One-for-All: Proposal Masked Cross-Class Anomaly Detection

Xincheng Yao[1], Chongyang Zhang[1,2*], Ruoqi Li[1], Jun Sun[1], Zhenyu Liu[3]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China
[3]Ningbo HTVision Digital Technology Co.,Ltd, Ningbo 315000, China
{i-Dover, sunny_zhang, nilponi, junsun}@sjtu.edu.cn[1], lzy409911075@163.com[3]

Abstract

One of the most challenges for anomaly detection (AD) is how to learn one unified and generalizable model to adapt to multi-class especially cross-class settings: the model is trained with normal samples from seen classes with the objective to detect anomalies from both seen and unseen classes. In this work, we propose a novel Proposal Masked Anomaly Detection (PMAD) approach for such challenging multi- and cross-class anomaly detection. The proposed PMAD can be adapted to seen and unseen classes by two key designs: MAE-based patch-level reconstruction and prototype-guided proposal masking. First, motivated by MAE (Masked AutoEncoder), we develop a patch-level reconstruction model rather than the image-level reconstruction adopted in most AD methods for this reason: the masked patches in unseen classes can be reconstructed well by using the visible patches and the adaptive reconstruction capability of MAE. Moreover, we improve MAE by ViT encoder-decoder architecture, combinational masking, and visual tokens as reconstruction objectives to make it more suitable for anomaly detection. Second, we develop a two-stage anomaly detection manner during inference. In the proposal masking stage, the prototype-guided proposal masking module is utilized to generate proposals for suspicious anomalies as much as possible, then masked patches can be generated from the proposal regions. By masking most likely anomalous patches, the "shortcut reconstruction" issue (*i.e.*, anomalous regions can be well reconstructed) can be mostly avoided. In the reconstruction stage, these masked patches are then reconstructed by the trained patch-level reconstruction model to determine if they are anomalies. Extensive experiments show that the proposed PMAD can outperform current state-of-the-art models significantly under the multi- and especially cross-class settings. Code will be publicly available at https://github.com/xcyao00/PMAD.

Introduction

Anomaly detection has widespread applications in diverse domains, such as industrial defect inspection (Bergmann et al. 2019a; Mishra et al. 2021; Defard et al. 2021; Roth et al. 2022; Yao, Zhang, and Li 2022), video surveillance (Acsintoae et al. 2022; Sultani, Chen, and Shah 2018), medical lesion detection (Tian et al. 2021; Zhang et al. 2021), and road anomaly detection (Vojir, Sipka, and Aljundi 2021; Biase et al. 2021). Due to the scarcity of anomalies, most previous anomaly detection studies have mainly devoted to unsupervised learning, *i.e.*, learning normal patterns by only utilizing anomaly-free data and treating anomalies as outliers. The current unsupervised AD methods are comprised of two main trends, *i.e.*, reconstruction-based (Bergmann et al. 2019b; Schlegl et al. 2017) and embedding-based methods (Defard et al. 2021; Roth et al. 2022). The former mainly utilizes AutoEncoders (Hinton and Salakhutdinov 2006) or GANs (Goodfellow et al. 2014) to generate reconstructed image and then employ reconstruction errors between the input and reconstructed image to localize anomalies. The latter aims to learn an embedding neural network for making normal data close to each other in the embedding space (Reiss et al. 2021; Roth et al. 2022; Li et al. 2021).

However, regarding the issue of class adaptability, we

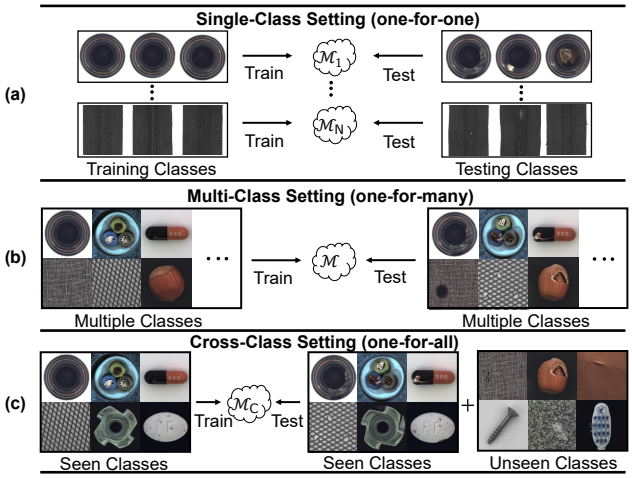

Figure 1: Different anomaly detection settings. (a) Single-Class Setting (one-for-one): most AD methods train a specific model for each class. (b) Multi-Class Setting (one-for-many): one unified model is trained and then used for multiple known classes. (c) Cross-Class Setting (one-for-all): one unified model is trained with normal data from seen classes, and aims to detect anomalies directly without any fine-tuning from both seen and unseen classes.

*Corresponding Author.

observe that previous methods often need to train a specific model for each object class. This one-for-one paradigm would require more computational and memory overhead, and more resources are required to store different model weights in real-world applications. Moreover, new classes usually appear in real-world scenarios, but these trained models cannot generalize directly to the new classes, which may cause the application system to fail in new scenarios. However, maintaining the system by retraining or fine-tuning is cost-ineffective. Thus, existing AD methods are still unsatisfactory for real-world scenarios. Therefore, class adaptability is a critical issue in the AD community, but it's still not been well studied in most AD literatures.

