AMAE: Adaptation of Pre-trained Masked Autoencoder for Dual-Distribution Anomaly Detection in Chest X-Rays

Behzad Bozorgtabar[1,3](✉), Dwarikanath Mahapatra[2], and Jean-Philippe Thiran[1,3]

[1] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{behzad.bozorgtabar,jean-philippe.thiran}@epfl.ch
[2] Inception Institute of AI (IIAI), Abu Dhabi, United Arab Emirates
dwarikanath.mahapatra@inceptioniai.org
[3] Lausanne University Hospital (CHUV), Lausanne, Switzerland

Abstract. Unsupervised anomaly detection in medical images such as chest radiographs is stepping into the spotlight as it mitigates the scarcity of the labor-intensive and costly expert annotation of anomaly data. However, nearly all existing methods are formulated as a one-class classification trained only on representations from the normal class and discard a potentially significant portion of the unlabeled data. This paper focuses on a more practical setting, dual distribution anomaly detection for chest X-rays, using the entire training data, including both normal and unlabeled images. Inspired by a modern self-supervised vision transformer model trained using partial image inputs to reconstruct missing image regions- we propose AMAE, a two-stage algorithm for adaptation of the pre-trained masked autoencoder (MAE). Starting from MAE initialization, AMAE first creates synthetic anomalies from only normal training images and trains a lightweight classifier on frozen transformer features. Subsequently, we propose an adaptation strategy to leverage unlabeled images containing anomalies. The adaptation scheme is accomplished by assigning pseudo-labels to unlabeled images and using two separate MAE based modules to model the normative and anomalous distributions of pseudo-labeled images. The effectiveness of the proposed adaptation strategy is evaluated with different anomaly ratios in an unlabeled training set. AMAE leads to consistent performance gains over competing self-supervised and dual distribution anomaly detection methods, setting the new state-of-the-art on three public chest X-ray benchmarks - RSNA, NIH-CXR, and VinDr-CXR.

Keywords: Anomaly detection · Chest X-ray · Masked autoencoder

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_19.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 195–205, 2023.
https://doi.org/10.1007/978-3-031-43907-0_19

1 Introduction

To reduce radiologists' reading burden and make the diagnostic process more manageable, especially when the number of experts is scanty, computer-aided diagnosis (CAD) systems, particularly deep learning-based anomaly detection [1,2,22], have witnessed the flourish due to their capability to detect rare anomalies for different imaging modalities including chest X-ray (CXR). Nonetheless, unsupervised anomaly detection methods [20,26] are strongly preferred due to the difficulties of highly class-imbalanced learning and the tedious annotation of anomaly data for developing such systems. Most current anomaly detection methods are formulated as a one-class classification (OCC) problem [18], where the goal is to model the distribution of normal images used for training and thus detect abnormal cases that deviate from normal class at test time. On this basis, image reconstruction based, e.g., autoencoder [9] or generative models [20], self-supervised learning (SSL) based, e.g., contrastive learning [26], and embedding-similarity-based methods [7] have been proposed for anomaly detection. Some recent self-supervised methods proposed synthetic anomalies using cut-and-paste data augmentation [12,19] to approximate real sub-image anomalies. Nonetheless, their performances lag due to the lack of real anomaly data. More importantly, these methods have often ignored readily available unlabeled images. More recently, similar to our method, DDAD [3] leverages readily available unlabeled images for anomaly detection, but it requires training an ensemble of several reconstruction-based networks. Self-supervised model adaptation on *unlabeled data* has been widely investigated using convolutional neural networks (CNNs) in many vision tasks via self-training [17], contrastive learning [22,26], and anatomical visual words [10]. Nonetheless, the adaptation of vision transformer (ViT) [8] architectures largely remains unexplored, particularly for anomaly detection. Recently, masked autoencoder (MAE) [11] based models demonstrated great scalability and substantially improved several self-supervised learning benchmarks [27].

