i^s\in X_c^s}\mathcal{H}_{i,j}^s$, 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 $X^t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ 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 $X^{t,m} \subset X^t, (m=1,\cdots,M)$ denote the m^{th} mini-batch of the target test data. We first obtain a subset of $X_c^{t,m} \subset X^{t,m}$, which includes all unlabeled test images in the current mini-batch $X^{t,m}$, which are assigned to class c by pseudo labeling: $\arg\max_c g_\omega\left(f_\phi\left(\mathbf{x}_i^t\right)\right)$. Then, for each test image $\mathbf{x}_i^t \in X_c^{t,m}$, similar to training data, we compute the *attention entropy* as in Equation 9:

$$\mathcal{H}_{i,j}^{t,m} = -\sum_{k=1}^{N} \hat{\mathbf{A}}_{j,k}\left(\mathbf{x}_i^t;\bar{\theta}\right) \log \hat{\mathbf{A}}_{j,k}\left(\mathbf{x}_i^t;\bar{\theta}\right), \qquad \mathbf{x}_i^t \in X_c^{t,m} \tag{9}$$

where $\hat{\mathbf{A}}\left(\mathbf{x}_i^t;\bar{\theta}\right)$ denotes the aggregated attention weight matrix parameterized by $\bar{\theta}$ during adaptation. Subsequently, the class-aware *mean* μ_c^s and *second central moment* $\mathbb{M}_c^{t,m}$ of the *attention entropy* are computed for m^{th} mini-batch of target test data as follows:

$$\mu_c^{t,m} = \underset{j\in N}{\texttt{Concat}}\left(\frac{1}{\left|X_c^{t,m}\right|}\sum_{\mathbf{x}_i^t\in X_c^{t,m}}\mathcal{H}_{i,j}^{t,m}\right), \qquad c=(1,2) \tag{10}$$

$$\mathbb{M}_c^{t,m} = \underset{j\in N}{\texttt{Concat}}\left(\frac{1}{\left|X_c^{t,m}\right|}\sum_{\mathbf{x}_i^t\in X_c^{t,m}}\left(\mathcal{H}_{i,j}^{t,m}-\mu_c^{t,m,j}\right)^2\right) \tag{11}$$

[2] $\hat{\mathbf{A}}$ denotes the learned attention weight aggregated over all heads without the first entry. Alternatively, the most representative head can be used.

where $\mu_c^{t,m,j} = \frac{1}{|X_c^{t,m}|} \sum_{\mathbf{x}_i^t \in X_c^{t,m}} \mathcal{H}_{i,j}^{t,m}$. Given statistics computed from source training data and mini-batch of test samples, a test-time adaptation loss $\mathcal{L}_{\text{TTA}}$ is defined as follows:

$$\mathcal{L}_{\text{TTA}} = \frac{1}{|C'|} \left(\frac{1}{\log N} \sum_{c \in C'} \left\| \mu_c^s - \mu_c^{t,m} \right\|_2 + \frac{1}{(\log N)^2} \sum_{c \in C'} \left\| \mathbb{M}_c^s - \mathbb{M}_c^{t,m} \right\|_2 \right) \tag{12}$$

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 $5e-4$. 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 224×224 pixels and eight local views of 96×96 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 $\lambda \in \{.01, \underline{.05}, .1, .5, 1\}$.

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 [29] 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 [5] 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 (1024×1024 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 ‡). The best average results are in **bold**.

	Method	Architecture	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	Hazelnut	Metal Nut	Pill	Screw	Toothbrush	Transistor	Zipper	Overall Average
			Textures					Objects										
			Image-Level AUROC (in %)															
SA.	CutPaste (3-way) [■]	ResNet18+Upsampling	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
SA.	FPI [■]	Residual Encoder-Decoder	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
SA.	PII [■]	Residual Encoder-Decoder	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
SA.	NSA [■]	Encoder-Decoder,ResNet18	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
SA.	DRAEM[■]	Encoder-Decoder	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
SSL & IN.	PaDiM‡ [■]	EfficientNet-B5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	97.9
SSL & IN.	PatchSVDD [■]	Convolutional Backbone	92.9	94.6	90.9	97.8	96.5	98.6	90.3	76.7	92.0	94.0	86.1	81.3	100.0	91.5	97.9	92.1
SSL & IN.	DifferNet‡ [■]	AlexNet	92.9	84.0	97.1	99.4	99.8	99.0	95.9	86.9	99.3	96.1	88.8	96.3	98.6	91.1	95.1	94.9
SSL & IN.	SPADE‡ [■]	ResNet-152	98.6	99.0	99.5	89.8	95.8	98.1	93.2	98.6	98.9	96.9	96.5	99.5	98.9	81.0	98.8	96.2
SSL & IN.	InTra [■]	Modified ViT-B/16	98.8	100.0	100.0	98.2	97.5	100.0	70.3	86.5	95.7	96.9	90.2	95.7	100.0	95.8	99.4	95.0
	Ours	ViT-S/16	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 %)															
SA.	CutPaste (3-way) [■]	ResNet-18+Upsampling	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
SA.	FPI [■]	Residual Encoder-Decoder	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
SA.	PII [■]	Residual Encoder-Decoder	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
SA.	NSA [■]	Encoder-Decoder,ResNet18	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
SA.	DRAEM [■]	Encoder-Decoder	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
SSL & IN.	PaDiM‡ [■]	EfficientNet-B5	99.1	97.3	99.2	94.1	94.9	98.3	96.7	98.5	98.2	97.2	95.7	98.5	98.8	97.5	98.5	97.5
SSL & IN.	PatchSVDD [■]	Convolutional Backbone	92.6	96.2	97.4	91.4	90.8	98.1	96.8	95.8	97.5	98.0	95.1	95.7	98.1	97.0	95.1	95.7
SSL & IN.	SPADE‡ [■]	ResNet-152	97.5	93.7	97.6	87.4	88.5	98.4	97.2	99.0	99.1	98.1	96.5	98.9	97.9	94.1	96.5	96.5
SSL & IN.	RIAD [■]	U-Net [■]	96.3	98.8	99.4	89.1	88.8	98.4	84.2	92.8	96.1	92.5	95.7	98.8	98.9	87.7	97.8	94.2
SSL & IN.	InTra [■]	Modified ViT-B/16	99.2	98.8	99.5	94.4	88.7	97.1	91.0	97.7	98.3	93.3	98.3	99.5	98.9	96.1	99.2	96.6
	Ours	ViT-S/16	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**

