Attention-Conditioned Augmentations for Self-Supervised Anomaly Detection and Localization

Behzad Bozorgtabar[1,2], **Dwarikanath Mahapatra**[3]

[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
behzad.bozorgtabar@epfl.ch, dwarikanath.mahapatra@inceptioniai.org

Abstract

Self-supervised anomaly detection and localization are critical to real-world scenarios in which collecting anomalous samples and pixel-wise labeling is tedious or infeasible, even worse when a wide variety of unseen anomalies could surface at test time. Our approach involves a pretext task in the context of masked image modeling, where the goal is to impose agreement between cluster assignments obtained from the representation of an image view containing saliency-aware masked patches and the uncorrupted image view. We harness the self-attention map extracted from the transformer to mask non-salient image patches without destroying the crucial structure associated with the foreground object. Subsequently, the pre-trained model is fine-tuned to detect and localize simulated anomalies generated under the guidance of the transformer's self-attention map. We conducted extensive validation and ablations on the benchmark of industrial images and achieved superior performance against competing methods. We also show the adaptability of our method to the medical images of the chest X-rays benchmark.

Introduction

Anomaly detection and localization are the centerpieces of many safety-critical applications, ranging from defect detection in industrial pipelines (Bergmann et al. 2019) to medical image inspection (Bozorgtabar et al. 2020; Spahr, Bozorgtabar, and Thiran 2021). Since it is challenging to gather anomalous examples for training, most deep learning-based anomaly detection and segmentation methods (Ruff et al. 2018; Baur et al. 2021; Akcay, Atapour-Abarghouei, and Breckon 2018) are formulated to explore the general patterns of normal samples using one-class classification setup.

Current state-of-the-art (SOTA) anomaly detection and localization methods (Roth et al. 2022; Lee, Lee, and Song 2022; Yu et al. 2021) often rely on deep representations from ImageNet classification. Instead, we focus on the self-supervised setting where we define a pretext task using a multi-view consistency strategy from a training set of only anomaly-free samples. Recently, self-supervised *vision transformers* (ViTs), e.g., (Caron et al. 2021) have performed tremendously on a downstream classification task, but they have not yet been fully explored for anomaly detection task. Nonetheless, these methods focus on learning image-level invariant representation, neglecting local context, making them suboptimal for the anomaly localization task. In particular, we are intrigued by the self-attention mechanisms of ViT-based models (Touvron et al. 2021; Dosovitskiy et al. 2021) and propose a pre-training strategy based on masked image modeling (MIM) (Bao, Dong, and Wei 2021; Zhou et al. 2022). Furthermore, we adopt the fine-tuning scheme using the transformer's attention-conditioned anomaly simulation (Fig. 1 (d)).

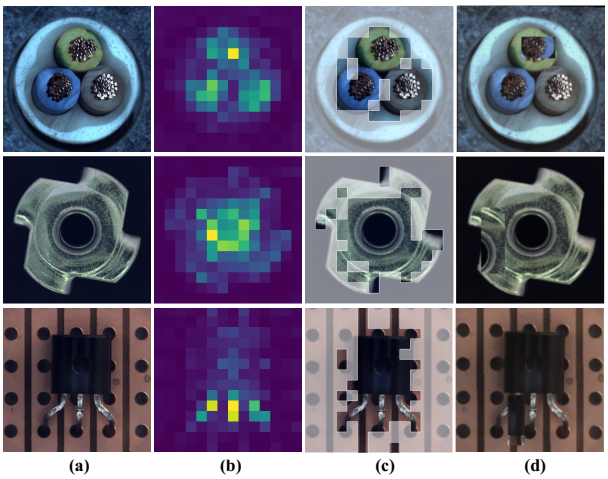


Figure 1: Given (a) the input image, we obtain (b) the average attention map of the transformer's heads, which is then leveraged for (c) attention-conditioned patch masking (APMask) by dropping the least attended patches and (d) attention-conditioned anomaly simulation by cutting and pasting local patch within the salient region.

In summary, we make the following contributions: 1) We propose a MIM strategy, namely **A**ttention-conditioned **P**atch **Mask**ing (**APMask**) (Fig. 1 (c)) for ViT-based pre-training by masking non-salient image patches that help the model to capture the local semantics while preserving the crucial structure associated with the foreground object. We conducted ablations to validate the effectiveness of our masking strategy. 2) We adopt the fine-tuning scheme using

anomaly simulation obtained from normal training samples. To further boost the performance, we utilize a simple test-time augmentation to produce scale-transformed versions of test images. 3) We empirically demonstrate on the challenging datasets of the MVTec AD (Bergmann et al. 2019) and the NIH chest X-rays (Wang et al. 2017) that our approach surpasses existing state-of-the-art self-supervised methods that rely on synthetic anomalies or even methods transferring pre-trained features from ImageNet.

Related Work

Deep Learning Methods for Anomaly Detection and Localization. As it is difficult to anticipate the expected anomaly types at test time, deep learning methods for anomaly detection and localization are typically trained only on normal data. They can be categorized into feature-based methods (Bergman and Hoshen 2020; Ruff et al. 2018; Yi and Yoon 2020), generative adversarial networks (GANs) based methods (Akcay, Atapour-Abarghouei, and Breckon 2018; Schlegl et al. 2017), normalizing flows based approaches (Rudolph, Wandt, and Rosenhahn 2021; Gudovskiy, Ishizaka, and Kozuka 2022; Yu et al. 2021), and reconstruction-based methods (Chen et al. 2022; Bozorgtabar et al. 2020; Nguyen et al. 2019; Baur et al. 2021).