This paper aims to address the two issues mentioned above, we propose and focus on two new but more challenging AD settings: multi-class and cross-class settings. As shown in Figure 1(b) and 1(c), under the multi-class setting, we follow the one-for-many paradigm to train one unified model with normal data from multiple classes, and the objective is to detect anomalies from the same known classes; under the cross-class setting, we follow the one-for-all paradigm to train one unified model with normal data from seen classes, and the objective is to detect anomalies from both seen and unseen classes. In this paper, we consider how to construct a class-adaptive AD model based on the popular reconstruction idea. However, modeling one class-adaptive reconstruction-based anomaly detector has two main challenges: 1. *How to obtain successful reconstruction for unseen classes?* As the model is only trained by normal samples from known classes, it may cause the model to fail when reconstructing samples of unseen classes (*i.e.*, both normal and abnormal regions are poorly reconstructed). 2. *How to effectively mitigate the "identical reconstruction" issue?* The reconstruction model can sometimes be overfitted, this will cause the "identical reconstruction" issue (Perera, Nallapati, and Xiang 2019), where both normal and anomalous regions can be well reconstructed. This will lead to lower anomaly scores in abnormal regions and thus failure of anomaly detection.

To address the class adaptability issue, we propose a novel **P**roposal **M**asked **A**nomaly **D**etection approach (PMAD), which consists of two key designs: MAE-based patch-level reconstruction and prototype-guided proposal masking. In MAE, He, *et al.* show that masked autoencoders are scalable vision learners. In this paper, we find that we can learn a unified and generalizable AD model based on MAE for this reason: The objective in MAE allows the model to learn how to utilize the contextual relationship in the image to infer the features of the masked patches. The model actually learns a contextual inference relationship in a single image, rather than the class-dependent reconstruction mode (*i.e*, generally learned in the conventional image-level reconstruction models). Thus, even in unseen classes, the masked patches can be reconstructed well by using the visible patches and the adaptive reconstruction capability of the model, then the anomalies can be detected by large reconstruction errors. To address the second challenge, we develop a two-stage anomaly detection manner during inference. Specifically, we propose a prototype-guided proposal masking approach to generate masked patches (suspicious anomalies) and then reconstruct these masked patches by the trained reconstruction model to decide if they are anomalies. We can address the second challenge by masking the main anomaly information: As shown in (He et al. 2022), the MAE-based reconstruction model is robust enough to reconstruct the masked patches. Thus, if a large amount of anomaly information is leaked, the model can generate good reconstruction for abnormal patches, causing failure of anomaly detection. However, our proposal masking approach is proposed to mask suspicious anomalies as much as possible. Thus, with the masked patch sequence, the trained patch-level reconstruction model can reconstruct these masked patches by normal patterns. Thereby, the anomalous patches wouldn't be well reconstructed and the "identical reconstruction" issue can be mostly avoided. At last, we indicate that our model can detect anomalies from unseen classes directly without any fine-tuning. The only requirement is to provide normal samples from unseen classes to generate nominal feature prototypes. We evaluate our model and other state-of-the-art models under the multi- and cross-class settings, extensive experiments on two widely-used anomaly detection datasets show the superior performance of our model.

In summary, the contributions of this work are as follows:

1. We propose a novel PMAD approach for challenging multi- and cross-class anomaly detection. Our class-adaptive AD method can achieve to train one unified and generalizable model and doesn't require retraining, fine-tuning, nor extra normal feature distribution modeling for new classes.

2. We develop a two-stage anomaly detection manner based on two key designs: prototype-guided proposal masking and MAE-based patch-level reconstruction. The former is conducive to effectively mitigate the "identical reconstruction" issue, and the latter makes our method adaptive well to unseen classes.

3. We perform comprehensive experiments on two real-world AD datasets. The results show that our model substantially outperforms previous state-of-the-art models under the multi- and especially cross-class settings. The results also establish new baselines for future work in this important emerging direction.

Related Work

Reconstruction-Based Anomaly Detection. These methods are the most relevant to our approach and are based on the assumption that reconstruction models trained by normal samples only can reconstruct normal regions, but fail in abnormal regions. Early works mainly aim to train AutoEncoders (Bergmann et al. 2019b; Yang, Shi, and Qi 2020; Hou et al. 2021), Variational AutoEncoders (Liu et al. 2020) and GANs (Schlegl et al. 2017; Akcay, Atapour-Abarghouei, and Breckon 2018; Pidhorsky et al. 2018; Sabokrou et al. 2018) by only normal samples. However, these methods may sometimes confront the overfitting problem and fall into the "identical reconstruction" issue (Perera, Nallapati, and Xiang 2019), where the anomalies are also well reconstructed. To address this issue, researchers adopt many techniques, such as introducing structural information (Bergmann et al. 2019b), reconstructing semantic features (Yang, Shi, and

Qi 2020), utilizing memory mechanism (Gong et al. 2019; Hou et al. 2021) and generating pseudo-anomaly (Zavrtanik, Kristan, and Skocaj 2021), *etc*. In our approach, we design a proposal masking and reconstructing two-stage detection manner to avoid the "identical reconstruction" issue.

Masked Image Modeling. Recently, He, *et al.* show that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. However, directly reconstructing the missing pixels for vision pre-training would push the model to focus on short-range dependencies and high-frequency details (Ramesh et al. 2021). Some other works based on masked image modeling (MIM) propose to not use raw pixels as objective, such as MaskFeat (Wei et al. 2022) and BEIT (Bao, Dong, and Wei. 2021). Wei, *et al.* present MaskFeat, their approach aims to predict the features of the masked regions. Bao, *et al.* present BEIT, the objective of their approach is to predict the original visual tokens based on the corrupted image patches. These works all show that MIM pre-training is quite generalizable to downstream tasks. In this work, we indicate that MIM-based patch-level reconstruction models are more adaptive and generalizable for unseen classes than the conventional image-level reconstruction models.