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.

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**

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 abnormal 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, 18, 37] for anomaly localization task on the NIH Chest X-rays [41]. The competitive self-supervised method, FPI [37], performs well on this task by using Poisson image editing [27] designed for localizing more subtle anomalies. Nonetheless, our method trained with the proposed attention-guided synthetic anomaly notably outperforms the second-best method, FPI [37], ($\sim$ 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** [40], pseudo label (**PL**) [16], 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

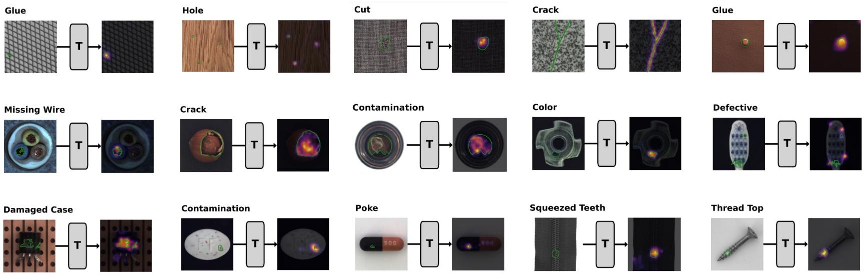


Figure 3: **Anomaly localization examples** from our method before test-time adaptation. The results are superimposed on the test images on the MVTec AD dataset. The green boundary denotes the ground-truth anomalies.

Table 3: **Ablation studies** on the TTA and augmentation strategies on the MVTec AD.

Tasks	Ours w TENT	Ours w PL	Ours w/o TTA	Ours w/o multi-crop	Aug-CutPaste	Aug-NSA	Aug-FPA	Ours
	Ablations for TTA			Ablations for Augmentation Strategy				
Image-Level AUROC (in %)	97.7	97.5	97.2	96.4	95.3	97.3	90.2	**98.4**
Pixel-Level AUROC (in %)	97.6	97.4	97.2	97.2	96.1	96.4	92.1	**98.2**

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 [40, 47], 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

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[2] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952, 2021.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE Computer Society, 2009.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[11] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018.

[12] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022.

[13] Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Robustifying vision transformer without retraining from scratch using attention based test-time adaptation. In *The 36th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 3S3IS2e04–3S3IS2e04, 2022.

[14] Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. Robustifying vision transformer without retraining from scratch by test-time class-conditional feature alignment. *arXiv preprint arXiv:2206.13951*, 2022.

[15] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[16] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[17] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *arXiv preprint arXiv:2206.04325*, 2022.

[18] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[19] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *International Conference on Learning Representations (ICLR)*, 2022.

[20] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[21] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2021.

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[23] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, pages 4800–4809. PMLR, 2019.

[24] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The Internetional Conference on Machine Learning*, 2022.

[25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[27] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[28] Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. *arXiv preprint arXiv:2104.13897*, 2021.

[29] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[31] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[32] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916, 2021.

[33] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[34] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[35] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Self-supervised out-of-distribution detection and localization with natural synthetic anomalies (nsa). *arXiv preprint arXiv:2109.15222*, 2021.

[36] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[37] Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. Detecting outliers with foreign patch interpolation. *arXiv preprint arXiv:2011.04197*, 2020.

[38] Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and Bernhard Kainz. Detecting outliers with poisson image interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 581–591. Springer, 2021.

[39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[41] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[42] Tao Yang, Shenglong Zhou, Yuwang Wang, Yan Lu, and Nanning Zheng. Test-time batch normalization. *arXiv preprint arXiv:2205.10210*, 2022.

[43] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[44] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

[45] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

[46] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021.

[47] Marvin Mengxin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[48] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[49] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, et al. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 2022.