Another line of research proposes to employ self-supervised pre-training schemes (Chen et al. 2020b; He et al. 2020; Caron et al. 2021, 2020). Their objective is based on image-level invariance-based representation (high-level abstract features) under different image transformations. Nonetheless, it has been shown that models pre-trained on ImageNet (Roth et al. 2022; Lee, Lee, and Song 2022; Yu et al. 2021; Schirrmeister et al. 2020) transcend current self-supervised pre-trained models (Schlüter et al. 2022; Golan and El-Yaniv 2018) on relatively small application-specific datasets. Furthermore, the inherent downside of image-level invariance-based representation approaches is that they often neglect local context, which can be beneficial for anomaly localization. More recently, self-supervised methods (Tan et al. 2020; Li et al. 2021a; Schlüter et al. 2022) have been proposed for creating synthetic anomalies through randomly pasting image patches within a single image (Li et al. 2021a) or blending patches from separate images (Schlüter et al. 2022). (Bozorgtabar, Mahapatra, and Thiran 2022) leveraged the transformer's attention map to guide sampling of patches for creating anomalies.

Masked Image Modeling (MIM). Inspired by the success of masked language modeling (MLM) (Devlin et al. 2019) for pre-training language models, recently, this strategy has been extended to the vision domain in the form of masked image modeling (MIM) (Bao, Dong, and Wei 2021; Zhou et al. 2022) for self-supervised learning methods. In the MIM setup, a portion of the input image is randomly masked, and the goal is to predict the missing image region based on its context. Recent approaches incorporate MIM for pre-training ViTs based on image embeddings (Caron et al. 2021; Assran et al. 2022) or encoder-decoder architectures (He et al. 2022; Chen et al. 2020a; Li et al. 2021b; Bao, Dong, and Wei 2021; El-Nouby et al. 2021), in which missing masked input values. e.g., image patches are predicted by reconstructing at pixel level (He et al. 2022) or utilizing an online tokenizer using patch-level loss (Zhou et al. 2022). Contrary to these approaches, our pre-training does not reconstruct the missing patches but instead imposes consistency between the *global* representation of the masked image view and the uncorrupted image view via clustering. Similar to our pre-training scheme, the recent mask denoising method (Assran et al. 2022) performs the denoising implicitly at the embedding space. Nevertheless, *random masking*, e.g., masking the patches of the salient region, deteriorates the representation for self-supervised vision transformers. Unlike random masking, we exploit the self-attention map extracted from the ViT's class token to mask the least attended patches without destroying the crucial structure, yielding a model to capture local semantics.

Method

Attention-Conditioned Patch Masking

We first describe the proposed self-supervised pre-training scheme, which combines a multi-view consistency strategy with attention-guided patch masking; see Fig. 2 for a schematic. We propose a distillation approach based on teacher and student networks. Given the training set of anomaly-free images $\mathcal{D}_u = (\mathbf{x}_i)_{i=1}^{M}$, we sample a mini-batch of B normal training images in each iteration. For each input image $\mathbf{x}$ in a mini-batch, we first apply two random data augmentations, yielding two augmented images. Then we patchify each augmented image by converting it into a sequence of non-overlapped $s \times s$ patches, yielding the patchified sequence for teacher network denoted as *teacher view* $\mathbf{x}_t$, and $V \geq 1$ patchified sequences for student network denoted as *student views* $\mathbf{x}_{s_j}, j \in \{1, \cdots, V\}$. The input augmented image to the student network is randomly cropped V-times into one large crop (global view) and V-1 small crops (local views) to generate student views. Subsequently, for the student view, we additionally apply the attention-conditioned masking strategy by dropping some patches that correspond to the least salient regions of an image to obtain the sequence of masked student patches denoted by $\hat{\mathbf{x}}_{s_j}$, while the target view remains unchanged.

Next, we forward the patchified views through the student and teacher network. Both networks have the same architecture, and the teacher parameters are updated by the exponential moving average of the student parameters following (Caron et al. 2021). The output representation associated with the class token (`[CLS]`) of the ViT encoder f_ϕ, parameterized by ϕ is fed to *multi-layer perceptron* (MLP) projection head g, parameterized by ω with a l_2-normalization bottleneck to obtain the feature representations for the teacher and masked student views. Let $\mathbf{z}_t \in \mathbb{R}^d$ denote the representation computed from the patchified *teacher view* $\mathbf{x}_t$, where d is the embedding dimension. Similarly, let $\hat{\mathbf{z}}_{s_j} \in \mathbb{R}^d$ denote the representation obtained from the patchified (and masked) *student view* $\hat{\mathbf{x}}_{s_j}$. Then, the objective is to encourage the networks to output consistent predictions for different image views. To enforce such behavior, we follow the online clustering of (Caron et al. 2020) that yields matching