In this paper, inspired by the success of the MAE approach, we propose a two-stage algorithm for "**A**daptation of pre-trained **M**asked **A**uto**E**ncoder" (AMAE) to leverage simultaneously normal and unlabeled images for anomaly detection in chest X-rays. As for **Stage 1** of our method, (i) AMAE creates synthetic anomalies from only normal training images, and the usefulness of pre-trained MAE [11] is evaluated by training a lightweight classifier using a proxy task to detect synthetic anomalies. (ii) For the **Stage 2**, AMAE customizes the recipe of MAE adaptation based on an unlabeled training set. In particular, we propose an adaptation strategy based on reconstructing the masked-out input images. The rationale behind the proposed adaptation strategy is to assign pseudo-labels to unlabeled images and train two separate modules to measure the distribution discrepancy between normal and pseudo-labeled abnormal images. (iii) We conduct extensive experiments across three chest X-ray datasets and verify the effectiveness of our adaptation strategy in apprehending anomalous features from unlabeled images. In addition, we evaluate the model with different

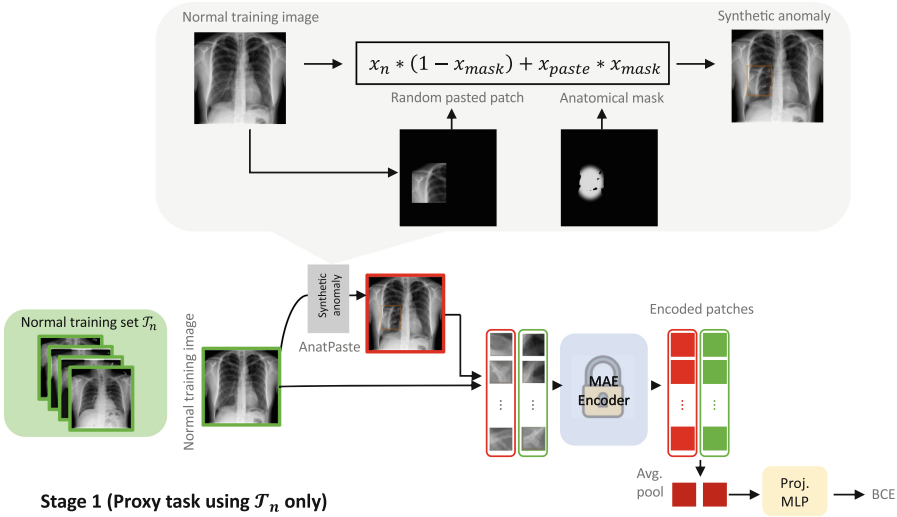

Fig. 1. Schematic overview of AMAE training (Stage 1). Top. Illustration of Anatpaste augmentation [19] generated from normal training images. **Bottom**. Starting from MAE initialization, only the MLP-based projection head (`Proj.`) is trained to classify synthetic anomalies.

anomaly ratios (ARs) in an unlabeled training set and show consistent performance improvement with increasing AR.

2 Method

Notation. We first formally define the problem setting for the proposed dual-distribution anomaly detection. Contrary to previous unsupervised anomaly detection methods, AMAE fully uses unlabeled images, yielding a training data $\mathcal{T}_{train} = \mathcal{T}_n \cup \mathcal{T}_u$ consisting of both normal $\mathcal{T}_n$ and unlabeled $\mathcal{T}_u$ training sets. We denote the normal training set as $\mathcal{T}_n = \{\boldsymbol{x}_{ni}\}_{i=1}^{N}$, with N normal images, and the unlabeled training set as $\mathcal{T}_u = \{\boldsymbol{x}_{ui}\}_{i=1}^{M}$, with M unlabeled images to be composed of both normal and abnormal images. At test time, given a test set $\mathcal{T}_{test} = \{(\boldsymbol{x}_{ti}, y_i)\}_{i=1}^{S}$ with S normal or abnormal images, where $y_i \in \{0, 1\}$ is the corresponding label to $\boldsymbol{x}_{ti}$ (0 for normal (negative) and 1 for abnormal (positive) image), the trained anomaly detection model should identify whether the test image is abnormal or not.

Architecture. Our architecture is ⋊-shaped: the ViT-small (ViT-S/16) [8] encoder f followed by a ViT head g and lightweight (3-layer) multilayer perception (MLP) based projection head h, simultaneously. Starting from the pre-trained MAE on 0.3M unlabeled chest X-rays and officially released checkpoints, we use exactly the same ViT encoder f and decoder g as MAE [27].

