Masked Attention ConvNeXt Unet with Multi-Synthesis Dynamic Weighting for Anomaly Detection and Localization

Shih-Chih Lin
leolin65@gapp.nthu.edu.tw
Ho-Weng Lee
howeng0901@gapp.nthu.edu.tw
Yu-Hsuan Hsieh
ss111062646@gapp.nthu.edu.tw
Cheng-Yu Ho
lessthan41@gapp.nthu.edu.tw
Shang-Hong Lai
lai@cs.nthu.edu.tw

National Tsing Hua University
Hsinchu, Taiwan

Abstract

In recent years, self-supervised models like Cutpaste, Mask, NSA, and Perlin have gained popularity for anomaly detection. These models generate synthetic data by employing various data augmentation strategies, demonstrating their potential for improving anomaly detection through learned representations. In this study, we introduce an algorithm called Multi-Synthesis Dynamic Weighting (MSdW) to leverage the advantages of diverse synthetic data. MSdW enables the model to learn various abnormal conditions during training, thereby enhancing accuracy. Our model architecture consists of reconstructive and discriminative subnetworks, both utilizing the UNet architecture. The encoders in both subnetworks employ modern ConvNets, specifically ConvNeXtV2, for proficient feature extraction. Additionally, we propose an attention mechanism known as Self-Supervised Predictive Convolutional Block with Multi-Attentions (SSPCBMA), which is seamlessly integrated into the reconstructive subnetwork to enhance feature extraction capabilities. We evaluate our proposed model on multiple datasets designed for anomaly detection and segmentation tasks, including MVTecAD, BTAD, and KSDD2. These datasets serve various purposes, and our model outperforms the state-of-the-art methods, particularly in terms of Pixel AP and PRO indices.

1 Introduction

Anomaly detection (AD) is a crucial task with various applications, such as industrial defect detection[1], medical detection[24], and video surveillance[7]. In unsupervised AD, no prior information about anomalies is available, and only a set of normal samples is provided for reference. To address this problem, previous studies have constructed various selfsupervision tasks on anomaly-free samples, including sample reconstruction[27], pseudo-outlier augmentation[11], and knowledge distillation[5]. Several studies focus on unsupervised anomaly detection from a reconstruction-based perspective, using generative models such as AutoEncoder(AE)[16, 23, 32], VAEs [38], Generative Adversarial Nets(GANs) [12] for sample reconstruction. Although these methods rely on the hypothesis that generative models trained on normal samples can successfully reconstruct anomaly-free regions but fail for anomalous regions, recent studies show that deep models generalize so well that even anomalous regions can be well-restored. To address this issue, memory mechanisms[22], image masking strategies[10, 33], pseudo-anomaly[11, 23], data augmentation strategies[14, 19, 32], and attention mechanism[15, 21] are incorporated into reconstruction-based methods.

In this study, we present a novel pipeline called Masked Attention ConvNeXt UNet with Multi-Synthesis Dynamic Weighting (MACoW), designed for the tasks of anomaly detection and segmentation. Our model capitalizes on the integration of several self-supervised methods that employ diverse data augmentation strategies, including Cutpaste[11], Mask[10, 33], NSA[23], and Perlin[32]. We introduce an algorithm named Multi-Synthesis Dynamic Weighting (MSdW) to harness the benefits of diverse synthetic data. Moreover, we enhance the effectiveness of information extraction by incorporating the latest ConvNeXtV2 and UNet architectures for feature extraction. To foster stronger relationships between features in the channel and spatial dimensions, we integrate a Self-Supervised Predictive Convolutional Block with Multi-Attention (SSPCBMA) mechanism into both the final encoder and initial decoder paths of our ConvNeXt UNet subnetwork. This integration significantly improves the model's ability to extract target-specific features, leading to promising results.

We evaluate our proposed approach on various datasets for anomaly detection and segmentation tasks, including MVTecAD [1], BTAD[18], and KSDD2[3]. The results demonstrate that the model with the SSPCBMA and MSdW outperforms other state-of-the-art(SOTA) methods. In summary, our paper makes the following contributions:

- We introduce a novel model, the Masked Attention ConvNeXt UNet with Multi-Synthesis dynamic Weighting (MACoW), and exploit multiple anomaly synthesis methods in our training framework to reach state-of-the-art(SOTA) performance.
- We propose the Self-Supervised Predictive Convolutional Block with a Multi-Attention (SSPCBMA) to boost accuracy performance.
- Our proposed model experimented on MVTecAD, BTAD, and KSDD2, surpassing other SOTA methods in anomaly detection and segmentation performance, especially in Pixel AP and PRO metrics.