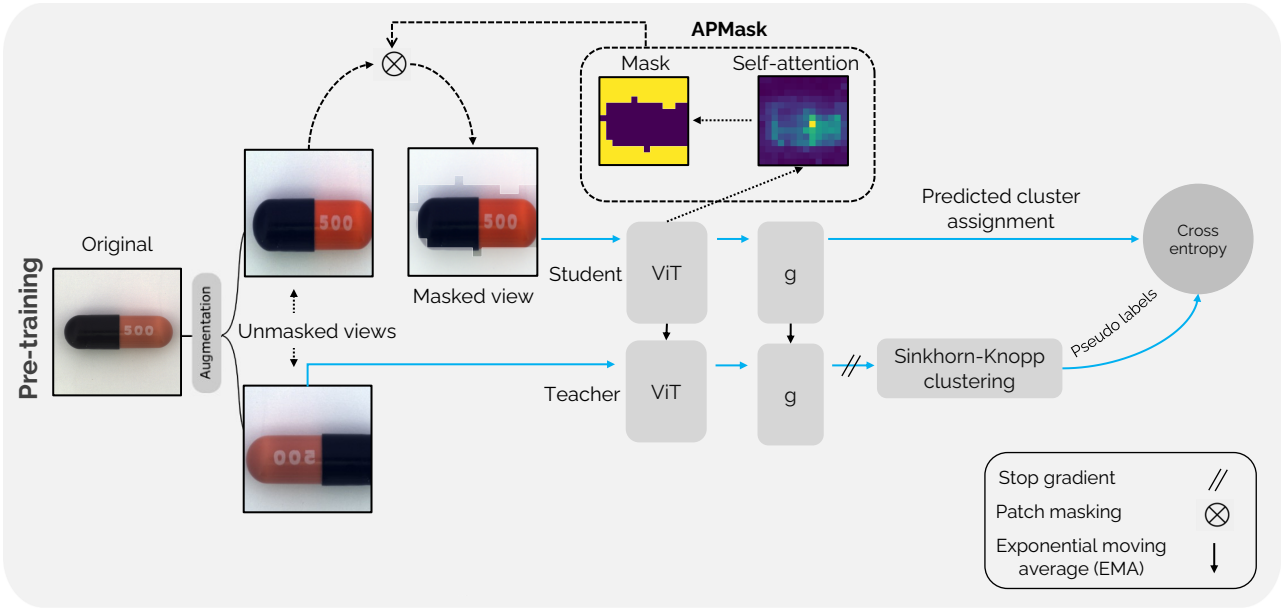


Figure 2: Overview of the proposed MIM pre-training. We set up pre-training from a DINO (Caron et al. 2021) initialization. Given two randomly augmented views, we patchify each augmented view, yielding two sequences of teacher and student views. We feed image views to the student and teacher networks. Subsequently, we apply attention-conditioned patch masking (APMask) to drop the least salient patches from the student view, yielding a masked student view. The networks are optimized using a multi-view consistency objective to enforce the cluster assignments made on the masked student view to match the pseudo labels predicted using the unmasked teacher view. For simplicity, we only illustrate the student's global view.

a soft distribution over a set of prototypes between multiple views of the same image. However, we adopt a multi-view consistency objective to enforce the cluster assignments made on each masked *student view* $\hat{\mathbf{x}}_{s_j}$ to match the optimal cluster assignments (pseudo labels) predicted using unmasked *teacher view* $\mathbf{x}_t$. Taking *teacher view* $\mathbf{x}_t$ as an example, we compute the pseudo labels (soft distribution over clusters) y_t of its features $\mathbf{z}_t \in \mathbb{R}^d$ to K learnable prototype vectors $[\mathbf{c}_1, \cdots, \mathbf{c}_K] = \mathbf{C} \in \mathbb{R}^{d \times K}$. We formulate this problem as an Optimal Transport solver (Asano, Rupprecht, and Vedaldi 2019) via the `Sinkhorn-Knopp` algorithm (Cuturi 2013) optimized over B features within a mini-batch. This solver enforces the uniform partition of the pseudo-labels over all clusters, avoiding degenerate solutions. We measure the multi-view consistency with the cross-entropy loss between the computed pseudo labels y_t of the *teacher view* and the softmaxed cluster assignment predictions p_{s_j} from masked *student view* $\hat{\mathbf{x}}_{s_j}$:

$$\mathcal{L}_{\text{multi-view}} = \frac{1}{|V|} \sum_{j=1}^{V} H\left(y_t, p_{s_j}\right), \qquad (1)$$

$$p_{s_j} = \sigma_\tau \left(\hat{\mathbf{z}}_{s_j}^\top \mathbf{C}\right)$$

where $H(a, b) = -a \log b$ and σ_τ denotes a `softmax` of the dot products of the masked *student view* features $\hat{\mathbf{z}}_{s_j}$ and all prototypes $\mathbf{C}$, which is scaled by temperature $\tau \in (0, 1)$. After each gradient step, the prototypes $\mathbf{C}$ are l_2- normalized. We optimize the loss in Eq. 1 by averaging across all V *student views* and B mini-batch features.

Attention-Conditioned Patch Masking

The transformer (Touvron et al. 2021; Dosovitskiy et al. 2021) processes input image $\mathbf{x} \in \mathbb{R}^{h \times w \times 3}$ by converting and embedding it to the sequence of length N patch tokens $\mathbf{x}_{\texttt{patches}} \in \mathbb{R}^{N \times d}$. A learnable `[CLS]` token $\mathbf{x}_{\texttt{[CLS]}} \in \mathbb{R}^d$ is then prepended to the sequence of patch tokens to form patch embedding $\mathbf{z} \in \mathbb{R}^{(N+1) \times d}$. In addition, to maintain positional information, a sequence of position embeddings is added to $\mathbf{z}$.