2.1 Stage 1- Proxy Task to Detect Synthetic Anomalies

AMAE starts the first training stage using only normal training images by defining a proxy task to detect synthetic anomalies shorn of real known abnormal images. For this purpose, we utilize the state-of-the-art (SOTA) anatomy-aware cut-and-paste augmentation, AnatPaste [19], to create synthetic anomalies from only a set of normal training images $\mathcal{T}_n$. AnatPaste integrates an anatomical mask $\boldsymbol{x}_{mask}$ created from unsupervised lung region segmentation, which guides generating anomalous images via cutting a patch from a normal chest radiograph $\boldsymbol{x}_n$ and randomly pasting it at another image location $\boldsymbol{x}_{paste}$ as:

$$\mathbf{Aug}\left(\boldsymbol{x}_n\right) = \boldsymbol{x}_n * \left(1 - \boldsymbol{x}_{mask}\right) + \boldsymbol{x}_{paste} * \boldsymbol{x}_{mask} \tag{1}$$

where $\mathbf{Aug}\left(\cdot\right)$ is the AnatPaste augmentation (see [19] for more details).

Given a normal training set $\mathcal{T}_n$, for each normal image $\boldsymbol{x}_n \sim \mathcal{T}_n$, we create a synthetic anomaly, denoted as $\mathbf{Aug}\left(\boldsymbol{x}_i\right)$. In preparation for input to the frozen ViT encoder, f_0 (obtained by MAE pre-training), each input image with the $h \times w$ spatial resolution is split into $T = (h/p) \times (w/p)$ patches of size $p \times p$. Then, for every input patch, a token is created by linear projection with an added positional embedding. The sequence of tokens is then fed to the frozen ViT encoder f_0 consisting stack of transformer blocks, yielding the embeddings of tokens $\boldsymbol{z}_i^1, \boldsymbol{z}_i^2 \in \mathbb{R}^{T \times d}$ corresponding to i^{th} normal and synthetic anomaly images. The returned embeddings $\boldsymbol{z}_i^1, \boldsymbol{z}_i^2 \in \mathbb{R}^{T \times d}$ are pooled via *average pooling* to form d-dimensional embeddings, which are fed to an MLP anomaly classifier projection head h (see Fig. 1 for schematic overview). Subsequently, we only train an anomaly classifier projection head h on top of the frozen embeddings to detect synthetic anomalies using the cross-entropy loss l_{ce} as follows:

$$h_0 = \arg\min_h \mathbb{E}_{\boldsymbol{x}_n \sim \mathcal{T}_n}\left[l_{ce}\left(h \circ f_0\left(\boldsymbol{x}_n\right), 0\right) + l_{ce}\left(h \circ f_0\left(\mathbf{Aug}\left(\boldsymbol{x}_n\right)\right), 1\right)\right] \tag{2}$$

We set the label for the normal image to 0 and 1 otherwise (synthetic anomaly). The above gradient-based optimization produces a trained classifier projection head h_0. Thus, the whole architecture can be trained with much fewer parameters while making only the classifier projection head specialized at recognizing anomalies without influencing the ViT encoder.

2.2 Stage 2- MAE Inter-Discrepancy Adaptation

The proposed MAE adaptation scheme is inspired by [3] to model the dual distribution of training data. Unlike [3], which treats all unlabeled images similarly, we propose assigning pseudo-labels to unlabeled images and formulating anomaly detection by measuring the distribution discrepancy between normal and pseudo-labeled abnormal images from unlabeled sets. We use a pre-trained anomaly classifier (**Stage 1**) to assign pseudo labels to unlabeled images. To begin with, for each unlabeled image $\boldsymbol{x}_u \sim \mathcal{T}_u$, we consider the anomaly detection model's confidence from **Stage 1** $(h_0 \circ f_0\left(\boldsymbol{x}_u\right))$. Those images on which the model is highly confident (normal or abnormal) are treated as reliable images