Class-Adaptive AD Methods. The goal in some previous class-adaptive AD methods shares some similarities with our cross-class setting, where we both focus on the model's class adaptability to novel classes. In (Lu et al. 2020; Wu et al. 2021), the authors have proposed class-adaptive AD methods based on meta-learning algorithms. However, these methods generally train a meta-model on a large dataset by meta-learning algorithms, and treat all classes in the AD dataset as novel classes. For each class in the AD dataset, these meta-learned models should be further fine-tuned to adapt to this class with few-shot supporting samples. In contrast, our approach doesn't require complex meta-learning algorithms, nor need to train a meta-model on a large dataset. In (Huang et al. 2022), the authors proposed a registration-based anomaly detection (RegAD) framework, their approach is generalizable and can be applied to novel classes without re-training and fine-tuning. However, the RegAD has to construct a normal feature distribution model for each novel class, and different novel classes require different normal feature distribution models. By contrast, our approach doesn't need to construct any normal feature distribution models for novel classes, and a unified model can be applied to all novel classes.

Approach

Problem Statement

We first formally define the multi-class and cross-class anomaly detection tasks. Under the multi-class setting, we denote its training set as $\mathcal{I}^s_{train} = \{I^{ns}_i\}^N_{i=1}$, all the normal samples I^{ns}_i belong to the multiple seen classes $\mathcal{S} \subset \mathcal{C}$, where $\mathcal{C}$ denotes all possible image classes. The test set is denoted as $\mathcal{I}^s_{test} = \{I^{ns}_i\}^{N'}_{i=1} \cup \{I^{as}_j\}^M_{j=1}$, all the normal samples I^{ns}_i and abnormal samples I^{as}_j are from the seen classes $\mathcal{S}$. Under the cross-class setting, the training set is also composed of normal images from multiple seen classes $\mathcal{S}$. The test set consists of images from unseen class $\mathcal{U} \subset \mathcal{C}, \mathcal{U} \cap \mathcal{S} = \varnothing$, which is denoted as $\mathcal{I}^u_{test} = \{I^{nu}_i\}^{N''}_{i=1} \cup \{I^{au}_j\}^{M'}_{i=1}$, all the normal samples I^{nu}_i and abnormal samples I^{au}_j are from the unseen classes $\mathcal{U}$. The goal is to learn a unified model $m : \mathcal{I} \to \mathbb{R}$ that can assign larger anomaly scores for anomalies than normal samples in both seen and unseen classes.

Model Overview

Figure 2 overviews our proposed approach. The model consists of three parts: MAE-based patch-level reconstruction, prototype-guided proposal masking, and visual tokens based anomaly scoring. In the training phase, to obtain a more generalizable reconstruction model, we develop and train a patch-level reconstruction model rather than the image-level reconstruction models adopted in most AD methods. In the testing phase, to avoid that abnormal patches are also well reconstructed, we propose a prototype-guided proposal masking approach to mask suspicious anomaly proposals as much as possible. With the masked patch sequence, we use the trained reconstruction model to reconstruct these masked patches to determine if they are anomalies. The anomalies can be detected by large reconstruction uncertainty.

MAE-Based Patch-Level Reconstruction

To obtain a unified and generalizable reconstruction model for better adapting multi- and cross-class settings, we first describe our MAE-based patch-level reconstruction model.

Network Architecture. Different from the asymmetric encoder-decoder architecture in MAE, we employ standard ViT structure as both the encoder and decoder. The reason for network architecture modification is that: The encoder in MAE has a larger model capacity for a more powerful representation ability. However, in the AD task, the decoder is more critical, because the decoder with a small model capacity may lead to poor reconstruction for normal regions, causing normal patch misclassification. Note that to represent the masked patches, we replace the masked patches with a special mask token $[M]$ in the input embedding sequence (see Figure 2), which is a shared and learnable vector.

Combinational Masking for Training. For training a robust and generalizable patch-level reconstruction model, we propose a combinational masking strategy to generate masked patches during the training phase. The combinational masking is based on random and blockwise masking strategies. Random masking is the most simple and straightforward strategy: we uniformly sample random patches to mask. However, the random masking can't simulate the anomaly occurrence well, because the anomalies are usually continuous regions in the image. Therefore, we also need to generate continuous masked regions during training, so as to ensure that the model will not misclassify the continuous masked regions during testing. To this end, we employ a blockwise masking strategy. In this strategy, a block of image patches is masked each step, we repeat the masking step until obtaining enough masked patches. The procedure of blockwise masking is summarized in Algorithm 1 in Appendix. Further considering that generating more masked patches in harder-to-reconstruct regions is