2 Related Works

Reconstruction-based approaches are commonly employed to detect anomalies in image space, using generative models, and involve two steps: (1) reconstructing the image and (2) comparing the original and reconstructed images to obtain anomaly maps[4, 26]. To reconstruct the image, previous works primarily utilized denoising autoencoders [10, 32, 33] to facilitate the network in capturing the normal distribution and avoiding identity mapping during training. In these methods, the original image is corrupted with specific noise to enable the network to eliminate it. It was also an attribute removal-and-restoration framework

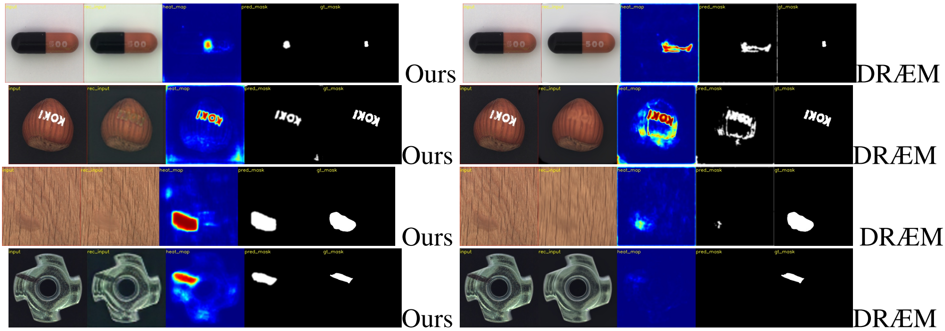

Figure 1: In the comparison between our model and DRÆM[32], the results are displayed side by side, with each row representing a different class—capsule, hazelnut, wood, and metal nut from top to bottom. The sequence of images, progressing from left to right, includes the input image, reconstruction result, anomaly map, predicted mask (PM), and ground truth mask (GT). It is apparent from the visual comparison that our model surpasses DRÆM[32] in terms of both reconstruction quality and the accuracy of anomaly localization.

[9]. This framework argues that the network can learn more robust features by restoring the removed attributes. These methods often utilize supervised ImageNet pre-trained [20, 25] as either feature extractors or initialization for fine-tuning. After reconstruction, anomalies can be detected by comparing the original and reconstructed images using various functions, such as L2 distance, L1 smooth distance, and structural similarity (SSIM) [28]. Given the complex nature of identifying abnormal patterns in images, a variety of models have been proposed that leverage both local and global information. One popular approach is to integrate the self-attention mechanism[15, 21] of the reconstruction model, which models long-range interactions between different regions of the image and demonstrates success in detecting abnormal patterns and has been actively explored.

Discriminative unsupervised anomaly detection methods [11, 32] employ synthetically generated anomalies to train a discriminative anomaly segmentation network. To alleviate overfitting on the synthetic anomaly appearance, a reconstruction network is utilized in [32] to restore the normal appearance of the synthetic anomalies. Subsequently, the discriminative network learns a distance function[13, 37] between the original image and its reconstruction to detect anomalies, typically using the Focal loss[32] or Dice loss[31] functions. However, the limited distribution of generated synthetic anomalies may cause the reconstruction network to overfit the synthetic appearance, reducing performance in detecting near-distribution anomalies [34]. In this paper, we employ four anomaly generation methods in our model training framework, thus preventing overfitting.

3 Proposed Method

Our proposed MACoW architecture comprises two subnetworks, reconstructive and discriminative, as depicted in Figure 2. The reconstructive subnetwork is responsible for reconstructing the normal regions and repairing the abnormal regions of input images that have been corrupted. On the other hand, the discriminative subnetwork performs further localization of anomalous areas by comparing the difference and consistency between the input