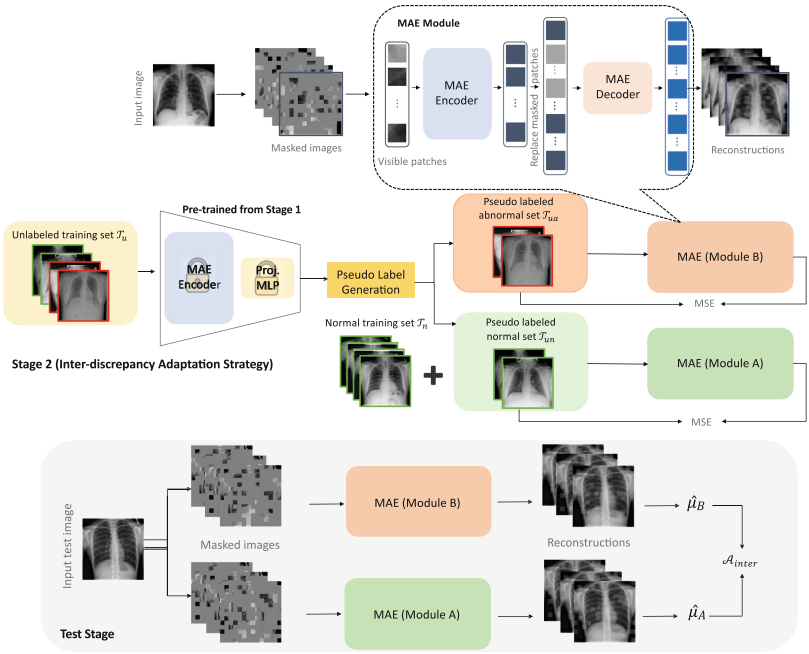

Fig. 2. Schematic overview of AMAE **training (Stage 2) and test stage**. **Top**. Our adaptation strategy first assigns pseudo labels to unlabeled images using a pre-trained classifier from **Stage 1** and uses two separate MAE modules to model the normative and anomalous distributions of pseudo-labeled images. **Bottom**. The discrepancy between the average of multiple reconstructions from two modules is used at test time to compute the anomaly score.

used for adaptation. For this purpose, we collect all output probabilities and opt for a threshold per class t_c corresponding to each class's top K-th percentile ($K = 50$) of all given confidence values. Those unlabeled images deemed reliable yield two subsets: a subset of pseudo-labeled normal images $\mathcal{T}_{un}$ and a subset of pseudo-labeled abnormal images $\mathcal{T}_{ua}$. We then utilize two MAE-based modules, **Module A** and **Module B** (see Fig. 2, **Top**), using the same MAE architecture and pixel-wise mean squared error (MSE) optimization in [11]. Within each module, the input patches for each image are masked out using a set of L random masks $\left\{\boldsymbol{m}^{(j)} \in \{0,1\}^T\right\}_{j=1}^{L}$ in which a different small subset of the patches (ratio of 25%) is retained each time to be fed to the ViT encoder f. The lightweight ViT decoder g receives unmasked patches' embeddings and adds learnable masked tokens to replace the masked-out patches. Subsequently, the full set of embeddings of visible patches and masked tokens with added positional embeddings to all tokens is processed by the ViT decoder g to reconstruct the missing patches of each image in pixels. This yields the reconstructed image $\hat{\boldsymbol{x}}^{(j)} = g \circ f\left(\boldsymbol{m}^{(j)}(\boldsymbol{x})\right)$, which is then compared against the input image $\boldsymbol{x}$ to optimize both ViT encoder

Table 1. Summary of dataset repartitions. Unlabeled image set $\mathcal{T}_u$ is constructed from the images presented in parentheses without using their annotations.

Dataset	Normal training set $\mathcal{T}_n$	**Repartition**	
		Unlabeled training set $\mathcal{T}_u$	Test set $\mathcal{T}_{test}$
RSNA	3851	4000 (4000 normal + 5012 abnormal)	1000 normal + 1000 abnormal images
VinDr-CXR	4000	4000 (5606 normal + 3394 abnormal)	1000 normal + 1000 abnormal images
NIH-CXR	3614	0	543 normal + 262 abnormal images

f and decoder g:

$$f^*, g^* = \arg\min_{f,g} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{T}_{train}} \left[\frac{1}{L} \sum_{j=1}^{L} l_{mse} \left(\boldsymbol{m}^{(j)} \left(\hat{\boldsymbol{x}}^{(j)} \right), \boldsymbol{m}^{(j)} (\boldsymbol{x}) \right) \right] \tag{3}$$