Consider a standard self-attention module with L heads, the $(N+1) \times (N+1)$ *attention matrix* $\mathbf{A}_j$ for each head $j \in [L]$ using a row-wise `softmax` is defined as:

$$\mathbf{A}_j = \text{softmax}\left(\mathbf{Q}_j \mathbf{K}_j^\top / \sqrt{d'}\right) \qquad (2)$$

where $\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j \in \mathbb{R}^{N+1 \times d'}$ denote *query*, *key*, and *value* sequences, respectively, and $d' = d/L$. We average matrices over all L heads to obtain the mean attention map $\bar{\mathbf{A}}$. We consider the attended patches by the `[CLS]` token denoted by $\bar{\mathbf{A}}^{\texttt{[CLS]}} \in [0, 1]^N$ as the first row of $\bar{\mathbf{A}}$, which is then reshaped to $(h/s) \times (w/s)$ attention map for a patch size of $s \times s$ pixels.

APMask Generation. To generate attention-conditioned mask, we sort the elements of $\bar{\mathbf{A}}^{\texttt{[CLS]}}$ in ascending order using a *permutation* $\pi\uparrow$: $\{1, \cdots, N\} \rightarrow \{1, \cdots, N\}$, such

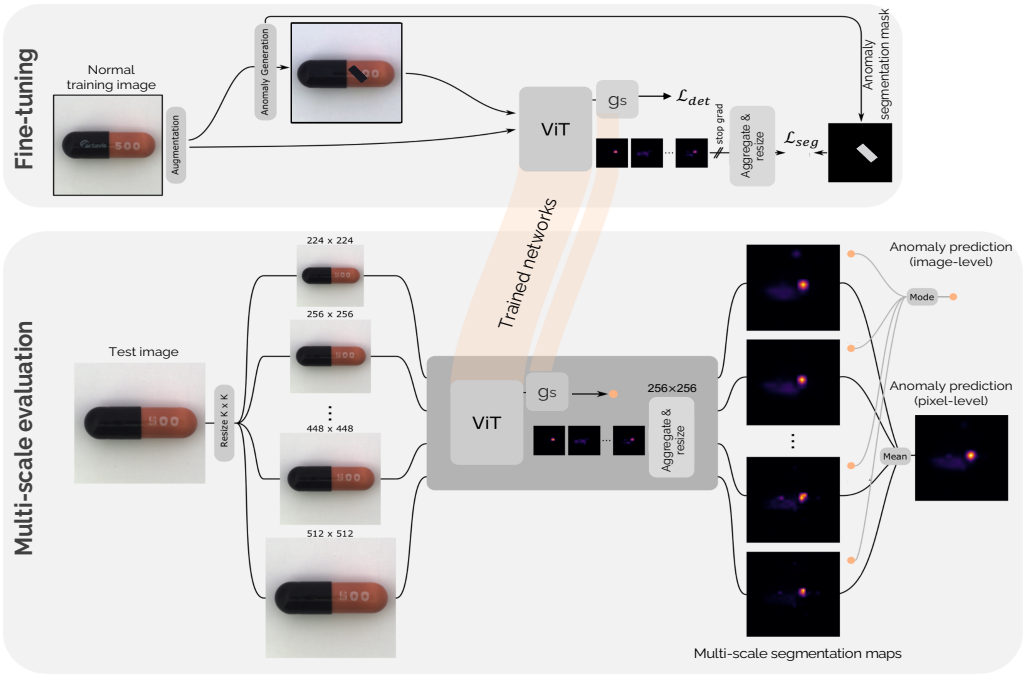


Figure 3: Schematic diagram of the proposed fine-tuning and multi-scale evaluation scheme. Top: The architecture is fine-tuned using normal training images and created synthetic anomalies with their corresponding segmentation masks. Bottom: We process a test image at multiple scales to generate multi-scale segmentation maps and image-level anomaly prediction scores.

that $\bar{A}^{[\mathtt{CLS}]}_{\pi\uparrow(i)} \leq \bar{A}^{[\mathtt{CLS}]}_{\pi\uparrow(j)}$ for $i < j$, where $\bar{A}^{[\mathtt{CLS}]}_{i}$ is the i^{th} element of $\bar{\mathbf{A}}^{[\mathtt{CLS}]}$.

Subsequently, we select a set of indices M^{Att} of the bottommost k elements (tokens with the lowest responses) that correspond to the *masking ratio* $r \in [0, 1]$, such that $k = \lfloor rN \rfloor$:

$$M^{Att} := \{\pi\uparrow(i), \cdots, \pi\uparrow(k)\} \tag{3}$$

We then obtain a binary attention-conditioned patch mask $\mathbf{m}^{Att}$, namely APMask with elements:

$$m_i^{Att} := 1_{M_i^{Att}}(i) = \begin{cases} 1 & \text{if } i \in M^{Att} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

for $i = 1, \cdots, N$. The proposed masking strategy exploits the non-uniform relevance of image patches by leveraging the attention map from `[CLS]` token of the transformer to drop the least attended patches.

Fine-Tuning Using Synthetic Anomalies

In the absence of prior knowledge of the types of anomalous features during training, we adopt synthetic anomalies (Bozorgtabar, Mahapatra, and Thiran 2022) and formulate a proxy task to detect and localize simulated anomalies for model fine-tuning (Fig. 3, top).