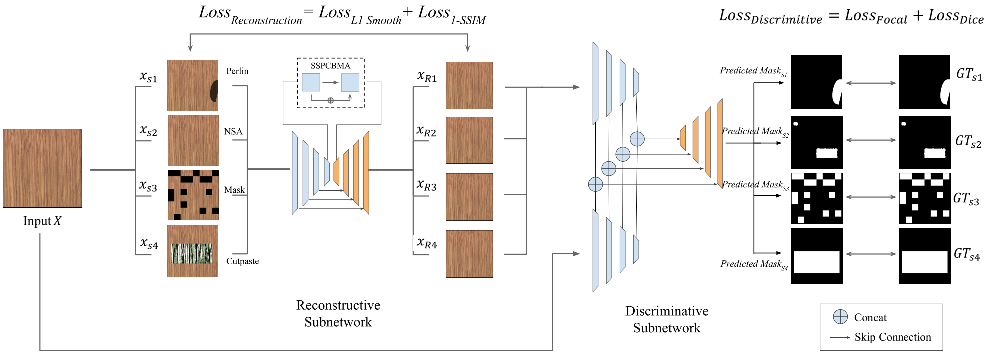

Figure 2: The architecture of our proposed model: MACoW. We train a reconstructor ConvNeXtUnetV2 X_{Rec} by including four anomaly synthesis methods using the L1 Smooth and SSIM[28], followed by a discriminator ConvNeXtUnetV2 for generating an anomaly map. The segmentation network, trained using the Focal Loss[13] and Dice Loss[17]. Best viewed by zooming in.

image and its flawless approximation, thereby enhancing the segmentation performance.

3.1 Reconstructive Subnetwork

This paper introduces the proposed Masked Attention ConvNeXtUNetV2 model architecture, as depicted in Figure 2. The model comprises two main parts: constructive and discriminative subnetworks. The downsampling operation on the constructive subnetwork utilizes the ConvNeXtV2[30] network as the backbone feature extraction network, with a stem operation executed before entering ConvNeXtV2[30]. For the upsampling process, we employ the bilinear upsampling method instead of the original transpose convolution process. This alternative mitigates the shadow problem associated with transpose convolution. Each upsampled feature map is concatenated with the feature obtained by the encoder via skip connections[8]. This method enhances the foreground target while minimizing the noise response generated during feature fusion across different channel numbers. The reconstructive ConvNeXtUnetV2 subnetwork, trained using the L1 Smooth and SSIM[28].

3.2 Discriminative Subnetwork

We introduce a discriminative subnetwork within our architecture, designed to function as a classifier for end-to-end anomaly detection while utilizing ConvNeXtV2 as the feature extraction backbone, as detailed in ConvNeXtV2[30]. The core functionality of this subnetwork is the generation of a predicted anomaly annotation map, where each pixel is assigned a binary label of either 0 (indicating normal) or 1 (indicating abnormal).

To enhance the efficiency and coherence of our architecture, we adopt a shared feature extraction process for both the input image and its reconstructed counterpart. These features, derived from each encoder layer, are concatenated and denoted as $concat_{feat}(i)$, with i representing the layer index (e.g., $i = 1$ corresponds to the input layer). During the decoding phase,

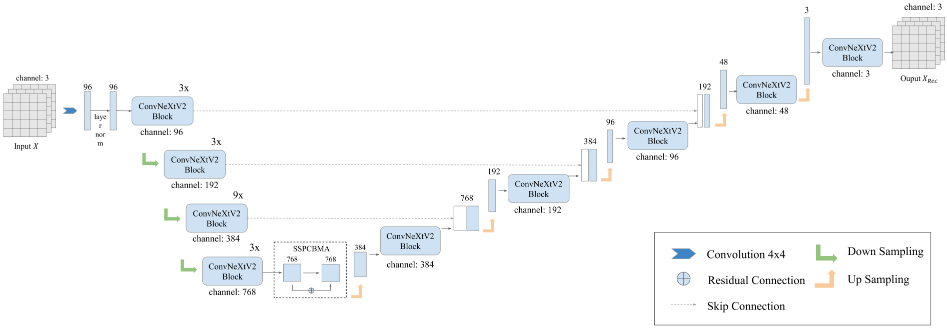

Figure 3: The structure of our reconstruction Unet. The encoder utilizes ConvNeXtV2[30] block for feature extraction. An attention mechanism of SSPCBMA is applied before upsampling. Best viewed by zooming in.

the input of the first decoder layer is $concat_{feat}(L)$, where L signifies the total number of encoder layers. Subsequently, the input for the ith decoder layer ($i = L$ indicating the final layer that outputs the segmentation map) is computed as $concat(concat_{feat}(i), \text{last layer's output})$, where the last layer's output is not considered if $i = 1$.

This approach optimizes the feature extraction process for both the input and reconstructed images, enhancing model performance for anomaly segmentation tasks. The discriminative network is trained using a combination of Focal loss[13] and Dice Loss[17] to effectively localize anomalous regions and produce an anomaly map, from which image-level anomaly scores are derived.