All pixels in the tth patch of both input image and reconstructed images are multiplied by $\left(\boldsymbol{m}^{(j)}\right)_t \in \{0, 1\}$. The above self-supervised loss term averages L pixel-wise mean squared errors for each image. **Module A** is trained on a combination of the normal training set $\mathcal{T}_n$ and pseudo-labeled normal images from the unlabeled set $\mathcal{T}_{un}$. In contrast, **Module B** is trained using only pseudo-labeled abnormal images from an unlabeled set $\mathcal{T}_{ua}$. Optimization for Eq. 3 always starts from pre-trained f_0 and g_0, and we reset the MAE weights to f_0 and g_0 before training each module. A high discrepancy between the reconstruction outputs of the two modules can indicate potential abnormal regions. Similar to the training stage, we apply L random masks to the test image $\boldsymbol{x}_t \sim \mathcal{T}_{test}$ to obtain L reconstructions (see Fig. 2, **Bottom**). Thus, the anomaly score based on the inter-discrepancy of the two MAE modules is computed as follows:

$$\mathcal{A}^p_{inter} = |\hat{\mu}^p_A - \hat{\mu}^p_B| \tag{4}$$

where p is the index of pixels, $\hat{\mu}_A$ and $\hat{\mu}_B$ are the *mean* maps of L reconstructed images from **Module A** and **Module B**, respectively. The pixel-level anomaly scores for each image are averaged, yielding the image-level anomaly score.

3 Experiments

Datasets. We evaluated our method on three public CXR datasets: 1) the RSNA Pneumonia Detection Challenge dataset[1], 2) the VinBigData Chest X-ray Abnormalities Detection Challenge dataset (VinDrCXR)[2] [15], and a subset of 3) the curated NIH dataset (NIH-CXR)[3] [21,25], by including only posteroanterior view images of both male and female patients aged over 18. We show a summary of each dataset's repartitions in Table 1. Except for NIH, where we use only the normal set $\mathcal{T}_n$ (OCC setting), for the other two datasets, we utilize both $\mathcal{T}_n$ and $\mathcal{T}_u$ and exact repartition files from [3] for model training.

[1] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.
[2] https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection.
[3] https://nihcc.app.box.com/v/ChestXray-NIHCC/file/371647823217.

Table 2. Performance comparison with SOTA methods (Avg. over four replicate). The two best results for each protocol are highlighted in **bold** and underlined. Note that "IN" refers to "ImageNet-Pretrained," and "e2e" refers to end-to-end training. The experimental results of competing methods† are taken from [4].

Method	Protocol	Taxonomy	CXR Datasest & Metrics					
			RSNA		VinDr-CXR		NIH-CXR	
			AUC %	AP %	AUC %	AP %	AUC %	AP %
AE†	N	Rec.	66.9	66.1	55.9	60.3	70.0	65.4
AE-U† [14]	N	Rec.	86.7	84.7	73.8	72.8	91.0	83.2
f-AnoGAN† [20]	N	Rec.	79.8	75.6	**76.3**	**74.8**	84.1	79.2
IGD† [6]	N	Rec.	81.2	78.0	59.2	58.7	85.2	80.3
DRAEM† [28]	N	Rec.+SSL	62.3	61.6	63.0	68.3	65.5	62.4
CutPaste^{e2e}† [21]	N	SSL	55.0	58.0	54.6	55.5	58.0	-
AnatPaste [19]	N	SSL	83.1	83.7	66.0	66.2	94.0	93.5
FPI† [23]	N	SSL	47.6	55.7	48.2	49.9	70.5	-
PII† [24]	N	SSL	82.9	83.6	65.9	65.8	92.2	-
NSA† [21]	N	SSL	82.2	82.6	64.4	65.8	94.1	-
DDAD-AE [3]	N	Rec.	69.4	-	60.1	-	73.0	71.5
AMAE^{IN} - Stage 1	N	SSL	83.3	83.8	65.9	66.0	94.1	93.7
AMAE - Stage 1	N	SSL	**86.8**	**84.9**	74.2	72.9	**95.0**	**94.9**
CutPaste^{e2e}† [21]	Y	SSL	59.8	61.7	59.2	60.0	-	
AnatPaste [19]	Y	SSL	84.4	85.5	67.1	67.5	-	
FPI† [23]	Y	SSL	46.6	53.8	47.4	49.4	-	
PII† [24]	Y	SSL	84.3	85.4	66.8	67.2	-	
NSA† [21]	Y	SSL	84.2	84.3	64.4	64.8	-	
DDAD-AE [3] ($\mathcal{A}_{iner}$)	Y	Rec.	81.5	81.0	71.0	-	-	
AMAE - Stage 2 ($\mathcal{A}_{inter}$)	Y	Rec.	**91.4**	**91.7**	**86.1**	**84.5**	-	