The central tenet of utilized synthetic anomaly is that most anomalous regions are associated with the salient objects within the image. Similarly, we benefit from the self-attention maps of a pre-trained ViT to delineate salient image regions used for creating synthetic anomalies (Fig. 4 (b)). More precisely, we use the softmax distribution of the $\bar{\mathbf{A}}^{[\mathtt{CLS}]}$ from the normal training image, which is resized to the input image dimensions followed by re-normalization to guide sampling of locations to cut and paste patches within the same image. The sizes for the patches' width r_w and height r_h are sampled from a uniform distribution $\sim \mathcal{U}(0.1W, 0.4W)$ for the image size of $W \times W$. Subsequently, we apply random rotation, resizing and jitter pixel values for the patches, yielding more diverse synthetic anomalies. Finally, unlike (Bozorgtabar, Mahapatra, and Thiran 2022), for the chest X-rays, we blend the pasted patches using Poisson blending (Tan et al. 2021; Pérez, Gangnet, and Blake 2003). This yields creating more close approximation of natural anomalies with fewer artificial discontinuities.

The fine-tuning stage follows the pre-training step to detect and localize simulated anomalies. The fine-tuning is formulated to simultaneously predict the image-level class $\mathbf{y}_c$ (normal/abnormal) and the segmentation mask $\mathbf{Y}_s$ corresponding to the anomalous regions for an input image. As the ground truth, we set the label $\mathbf{y}_{ci}$ to zero for the normal training image and to one for the simulated anomaly created from a normal training image. Furthermore, we utilize pasted patches' locations of the synthetic anomalies as the ground truth segmentation masks $\mathbf{Y}_s$.

Our model architecture consists of a small ViT encoder f_ϕ and MLP projection head g_s for image-level classification. The projection head g_s takes the last `[CLS]` token of the pre-trained ViT encoder and outputs two neurons. The projection heads g (from pre-training) and g_s share the same architectures except for the last layer. Similar to (Bo-

zorgtabar, Mahapatra, and Thiran 2022), we formulate the image-level anomaly detection training objective $\mathcal{L}_{\text{det}}$ using binary cross-entropy loss to distinguish between normal training image and synthetic anomaly. In practice, we apply data augmentations to the normal training image before creating synthetic anomalies. On the other hand, for anomaly segmentation, we aggregate the attention maps of the ViT's last layer using the learned weights to obtain a single anomaly segmentation prediction $\hat{\mathbf{Y}}_s$. The weights of attention maps are learned using Dice loss for anomaly segmentation training objective $\mathcal{L}_{\text{seg}}$. The total cost function for fine-tuning is defined as follows:

$$\mathcal{L}_{\text{fine-tune}} = \lambda_{det}\mathcal{L}_{\text{det}} + \lambda_{seg}\mathcal{L}_{\text{seg}} \tag{5}$$

where λ_{det} and λ_{seg} are hyperparameters for losses.

Multi-Scale Evaluation. The resolution of 2D attention maps is equal to $(h/s) \times (w/s)$, which is restricted by the patch size of $s \times s$ pixels. While high-resolution attention maps capture detailed information, we cannot expect the localization of fine-grained local anomalies with relatively low-resolution attention maps. To cope with this issue, we utilize a simple *test-time augmentation* to produce *scale* transformed versions per test image. In particular, we process a test image $\mathbf{x}_{\text{test}}$ at T different scales using scale augmentation $\texttt{aug}(\cdot)$. We feed the multi-resolution images to the model parameterized by Φ and obtain T anomaly segmentation maps and T image-level anomaly prediction scores. The final pixel-level anomaly prediction $\hat{\mathbf{Y}}_s$ is computed by *averaging* the T segmentation maps, which are resized to the original test image resolution (Fig. 3, bottom). To compute the final image-level anomaly prediction $\hat{\mathbf{y}}_c$, we check consistency among predictions of T scale transformed versions $\{\Phi(\texttt{aug}_t(\mathbf{x}_{\text{test}}))\}_{t=1}^{T}$. We use a *majority voting* scheme to return prediction for a majority of augmented test images as the final output.

Experiments

Training Setup and Evaluation Metrics. We use PyTorch 1.9 (Paszke et al. 2019) 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 (Caron et al. 2021). We use an MLP-based (Caron et al. 2021) projection head g with output dimension 256, fed to the l_2-bottleneck and the prototypes $\mathbf{C}$ to obtain cluster assignments. Our model is optimized by AdamW (Loshchilov and Hutter 2018) during pre-training for 300 epochs and is fine-tuned for 100 epochs with a batch size of 16. The learning rate linearly increases from $1e$-5 to $5e$-4 over the first 10 epochs and then follows a cosine annealing profile (Loshchilov and Hutter 2016). We follow the data augmentations of (Caron et al. 2021). Then for each student view, except the global view (224×224 pixels), eight local views (96×96 pixels) are generated by random region cropping followed by resizing. We find that the best performance is achieved with the masking ratio $r = 0.3$ (Eq. 3), temperature $\tau = 0.1$, and 20 prototypes. We set $\lambda_{det} = 0.95$ and $\lambda_{seg} = 0.05$ (Eq. 5). For the multi-scale evaluation, we use six different scales ($T = 6$) for the test image $W \in \{224, 320, 384, 416, 448, 512\}$, of the image size of $W \times W$ pixels. 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.