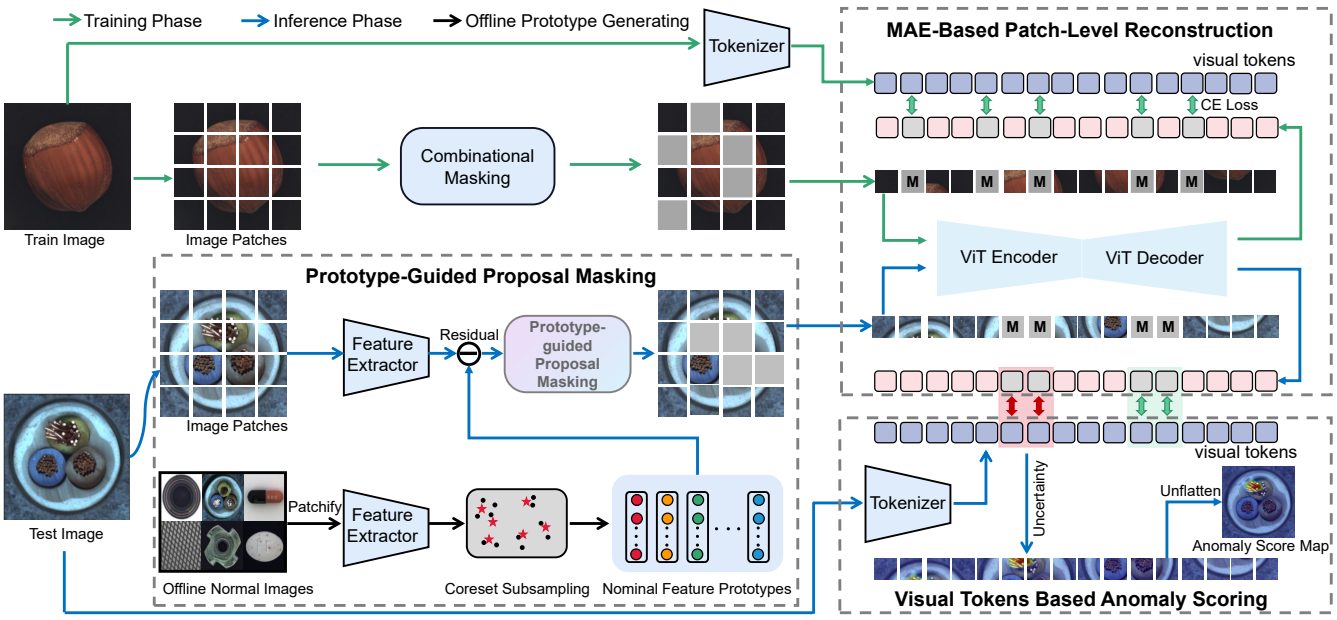

Figure 2: Model overview. The model is composed of three parts: MAE-based patch-level reconstruction, prototype-guided proposal masking, and visual tokens based anomaly scoring. During training, the masked patches are replaced with a special mask token $[M]$. The learning objective is to predict the visual tokens of the masked patches based on the encoded embeddings of the unmasked patches. During inference, we first generate nominal prototypes by Coreset Subsampling for both seen and unseen classes. Then, we employ a prototype-guided proposal masking module to mask suspicious anomaly proposals as much as possible. The uncertainty in the prediction of visual tokens can be utilized as anomaly score.

conducive to train more generalizable models with stronger reconstruction ability. We thus propose a frequency-based masking strategy to generate more masked patches for high-frequency regions, a dynamic masking strategy to generate masked patches based on the current reconstruction state of the model, and a region-limited masking strategy to generate more masked patches for the foreground regions. Finally, we combine these masking strategies with the basic random and blockwise masking strategies to formulate the combinational masking strategy (see details in Appendix).

Model Adaptability to Unseen Classes. Our model's training procedure is to allow the model to learn how to utilize the contextual relationship in the image to infer the features of the masked patches. Thus, even in unseen classes, the masked patches can be reconstructed well by employing the non-masked patches and the adaptive reconstruction capability of the model, so anomalies can still be detected by large reconstruction errors. Because there are always visible patches that can be exploited within a single image, our patch-level reconstruction model is more adaptive to unseen classes (as validated in the Experiments).

Prototype-Guided Proposal Masking

Random and blockwise masking are not suitable for inference, because these strategies may leak a large amount of abnormal information, thus causing the "identical reconstruction" issue. To address this issue, we propose a prototype-guided proposal masking approach to mask suspicious anomaly proposals as much as possible. With the masked patch sequence, the trained patch-level reconstruction model can reconstruct these masked patches by normal patterns. Thereby, the anomalous patches will not be well reconstructed. As shown in Figure 2, we first generate nominal feature prototypes for each object class. Specifically, we utilize an ImageNet pre-trained network ϕ to extract normal features $\mathcal{F}_N = \bigcup_{x_i \in \mathcal{X}_N} \phi(x_i)$ from normal samples $\mathcal{X}_N$, and then employ the coreset subsampling mechanism (Sener and Savarese 2018; Sinha et al. 2020) to generate nominal feature prototypes $\mathcal{P}$ for each object class. The test image will be divided into patches, and then the image patches will be sent into the same pre-trained network to extract test features $\mathcal{F}_T$. Then, both the test features $\mathcal{F}_T$ and nominal prototypes $\mathcal{P}$ will be fed into the Proposal Masking module to generate abnormality ranking for all test image patches. In the Proposal Masking module, each test patch feature $f_p^i \in \mathcal{F}_T, f_p^i \in \mathbb{R}^d$ will match a corresponding nearest nominal feature prototype $f_n^i = \underset{f \in \mathcal{P}}{\operatorname{argmin}} ||f - f_p^i||_2$ from the feature prototype pool $\mathcal{P}$. The residual feature $f_r^i = f_p^i - f_n^i, f_r^i \in \mathbb{R}^d$ will be sent into a normalizing flow (NF) model φ_θ (Dinh, Sohl-Dickstein, and Bengio 2016) to obtain normalized residual feature $\varphi_\theta(f_r^i)$. More details of the Proposal Masking module are provided in Appendix. We can utilize the following function to evaluate the abnormality of each patch:

$$a(f_r^i) = \max_{f_r^j \in \mathcal{F}_r} (\exp(\mathrm{ll}(f_r^j))) - \exp(\mathrm{ll}(f_r^i)) \quad (1)$$

where $\mathcal{F}_r = \{f_r^i\}_{i=1}^M$ denotes all residual features and the $\mathrm{ll}(f_r^i)$ means the log-likelihood of f_r^i. $\mathrm{ll}(f_r^i)$ can be evaluated by the NF model φ_θ as follows (Gudovskiy et al. 2022):

$$\mathrm{ll}(f_r^i) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\varphi_\theta(f_r^i)^T\varphi_\theta(f_r^i) + \log\left|\det J\right| \quad (2)$$

where d is the feature dimension and $J = \frac{\partial\varphi_\theta(f_r^i)}{\partial f_r^i}$ is the Jacobian matrix. We then can form an abnormality ranking of image patches by ranking all the abnormality scores. With the pre-defined mask ratio m, we select the top m percent of the image patches in the abnormality ranking as masked patches. Finally, we indicate that the masking method may have the "mis-masking" issue (*i.e.*, normal patches are incorrectly masked in unseen classes), due to the normal patterns of unseen classes being significantly different from the known normal patterns. However, our approach can avoid this issue with the guidance of nominal prototypes, because the distribution of normal residual features would not remarkably shift from the learned distribution even in unseen classes (see intuitive explanation in Appendix). Thus, not too many normal patches will be masked by our proposal masking approach. Note that for unseen classes, our method only requires generating nominal feature prototypes offline by normal samples from unseen classes.

Visual Tokens Based Anomaly Scoring

After obtaining the prediction results, we can then design anomaly scoring function for anomaly detection.

Reconstruction Objective. The most straightforward objective is obviously the raw pixels as used in MAE (He et al. 2022). However, this objective is not suitable for anomaly detection, because raw pixels as targets have a potential risk of overfitting to local statistics and high-frequency details (Ramesh et al. 2021). Moreover, when we use the raw pixels to evaluate the reconstruction error, it would be affected by the image details (*i.e.* normal patches with rich details may also have large reconstruction errors). The above mentioned issues may cause degraded anomaly detection performance. Thus, we propose to employ visual tokens as the reconstruction objective, this is described as follows:

Visual Tokens. We follow DALL-E (Ramesh et al. 2021) to compress an image with a dVAE codebook. In particular, each patch is encoded into a discrete visual token, and the vocabulary size is set to $|\mathcal{V}| = 8192$. As shown in Figure 2, we first tokenize each image to 14×14 grid of visual tokens by a pre-trained publicly available image tokenizer described in DALL-E. With the encoded visual tokens, the task is converted to predict the visual token distribution of the masked patches by optimizing a cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|\mathcal{M}_p|}\sum_{m=1}^{|\mathcal{M}_p|}\sum_{i=1}^{|\mathcal{V}|} p_m^i \log(p_m^i) \quad (3)$$

where $|\mathcal{M}_p|$ is the number of masked patches in each image, and p_m^i indicates the probability that the mth masked patch belongs to the ith visual token.

Anomaly Scoring. For visual tokens, the dimension of output vectors is 8192, where each dimension p_i indicates the probability that the patch belongs to a special visual token. Thus, we can calculate cross-entropy to measure the uncertainty of each patch. The larger the uncertainty, the more likely the patch is to be abnormal. The anomaly scoring function is as follows:

$$s = -\sum_{i=1}^{|\mathcal{V}|} p_i \log(p_i) \quad (4)$$

We then multiply s with the abnormality scores produced at the masking stage to get final anomaly scores, which we find are more robust to detect and evaluate anomalies.

Experiments

Datasets and Metrics

We evaluate the proposed approach on two widely used industrial anomaly detection datasets: the MVTecAD (Bergmann et al. 2019a) and BTAD (Mishra et al. 2021).

MVTecAD. This dataset contains 5354 high-resolution images (3629 images for training and 1725 images for testing) of 15 different categories. 5 classes consist of textures and the other 10 classes contain objects. A total of 73 different defect types are presented and almost 1900 defective regions are manually annotated in this dataset.

BTAD. This dataset contains 2830 real-world images of 3 industrial products. Product 1, 2 and 3 of this dataset contain 400, 1000 and 399 training images respectively.

Evaluation Metrics. The performance of our PMAD and all comparison methods are evaluated by the area under the curve (AUC) of the receiver operating characteristic (ROC) at the image or pixel level (AUROC).

Implementation Details

We mainly follow the hyperparameters in (Bao, Dong, and Wei. 2021) to train the reconstruction model. All training hyperparameters are listed in Appendix. Because the sizes of anomalies in different classes are generally different, the same mask ratio for all classes cannot achieve the optimal results. We select a suitable mask ratio for each class through extensive experiments (see Appendix for details).

Results under the Single-Class Setting

Setup. Existing anomaly detection algorithms are almost evaluated under this paradigm, where a specific model is trained for each object class.