Implementation Details. We adopt AdamW [13] optimizer and set the learning rate (lr) and batch size to 2.5e-4 and 16, where we linearly warm up the lr for the first 20 epochs and decay it following a cosine schedule thereafter till 150 epochs. We follow the exact recipe as [27] for other hyperparameters (see **Supplementary Material**). The number of generated masks L per image is set to 2 and 4 for the adaptation and test stages (see ablation in **Supplementary Material**). We use PyTorch 1.9 [16] and train each model on a single GeForce RTX 2080 Ti GPU. We use the area under the ROC curve (AUC) and average precision (AP) for the evaluation metrics.

Comparison with SOTA Methods. Table 2 compares AMAE with a comprehensive collection of SOTA methods, including self-supervised synthetic anomaly and reconstruction (Rec.) based methods using their official codes and under two experimental protocols. We use the **Y** protocol to indicate if access to

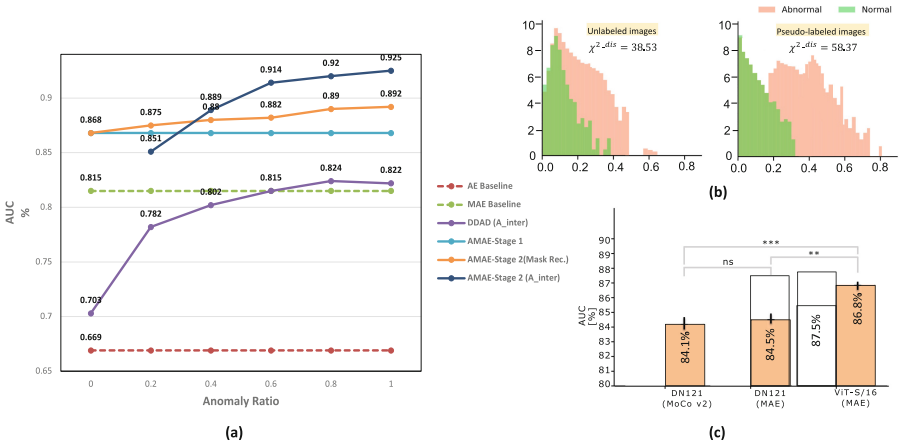

Fig. 3. Ablations on the RSNA dataset. **(a)** Performance comparison with a varying AR of $\mathcal{T}_u$. **(b)** The $\chi^{2-\text{distance}}$ of AS histograms *with* and *without* pseudo labeling. **(c)** Ablation for different backbones and pre-training schemes. Test AUC performances are presented as $mean \pm 1.96 std$ averaged over four replicates. Samples are statistically tested for H_0*: means are similar*, using a bilateral Welch t-test. *ns* non-significant, ** $pvalue \in [0.05, 0.01]$, *** $pvalue \in [0.01, 0.001]$.

the unlabeled images is possible in which an AR of 60% of $\mathcal{T}_u$ is assumed in the experiments; otherwise, we use **N**. Under protocol **N** (OCC setting), except for VinDr-CXR, AMAE-**Stage 1** achieves SOTA results on two CXR benchmarks, demonstrating the effectiveness of pre-trained ViT using MAE and synthetic anomalies. In particular, AMAE-**Stage 1** surpasses the best-performing synthetic anomaly-based method, AnatPaste [19], with the same synthesis approach as ours but using ResNet18 as a feature extractor. Furthermore, outperforming MAE pre-training on ImageNet (e.g., improved AUC from 65.9% to 74.2% on the VinDr-CXR) indicates the importance of in-domain adaptation. Under the **Y** protocol, our adaptation strategy, AMAE- **Stage 2** ($\mathcal{A}_{inter}$), outperforms the current SSL methods by a larger margin, underlining the importance of modeling the dual distribution and leveraging unlabeled images more effectively.