3.3 Self-Supervised Predictive Convolutional Block with Multi-Attention(SSPCBMA)

Our work draws inspiration from the self-supervised predictive convolutional attentive block (SSPCAB)[21], which focuses on the task of predicting or reconstructing masked information by leveraging contextual cues. In our approach, we introduce a novel modification to this module with the objective of enhancing its performance.

Firstly, we incorporate residual connections that establish links both before the masked convolution operation and after the spatial attention operations. These connections serve the purpose of facilitating information flow and gradient propagation within the module.

Secondly, we introduce spatial attention mechanisms inspired by CBAM[29] to the module, enabling it to emphasize critical features while simultaneously suppressing irrelevant ones. This spatial attention component contributes to enhancing the overall effectiveness of the module in capturing relevant information.

The resultant module, denoted as the Self-Supervised Predictive Convolutional Block with Multi-Attention (combining both channel and spatial attention), abbreviated as SSPCBMA, is illustrated in Figure 4. We seamlessly integrate SSPCBMA into our ConvNeXtUNetV2 architecture, as depicted in Figure 3. This integration equips our architecture with the capability to learn and reconstruct masked information while simultaneously providing valuable features for downstream neural layers.

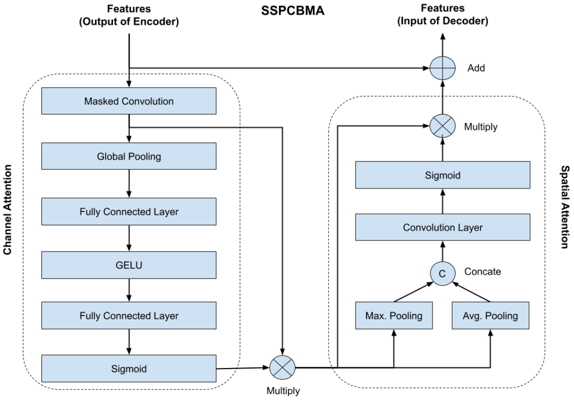

Figure 4: The structure of SSPCBMA.

3.4 Simulated Anomaly Generation

Our research aims to leverage SOTA synthesis methods to simulate anomalous images, thereby enhancing the training of our MACoW model for anomaly detection. We further explore novel approaches by drawing inspiration from prior works such as DRÆM[32], which introduces external data sources as the anomaly source and incorporates Perlin masks to create masks with distinctive, non-uniform shapes. Specifically, Cutpaste[11] generates anomalies by selecting image patches from the same dataset and placing them in different regions. NSA[23], on the other hand, seamlessly blends scaled patches of varying sizes from separate images using Poisson image editing, yielding anomalies that closely mimic real-world anomalies in visual appearance.

Our approach builds upon these established synthesis methods and incorporates four distinct anomaly simulators: Perlin noise[32], NSA[23], Mask[10, 33], and CutPaste[11]. We adapt and fine-tune certain algorithms and parameters to enhance model accuracy. For example, we introduce a mask method where the number and scale of mask regions are parameterized and refined to optimize model performance. A visual representation of generated images can be observed in Figure 2.

3.5 Multi-Synthesis dynamic Weighting

Our model leverages multiple losses to reconstruct images or discriminate pixels, thus defining a multi-objective problem. Typically, the total loss, denoted as $\mathcal{L}_T$, is a linear combination of different losses, denoted as $\mathcal{L}_i$, in the following manner:

$$\mathcal{L}_T = \sum_i \alpha_i \mathcal{L}_i + \mathbb{R}(\alpha) \tag{1}$$

where α denotes a set of weights and $R(\cdot)$ imposes some regularization on these weights. The loss function commonly assumes equal weights for individual loss terms, but varying the weights can considerably affect the model's performance. Adjusting the weights through grid search for optimal α values is computationally expensive and still static. To address this challenge, we implemented a dynamic weighting strategy, inspired by the coefficient of variations weighting method[6]. This strategy involves assigning weights, denoted as α, to four types of different loss terms, and these weights evolve as a function of the number of epochs (t) for each specific synthesis method. In this context, we denote the synthesis images as X_{Si}, with i representing one of the four synthesis methods under consideration. Furthermore, we use X_{Ri} to denote the corresponding reconstructed images, as illustrated in Figure 2. The loss function employed in our framework encompasses L1 Smooth, SSIM[28], Focal Loss[13], and Dice Loss[17]. Additional details are presented in Equation (2):

$$\begin{aligned} Multi\_Loss(t)_{S_i} = \; & \alpha_1(t) \cdot Loss_{L1smooth}(X_{Si}, X_{Ri}) \quad + \\ & \alpha_2(t) \cdot Loss_{1-SSIM}(X_{Si}, X_{Ri}) \quad + \\ & \alpha_3(t) \cdot Loss_{Focal}(PredictedMask_{S_i}, GT_{Si}) \quad + \\ & \alpha_4(t) \cdot Loss_{Dice}(PredictedMask_{S_i}, GT_{Si}) \end{aligned} \tag{2}$$

where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$.