Baselines. We compare our approach with the state-of-the-art AD methods, including DFR (Yang, Shi, and Qi 2020), PaDiM (Defard et al. 2021), PatchSVDD (Yi and Yoon 2021), DRAEM (Zavrtanik, Kristan, and Skocaj 2021), MSFD (Salehi et al. 2021), and CFLOW (Gudovskiy et al. 2022). All these methods are representative methods in the AD community. The DFR and DRAEM are conventional image-level reconstruction models. Results of these methods are reproduced using the public implementations.

Quantitative Results. The left part of Table 1 shows the comparison results under the single-class setting. Our PMAD can achieve comparable results with the state-of-the-art methods under the single-class setting on both MVTecAD and BTAD datasets.

Method	Single-Class Setting		Multi-Class Setting	
	MVTecAD	BTAD	MVTecAD	BTAD
DFR	0.942/0.953	**0.950**/0.971	0.867/0.916	**0.948**/0.969
PaDiM	0.966/**0.971**	0.943/0.962	0.894/0.954	0.938/0.966
PatchSVDD	0.892/0.899	0.852/0.765	0.691/0.798	0.801/0.746
DRAEM	**0.971**/0.967	0.944/0.922	0.918/0.891	0.912/0.919
MSFD	0.907/0.949	0.916/0.962	0.888/0.944	0.897/0.962
CFLOW	0.901/0.935	0.931/0.971	0.890/0.940	0.930/0.966
PMAD (ours)	0.961/0.968	0.936/**0.974**	**0.945/0.956**	0.938/**0.973**

Table 1: AUROC results on two real-world AD datasets under the single- and multi-class settings. $\cdot/\cdot$ means image-level AUROC and pixel-level AUROC, respectively.

Results under the Multi-Class Setting

Setup. Normal samples from multiple classes are simultaneously used to train a unified model, and the trained model is utilized to detect anomalies from the same trained classes.

Baselines. Under the multi-class setting, we use the same baseline methods mentioned in the last section but train these methods with multiple classes simultaneously.

Quantitative Results. As shown in the right part of Table 1, our PMAD can achieve better results compared with the SOTA methods under the multi-class setting. It can be found that the performances of all baseline methods drop dramatically under the multi-class setting. The previous SOTA, DRAEM, suffers from a drop of near 5.3% and 7.6% at the image-level and the pixel-level respectively. For another SOTA, PaDiM, has a performance drop of 7.2% at the image-level. The PatchSVDD has the largest performance degradation, which is as large as 20.1% and 11%. However, our PMAD has only a small performance drop from the single-class setting to the multi-class setting (-1.6%/-1.2%). Moreover, we beat the best competitor (DRAEM) under the multi-class setting by a large margin (2.7%) at the image-level, demonstrating the superiority of our approach. For the BTAD dataset, all the classes belong to texture classes, which have much simpler normal patterns. Thus, even in the multi-class setting, most methods have no significant performance degradation. Our method has almost no performance degradation.

Results under the Cross-Class Setting

Setup. Normal samples are limited to be drawn from partial classes only, and all samples from these classes are removed from the test set to ensure that the test set contains only samples from unseen classes. To validate the multi- and cross-class performance of the models simultaneously, we adopt the grouping method to divide the dataset. That is to select some classes as training classes and the remaining classes for testing. On the MVTecAD dataset, we divide the training and testing classes separately for texture and object categories. On the BTAD dataset, since there are only 3 classes, we adopt another setup: one class for training and the remaining classes for testing.

Quantitative Results. The detection results are presented in Table 2. All these methods are directly without any fine-tuning utilized to detect anomalies in unseen classes. As shown in Table 2, our method can outperform these SOTA [1] methods significantly under the cross-class setting. It can be found that without re-training or fine-tuning, most of these SOTA methods fail completely in unseen classes, but our method still has good anomaly detection results. For texture classes, our approach can outperform the best SOTA method by (2.5%/0.7% and 0.8%/0.1%). For more complex object classes, our approach can outperform the best SOTA method by a significantly large margin (17.4%/4.7% and 14.7%/8.4%). For the BATD dataset, our approach can also outperform the SOTA methods by a significant margin (5.7%/6.9%, 16.1%/22.7%, and 6.2%/11.4% for Product 1, 2, and 3, respectively). Moreover, we also compare our approach with a registration-based class-adaptive AD model (RegAD). In (Huang et al. 2022), the RegAD is evaluated by a similar experimental setting to our cross-class setting. Our PMAD can achieve much better results than RegAD on both MVTecAD and BTAD datasets. In Appendix, we also show the cross-class detection results from objects to textures and from textures to objects on the MVTecAD dataset. Our approach can outperform the SOTA methods by a significant margin (7.5%/3.0% and 11.3%/2.9%), and also achieve much better results than RegAD. All these results reflect the superior class adaptability of our model.

Ablation Study

All the ablation study results are shown in Table 3.

Network Architecture (NA). The ViT architecture can achieve better detection results than the asymmetric architecture. This means that the architecture designed in MAE is not suitable for the AD task, and decoders with larger model capacity are more conducive for reconstruction in AD tasks.

Masking Strategy in Training (MST). Masking strategy in training doesn't have much effect on the detection results, even the simplest random masking strategy can achieve good detection results. Compared with random masking and blockwise masking, our combinational masking strategy can enable the network to learn better reconstruction capabilities, thus achieving better detection results.

Reconstruction Objective (RO). It can be found that raw pixels will result in much worse detection performance. Because when we use the raw pixels to evaluate the reconstruc-

[1] This only represents the unsupervised SOTA methods, except RegAD. RegAD belongs to the class-adaptive AD method.