Datasets and Experimental Results

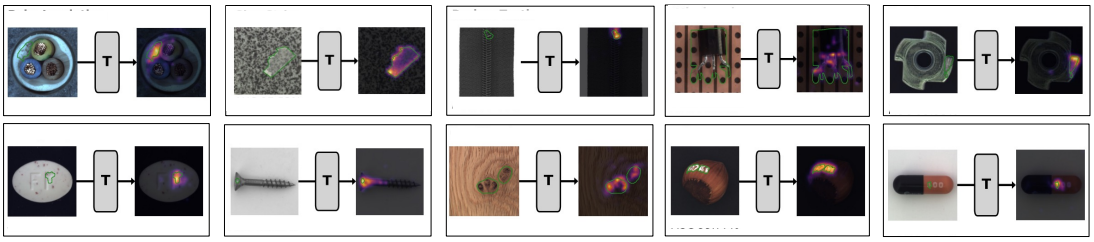
MVTec AD Dataset (Bergmann et al. 2019) comprises 10 object categories and 5 texture categories of industrial images. This dataset contains 3,629 defect-free training images and 1,725 test images, including pixel-level annotations for all defect types. Each category includes high-resolution images ($700 \times 700 \sim 1024 \times 1024$ pixels).

Table 1 compares our method to the SOTA methods, including CutPaste (Li et al. 2021a), NSA (Schlüter et al. 2022), DRAEM (Zavrtanik, Kristan, and Skočaj 2021a), PaDiM (Defard et al. 2021), DifferNet (Rudolph, Wandt, and Rosenhahn 2021), InTra (Pirnay and Chai 2021), SPADE (Cohen and Hoshen 2020), and RIAD (Zavrtanik, Kristan, and Skočaj 2021b). Our method consistently outperforms existing methods and achieves the highest average AUROC (**98.3%** image-level AUROC and **98.2%** pixel-wise AUROC) on the MVTec AD dataset. Our method surpasses existing self-supervised methods relying on synthetic anomalies at random image locations (Li et al. 2021a; Schlüter et al. 2022) or even methods transferring pre-trained features from ImageNet (Zavrtanik, Kristan, and Skočaj 2021a; Defard et al. 2021; Rudolph, Wandt, and Rosenhahn 2021). Our proposed attention-conditioned MIM strategy helps the transformer to better capture the *local semantics*. In addition to quantitative results, the qualitative anomaly localization results (Fig. 4 (a)) on the MVTec AD test set show the precise localization of anomalous regions of varying structures. Furthermore, from the t-SNE (Van der Maaten and Hinton 2008) projection results (see Fig. 5) on example categories of MVTec AD, it can be clearly seen that our trained method using only normal training samples tends to group normal features while separating feature distribution of normal class from defect types.

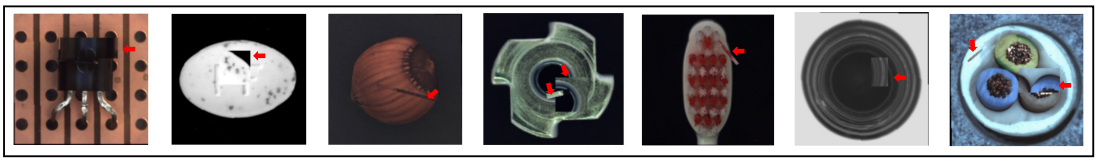
NIH Dataset (Wang et al. 2017) is comprised of 112,120 X-ray images (1024×1024 pixels) annotated either as normal or with 14 categories of thoracic disease labels from 30,805 unique patients. The training set consists of 50,500 normal X-ray images, and the test set includes 25,595 X-rays (15,735 normal and 9,860 abnormal images). In addition, the ground truth anomalous regions in the form of bounding boxes for 880 X-ray images (503 for male and 377 for female patients) are provided. In Table 2, we present the anomaly localization results of the model *fine-tuning* experiment using a ViT encoder initialized from DINO weights (Caron et al. 2021) without our pre-training scheme. Compared to industrial images, anomaly localization on chest X-rays is more challenging due to high inter-sample variability in the nominal data, mistakenly localizing normal variations as an anomaly. Competitive self-supervised methods, FPI (Tan et al. 2020), and PII (Tan et al. 2021) per-

	Carpet	*Grid*	*Leather*	*Tile*	*Wood*	*Bottle*	*Cable*	*Capsule*	*Hazelnut*	*Metal Nut*	*Pill*	*Screw*	*Toothbrush*	*Transistor*	*Zipper*	**Overall Average**
	Textures					**Objects**										
Image-Level Anomaly Classification AUROC (in %)																
CutPaste (3-way)	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
NSA	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	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
PaDiM	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	97.9
DifferNet	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
InTra	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	100.0	99.9	99.8	99.8	96.2	99.1	95.8	97.2	99.6	99.9	98.1	96.4	98.4	95.5	99.7	**98.3**
Pixel-Level Anomaly Localization AUROC (in %)																
CutPaste (3-way)	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
NSA	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	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
PaDiM	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
SPADE	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
RIAD	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
Ours	99.0	98.6	99.3	97.3	96.9	97.8	98.2	98.8	99.1	98.6	99.0	99.4	98.0	94.3	98.9	**98.2**

Table 1: Performance comparison of our method with various methods for anomaly detection (image-level AUROC %) and anomaly localization (pixel-level AUROC %) on the MVTec AD dataset. The best average results are in bold.