Ablation Studies. To understand the effectiveness of AMAE in capturing abnormal features from unlabeled images, we conduct ablation experiments on the RSNA with the AR of $\mathcal{T}_u$ varying from 0 to 100% (Fig. 3 (**a**)). Concerning reconstruction-based methods, the baseline aggregating multiple reconstructions via MAE achieves consistent improvement compared with the AE baseline (+14.6% AUC), implying better capturing of fine-grained texture information. With an increasing AR of $\mathcal{T}_u$, our MAE adaptation strategy (AMAE - **Stage 2** ($\mathcal{A}_{inter}$)) performs favorably against the SOTA method (DDAD [3]), and AMAE without adaptation (**Stage 1**). Furthermore, we consider an additional baseline for model adaptation based on patch reconstruction in [11] on pooled

normal and unlabeled images, denoted as (AMAE - **Stage 2** (Mask Rec.)). AMAE - **Stage 2** ($\mathcal{A}_{inter}$) rises with a steeper slope than AMAE - **Stage 2** (Mask Rec.), e.g., improved AUC from 89% to 92% on AR=80%, suggesting high-quality pseudo-labeled images. We also analyze the discriminative capability of our adaptation *with* and *without* pseudo labeling by levering all unlabeled images in **Module B**. We utilize the $\chi^{2-\text{distance}}$ between the histograms of anomaly scores (AS) of normal and abnormal images in the RSNA test set (Fig. 3 (**b**)), showing a more substantial discriminative capability of incorporating pseudo labeling (improved $\chi^{2-\text{distance}}$ from 38.53 to 58.37). Finally, the ViT encoder obtained by MAE pre-training (**Stage 1**) surpasses DenseNet-121 (DN121), either pre-trained by MAE [27] or an advanced contrastive learning method (MoCo v2 [5]) on the RSNA dataset (Fig. 3 (**c**)).

4 Conclusion

We present AMAE, an adaptation strategy of the pre-trained MAE for dual distribution anomaly detection in CXRs, which makes our method capable of more effectively apprehending anomalous features from unlabeled images. Experiments on the three CXR benchmarks demonstrate that AMAE is generalizable to different model architectures, achieving SOTA performance. As for the limitation, an adequate number of normal training images is still required, and we will extend our pseudo-labeling scheme in our future work for robust anomaly detection bypassing any training annotations.

References

1. Bozorgtabar, B., Mahapatra, D.: Attention-conditioned augmentations for self-supervised anomaly detection and localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14720–14728 (2023)
2. Bozorgtabar, B., Mahapatra, D., Vray, G., Thiran, J.-P.: SALAD: self-supervised aggregation learning for anomaly detection on X-Rays. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 468–478. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_46
3. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy for anomaly detection in chest x-rays. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III. pp. 584–593. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_56
4. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Med. Image Anal. **86**, 102794 (2023)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Chen, Y., Tian, Y., Pang, G., Carneiro, G.: Deep one-class classification via interpolated gaussian descriptor. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 383–392 (2022)

7. Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDiM: a patch distribution modeling framework for anomaly detection and localization. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12664, pp. 475–489. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68799-1_35
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Gong, D., et al.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1705–1714 (2019)
10. Haghighi, F., Taher, M.R.H., Zhou, Z., Gotway, M.B., Liang, J.: Transferable visual words: exploiting the semantics of anatomical patterns for self-supervised learning. IEEE Trans. Med. Imaging **40**(10), 2857–2868 (2021)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
12. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9664–9674 (2021)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
14. Mao, Y., Xue, F.-F., Wang, R., Zhang, J., Zheng, W.-S., Liu, H.: Abnormality detection in chest X-ray images using uncertainty prediction autoencoders. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 529–538. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_51
15. Nguyen, H.Q., et al.: VinDr-CXR: an open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)
16. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
17. Prabhu, V., Khare, S., Kartik, D., Hoffman, J.: Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8558–8567 (2021)
18. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International Conference on Machine Learning, pp. 4393–4402. PMLR (2018)
19. Sato, J., et al.: Anatomy-aware self-supervised learning for anomaly detection in chest radiographs. arXiv preprint arXiv:2205.04282 (2022)
20. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGan: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019)
21. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: 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. pp. 474–489. Springer (2022). https://doi.org/10.1007/978-3-031-19821-2_27
22. Spahr, A., Bozorgtabar, B., Thiran, J.P.: Self-taught semi-supervised anomaly detection on upper limb x-rays. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1632–1636. IEEE (2021)
23. Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting outliers with foreign patch interpolation. arXiv preprint arXiv:2011.04197 (2020)

24. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 581–591. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_56
25. Tang, Y.X., et al.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digital Med. **3**(1), 70 (2020)
26. Tain, Yu., et al.: Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 128–140. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_13
27. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3588–3600 (2023)
28. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8330–8339 (2021)