Method	Cross-Class Setting						
	Seen Classes (On MVtecAD)				Seen Classes (On BTAD)		
	Textures(1)	Objects(1)	Textures(2)	Objects(2)	Product 1	Product 2	Product 3
DFR	0.792/0.502	0.595/0.799	0.585/0.499	0.409/0.733	0.872/0.778	0.611/0.550	0.672/0.696
PaDiM	0.870/0.773	0.473/0.827	0.989/0.985	0.536/0.752	0.648/0.778	0.531/0.723	0.556/0.736
PatchSVDD	0.920/0.773	0.721/0.847	0.911/0.852	0.697/0.848	0.756/0.835	0.828/0.660	0.823/0.676
DRAEM	0.766/0.676	0.549/0.696	0.804/0.709	0.513/0.617	0.576/0.542	0.709/0.629	0.644/0.553
MSFD	0.720/0.647	0.692/0.864	0.982/0.985	0.607/0.814	0.721/0.907	0.611/0.368	0.686/0.767
CFLOW	0.917/0.889	0.565/0.804	0.985/0.985	0.525/0.795	0.802/0.892	0.605/0.753	0.836/0.846
RegAD	0.874/0.838	0.667/**0.911**	0.900/0.924	0.668/0.917	0.679/0.779	0.665/0.793	0.666/0.773
PMAD (ours)	**0.945/0.896**	**0.895/0.911**	**0.997/0.986**	**0.844/0.932**	**0.929/0.976**	**0.989/0.980**	**0.898/0.960**

Table 2: AUROC results on two real-world AD datasets under the cross-class setting. Textures(1) contains carpet and leather as seen classes, and other texture classes in the MVTecAD dataset as unseen classes. Objects(1) contains bottle, cable, capsule, screw, and transistor as seen classes, and other object classes in the MVTecAD dataset as unseen classes. The seen classes in Textures(2) and Objects(2) are the unseen classes in Textures(1) and Objects(1), respectively.

Ablations		Multi-Class Setting
		MVTecAD
NA	Asymmetric Architecture	0.918/0.937
	ViT structure	**0.945/0.956**
MST	Random Masking	0.929/0.945
	Blockwise Masking	0.939/0.950
	Combinational Masking	**0.945/0.956**
RO	Raw Pixels	0.773/0.712
	Deep Features	0.844/0.867
	Visual Tokens	**0.945/0.956**
IMS	Random Masking	0.730/0.667
	Blockwise Masking	0.749/0.700
	Proposal Masking	**0.945/0.956**

Table 3: Ablation study results. Best results are highlighted. IMS means inference masking strategy.

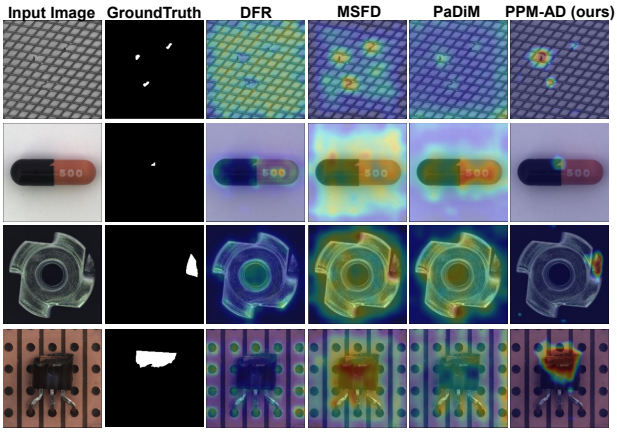

Figure 3: Qualitative results. The anomaly score maps are generated under the cross-class setting, where the training set doesn't contain the shown classes.

tion errors, it would be affected by the image details (normal patches with rich details may also have large reconstruction errors). Compared with raw pixels, higher-level and more semantic visual representation objectives can achieve a significant performance improvement, such as visual tokens and deep features. Moreover, visual tokens can achieve better detection results compared to deep features.

The Effect of Inference Proposal Masking. The results in Table 3 show that the proposal masking strategy is much crucial for achieving better detection results. When we use random or blockwise masking in the inference phase, the detection results will drop significantly (about -20%/-25%). This is because random and blockwise masking may generally leak a large amount of anomaly information, further causing the abnormal patches to be also well reconstructed. By contrast, our proposal masking strategy can achieve a significant performance improvement, because the suspicious abnormal patches will be masked as much as possible.

Qualitative Results

Qualitative Results. We visualize some anomaly localization results in Figure 3 with the MVTecAD dataset and under the cross-class setting. It can be found that most SOTA methods fail to generate anomaly localization maps for unseen classes, while our PMAD can still generate good anomaly localization results.

Conclusion

Class adaptability is a critical but still not well-studied issue in the anomaly detection community. Considering this issue, we propose a novel and class-adaptive PMAD approach based on two key designs: MAE-based patch-level reconstruction and prototype-guided proposal masking. Under the multi- and cross-class settings, our model illustrates better class adaptability than the SOTA models. We expect our results can establish new baselines for future work in this important emerging direction. One main limitation of our approach is that the ViT model can only reconstruct 16×16 image patches, but cannot reconstruct more fine-grained image patches. Therefore, the anomaly localization ability of our model is limited. In the future, we plan to investigate how to employ hierarchical transformers and design a multi-scale masking strategy to further improve our method.

Acknowledgements

This work was supported in part by the National Natural Science Fund of China (61971281), the National Key R&D Program of China (2021YFD1400104), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Science and Technology Commission of Shanghai Municipality (18DZ2270700).