(a) Anomaly localization examples



(b) Examples of synthetic anomalies

Figure 4: (a) Anomaly localization results of our method on the test images on the MVTec AD. The green boundary denotes the ground truth. (b) Examples of the synthetic anomalies on the MVTec AD dataset. Red arrows highlight pasted patches.

form well on this task by using Poisson image editing (Pérez, Gangnet, and Blake 2003) designed for localizing more subtle anomalies. Nonetheless, incorporating (Pérez, Gangnet, and Blake 2003) into a synthetic anomaly scheme, our fine-tuned method using synthetic anomalies notably outperforms the second-best method PII (Tan et al. 2020) (**+5.4**% pixel-level AUROC) and creates more diverse synthetic anomalies. We argue that the superior performance is a direct effect induced by the proposed synthetic anomaly.

Ablation Studies

Pre-Training. We investigate the impact of the pre-training on the model's performance by using the same fine-tuning and multi-scale evaluation scheme for all methods and only altering the pre-training. Similar to our pre-training, SwAV (Caron et al. 2020) and MSN (Assran et al. 2022) leverage clustering schemes for representation learning. The former utilizes multi-crop training, while the latter applies random masking to a global image view. The ablation results (Table 3) show that our full method outperforms the other model variants pre-trained using a global semantic level (Caron et al. 2021) (**+1.3**% pixel-level AUROC) or *random* masked image modeling (Zhou et al. 2022; Assran et al. 2022), e.g., block-wise masked patches in iBOT (Zhou et al. 2022) (**+1.2**% image-level AUROC). This stems from better capturing local semantics via our proposed APMask used for pre-training. Since our method only requires storing the activation associated with the masked image view, dropping the least attended patches during pre-training significantly reduces the computational and memory requirements; i.e., masking 30% of patches uses $\sim$ 70% of the computation compared to an unmasked model variant.

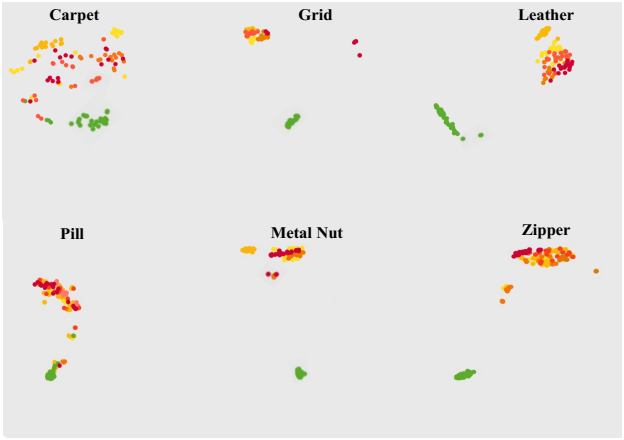


Figure 5: The t-SNE visualization of the learned features (before the projection head) on the MVTec AD categories. The green dots represent features of normal samples, while each anomaly type is shown with a different color.

Method	**Pixel-Level AUROC**%	
	Male	Female
CutPaste	52.6±1.3	51.8±1.2
PaDiM	54.2±0.8	53.8±0.9
FPI	63.4±0.9	62.9±1.1
PII	65.2±0.9	65.1±1.1
Ours	**70.9±0.5**	**70.5±0.8**

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

Fine-Tuning. In addition, we conduct an ablation experiment to assess the impact of the synthetic anomaly for our model variants using the same proposed pre-training but fine-tuned using various synthetic anomaly methods (same fine-tuning setup). The superior ablation results over competitive baselines (Schlüter et al. 2022; Li et al. 2021a) in Table 3 show the importance of sampling patches from informative image regions for creating simulated anomalies. Unlike (Tan et al. 2020; Schlüter et al. 2022; Li et al. 2021a), we leverage the attention map of a transformer to provide an incentive for the model to focus on salient image regions to smoothly blend the patches of various sizes within the same image, yielding more close approximation of real anomalies.

Masking Strategies. In Table 4 (top), we investigate the impact of the masking strategy used for model pre-training on the final performance using the MVTec AD dataset. We pre-train the model with the same training objective using the ViT-S/16 encoder when using a) *No Masking*, b) *Random Masking*, and the proposed c) *APMask*. For all baselines, we use the same masking ratio of 0.3 for a fair comparison. Consequently, all models are fine-tuned using the same fine-tuning setup. The ablation results show that applying *Random Masking* performs better than the *No Masking* baseline. Nonetheless, *APMask* improves upon the *Random Masking* scheme. This is because a random masking strategy may drop relevant tokens of the salient image region, which are necessitous for recognizing anomalies.

Methods	**Pre-training Methods**	
	Pixel-Wise AUROC%	Image-Level AUROC%
DINO	96.9±0.9	97.0±0.7
iBOT	97.2±1.1	97.1±0.9
MSN	97.8±0.6	97.9±0.8
	Augmentation Methods (Fine-tuning)	
	Pixel-Wise AUROC%	Image-Level AUROC%
CutPaste	96.1±1.1	95.4±0.4
NSA	96.4±0.9	97.3±0.5
FPI	92.2±0.7	90.6±0.9
Ours	**98.2**±0.4	**98.3**±0.2

Table 3: Effect of pre-training and fine-tuning scheme. Ablation experiments on the MVTec AD dataset. All methods share the same backbone (ViT-S/16).

Baselines	**Different Masking Strategy**	
	Pixel-Wise AUROC%	Image-Level AUROC%
No Masking	96.2	95.1
Random Masking	97.1	96.6
APMask	**98.2**	**98.3**
	Attention-Conditioned Masking Ratio	
	Pixel-Wise AUROC%	Image-Level AUROC%
$r = 0.15$	97.9	98.0
$r =$ **0.3**	**98.2**	**98.3**
$r = 0.5$	97.8	98.0

Table 4: Effect of different masking strategies and masking ratios. Ablation experiments on the MVTec AD dataset.

Masking Ratio. Since the scale of the foreground object varies across image samples, the choice of masking ratio r needs to be investigated. Without prior knowledge about the size of foreground objects, we thus conduct an ablation study to examine the effect of the *masking ratio* $r \in [0, 1]$ used for the *APMask* on the final performance. Table 4 (bottom) reports the anomaly detection and localization performance for the same model pre-trained with various masking ratios on the MVTec AD dataset. We observe the performance drop using more aggressive masking ratio $r > 0.3$. This observation confirms that by increasing the masking ratio $r > 0.3$, some crucial tokens may be dropped, yielding a decline in performance.

Conclusion

This paper presents a new self-supervised anomaly detection and segmentation method using only normal training samples. The proposed pre-training scheme involves multi-image view consistency learning via patch masking, where we benefit from the attention map of a transformer for patch masking. The model fine-tuned with synthetic anomalies yields better generalization compared to previous methods on two challenging benchmarks. We plan to study the use of auto-encoder architectures in the pixel-wise segmentation of anomalies and to investigate new self-supervised learning approaches to improve further generalizability.

Acknowledgments

The authors are especially grateful to Antoine Spahr for his work in implementing the synthetic anomaly-based model training and also for his plots used for the schematic diagram in the paper.

References

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, 622–637. Springer.

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*.

Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Bordes, F.; Vincent, P.; Joulin, A.; Rabbat, M.; and Ballas, N. 2022. Masked Siamese Networks for Label-Efficient Learning. *arXiv preprint arXiv:2204.07141*.

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Baur, C.; Denner, S.; Wiestler, B.; Navab, N.; and Albarqouni, S. 2021. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Medical Image Analysis*, 69: 101952.

Bergman, L.; and Hoshen, Y. 2020. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Bozorgtabar, B.; Mahapatra, D.; and Thiran, J.-P. 2022. Anomaly detection and localization using attention-guided synthetic anomaly and test-time adaptation. In *British Machine Vision Virtual Conference*.

Bozorgtabar, B.; Mahapatra, D.; Vray, G.; and Thiran, J.-P. 2020. SALAD: Self-supervised aggregation learning for anomaly detection on x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 468–478. Springer.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020a. Generative pretraining from pixels. In *International conference on machine learning*, 1691–1703. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Y.; Tian, Y.; Pang, G.; and Carneiro, G. 2022. Deep one-class classification via interpolated gaussian descriptor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 383–392.

Cohen, N.; and Hoshen, Y. 2020. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, 475–489. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

El-Nouby, A.; Izacard, G.; Touvron, H.; Laptev, I.; Jegou, H.; and Grave, E. 2021. Are Large-scale Datasets Necessary for Self-Supervised Pre-training? *arXiv preprint arXiv:2112.10740*.

Golan, I.; and El-Yaniv, R. 2018. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31.

Gudovskiy, D.; Ishizaka, S.; and Kozuka, K. 2022. 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*, 98–107.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Lee, S.; Lee, S.; and Song, B. C. 2022. CFA: Coupled-hypersphere-based Feature Adaptation for Target-Oriented Anomaly Localization. *arXiv preprint arXiv:2206.04325*.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021a. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674.

Li, Z.; Chen, Z.; Yang, F.; Li, W.; Zhu, Y.; Zhao, C.; Deng, R.; Wu, L.; Zhao, R.; Tang, M.; et al. 2021b. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34: 13165–13176.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Nguyen, D. T.; Lou, Z.; Klar, M.; and Brox, T. 2019. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, 4800–4809. PMLR.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, 313–318.

Pirnay, J.; and Chai, K. 2021. Inpainting transformer for anomaly detection. *arXiv preprint arXiv:2104.13897*.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Rudolph, M.; Wandt, B.; and Rosenhahn, B. 2021. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1907–1916.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International conference on machine learning*, 4393–4402. PMLR.

Schirrmeister, R.; Zhou, Y.; Ball, T.; and Zhang, D. 2020. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33: 21038–21049.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, 146–157. Springer.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, 474–489. Springer.

Spahr, A.; Bozorgtabar, B.; and Thiran, J.-P. 2021. Self-taught semi-supervised anomaly detection on upper limb x-rays. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1632–1636. IEEE.

Tan, J.; Hou, B.; Batten, J.; Qiu, H.; and Kainz, B. 2020. Detecting outliers with foreign patch interpolation. *arXiv preprint arXiv:2011.04197*.

Tan, J.; Hou, B.; Day, T.; Simpson, J.; Rueckert, D.; and Kainz, B. 2021. Detecting outliers with poisson image interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 581–591. Springer.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. 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*, 2097–2106.

Yi, J.; and Yoon, S. 2020. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; and Wu, L. 2021. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021a. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021b. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112: 107706.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*.