References

Acsintoae, A.; Florescu1, A.; Georgescu, M.-I.; Mare, T.; Sumedrea1, P.; Ionescu, R. T.; Khan, F. S.; and Shah, M. 2022. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. *In CVPR*.

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. GANomaly: Semi-supervised anomaly detection via adversarial training. *In ACCV*, 622–637.

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

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019a. Mvtec AD - A comprehensive real-world dataset for unsupervised anomaly detection. *In CVPR*.

Bergmann, P.; Löwe, S.; Fauser, M.; and Sattlegger, D. 2019b. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *In ICCVTA*.

Biase, G. D.; Blum, H.; Siegwart, R.; and Cadena, C. 2021. Pixel-wise Anomaly Detection in Complex Driving Scenes. *In CVPR*.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization. *In The 1st International Workshop on Industrial Machine Learning*, 475–489.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*.

Gong, D.; Liu, L.; Le, V.; and Saha, B. 2019. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. *In ICCV*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *In Conference and Workshop on Neural Information Processing Systems*.

Gudovskiy, D.; Ishizaka, S.; ; and Kozuka, K. 2022. CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows. *In WACV*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollar, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. *In CVPR*.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *In Science 313, 5786*.

Hou, J.; Zhang, Y.; Zhong, Q.; Xie, D.; Pu, S.; and Zhou, H. 2021. Divide-and-Assemble: Learning Block-wise Memory for Unsupervised Anomaly Detection. *In ICCV*.

Huang, C.; Guan, H.; Jiang, A.; Zhang, Y.; Spratling, M.; and Wang, Y.-F. 2022. Registration based Few-Shot Anomaly Detection. *In ECCV*.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. *In CVPR*.

Liu, W.; Li, R.; Zheng, M.; Karanam, S.; Wu, Z.; Bhanu, B.; Radke, R. J.; and Camps, O. 2020. Towards visually explaining variational autoencoders. *In CVPR*.

Lu, Y.; Yu, F.; Reddy, M. K. K.; and Wang, Y. 2020. Few-Shot Scene-Adaptive Anomaly Detection. *In ECCV*.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: Avision Transformer Network for Image Anomaly Detection and Localization. *arXiv preprint arXiv:2104.10036*.

Perera, P.; Nallapati, R.; and Xiang, B. 2019. OCGAN: one-class novelty detection using gans with constrained latent representations. *In CVPR*.

Pidhorsky, S.; Almohsen, R.; A.Adjeroh, D.; and Doretto, G. 2018. Generative probabilities novelty detection with adversarial autoencoders. *In NeurIPS*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. *In ICML*.

Reiss, T.; Cohen, N.; Bergman, L.; and Hoshen, Y. 2021. PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation. *In CVPR*.

Roth, K.; Pemula, L.; Zepeda, J.; Scholkopf, B.; Brox, T.; and Gehler, P. 2022. Towards Total Recall in Industrial Anomaly Detection. *In CVPR*.

Sabokrou, M.; Khaloori, M.; Fathy, M.; and Adeli, E. 2018. Adversarially learned one-class classifier for novelty detection. *In CVPR*.

Salehi, M.; Sadjadi, N.; Rohban, S. H.; and R.Rabiee, H. 2021. Multiresolution Knowledge Distillation for Anomaly Detection. *In CVPR*.

Schlegl, T.; Seebock, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *In ICIPMI*, 146–157.

Sener, O.; and Savarese, S. 2018. Active learning for convolutional neural networks: A core-set approach. *In ICLR*.

Sinha, S.; Zhang, H.; Goyal, A.; Bengio, Y.; Larochelle, H.; and Odena, A. 2020. Small-GAN: Speeding up GAN training using core-sets. *In ICML*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world Anomaly Detection in Surveillance Videos. *In CVPR*.

Tian, Y.; Pang, G.; Liu, F.; Chen, Y.; Shin, S. H.; Verjans, J. W.; Singh, R.; and Carneiro., G. 2021. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. *Medical Image Computing and Computer Assisted Intervention*, 128–140.

Vojir, T.; Sipka, T.; and Aljundi, R. 2021. Road Anomaly Detection by Partial Image Reconstruction with Segmentation Coupling. *In ICCV*.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and Feichtenhofer, C. 2022. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *In CVPR*.

Wu, J.-C.; Chen, D.-J.; Fuh, C.-S.; and Liu, T.-L. 2021. Learning Unsupervised Metaformer for Anomaly Detection. *In ICCV*.

Yang, J.; Shi, Y.; and Qi, Z. 2020. DFR: Deep Feature Reconstruction for Unsupervised Anomaly Segmentation. *arXiv preprint arXiv: 2012.07122*.

Yao, X.; Zhang, C.; and Li, R. 2022. Explicit Boundary Guided Semi-Push-Pull Contrastive Learning for Better Anomaly Detection. *arXiv preprint arXiv:2207.01463*.

Yi, J.; and Yoon, S. 2021. Patch SVDD: Patch-Level SVDD for Anomaly Detection and Segmentation. *In ACCV*.

Zavrtanik, V.; Kristan, M.; and Skocaj, D. 2021. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. *In ICCV*.

Zhang, J.; Xie, Y.; Pang, G.; Liao, Z.; Verjans, J.; Li, W.; Sun, Z.; He, J.; Li, Y.; Shen, C.; and Xia, Y. 2021. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. *Medical Imaging*, 879–890.