Our model also applies [6] for calculating the weight of multiple anomaly synthesis methods, named Multi-Synthesis dynamic Weighing (MSdW). The output of each synthesis method from the *Multi_Loss(t)* quotation is treated as an input for the *Synthesis_Loss(t)* quotation. During the training process, the weight β is assigned to each synthesis method based on the number of epochs (t), thereby eliminating the need for an additional optimization process. Additional details are presented in Equation (3):

$$\begin{aligned} Synthesis\_Loss(t)_{Total} = \; & \beta_1(t) \cdot Multi\_Loss_{s_1} \quad + \\ & \beta_2(t) \cdot Mutli\_Loss_{s_2} \quad + \\ & \beta_3(t) \cdot Mutli\_Loss_{s_3} \quad + \\ & \beta_4(t) \cdot Mutli\_Loss_{s_4} \end{aligned} \tag{3}$$

where $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$.

4 Experiments

Our MACoW model is extensively evaluated and compared with the recent unsupervised SOTA methods. Additionally, individual components of the proposed method and the effectiveness of training on simulated anomalies are evaluated with an ablation study. Finally, our findings demonstrate superior performance when compared to other SOTA methods.

4.1 Implementation Details

MVTecAD dataset. It was introduced in [1], and is widely adopted as a standard benchmark for evaluating the effectiveness of anomaly detection algorithms in industrial inspection images, including 15 categories, comprising ten object categories and five texture categories. The BTAD dataset [18] is also a real-world industrial anomaly dataset consisting of three industrial products. The KSDD2 dataset [3] was also developed using images of defected production items and comprised 356 images with visible defects and 2,979 without defects. In cases where the training data contained defective samples, the aforementioned datasets were processed by removing the abnormal samples entirely, leaving only the **normal data** for our unsupervised training.

Evaluation Metrics. The evaluation of outcomes in prior research is based on the area under the receiver operating characteristic curve at both the image level (Image-AUROC) and pixel level (Pixel-AUROC). However, it has been observed that abnormal regions occupy only a tiny proportion of the entire image. Therefore, the Pixel-AUROC metric needs to reflect localization accuracy precisely. Furthermore, many non-anomalous pixels primarily influence the false positive rate, leading to low false favorable detection rates. To obtain a comprehensive measurement of localization performance, we also employ the Per Region Overlap (PRO) score[2] and pixel-level Average Precision (Pixel-AP)[32, 34] as the evaluation metrics. The PRO score provides equal consideration to anomaly regions of varied sizes[5]. In contrast, Pixel-AP is more appropriate for assessing highly imbalanced classes, especially in industrial anomaly localization, where accuracy plays a critical role[32].

4.2 Anomaly Detection and Localization on MVTecAD

Table 1 presents the results of anomaly detection and localization on the MVTecAD dataset. Our proposed method outperforms other SOTA techniques in terms of image AUROC (detection) in **7** out of 15 classes. Specifically, our method achieves the best average pixel AUROC performance **98.2%** compared to the unsupervised SOTA method. Table 2 presents the results of PRO and Pixel AP on the MVTec dataset. Our method achieves the highest PRO in **9** out of 15 classes. The average PRO results show that our method outperforms unsupervised SOTA by **1.5%**. Specifically, our proposed method demonstrates superior average pixel AUROC (localization) performance compared to other SOTA models for both Texture and Object. And Total average pixel AP performance **73.4%** compared to the unsupervised SOTA method and outperforms unsupervised SOTA by **2.3%**. This confirms the effectiveness of our approach in simultaneously localizing anomalous regions of varying sizes. Moreover, our method demonstrates excellent anomaly localization capability on the more challenging AP metrics. These results also demonstrate the discriminative power of our proposed approach in differentiating between normal and abnormal pixels, thereby improving the AP metric's performance. It also proved that our method can leverage all kinds of anomaly synthesis algorithms to reach the best performance for unsupervised reconstruction-based methods.

Category	**DRÆM**				**SSPCAB**				**RD**				**Patchcore**				**Ours**			
	I	P	O	A	I	P	O	A	I	P	O	A	I	P	O	A	I	P	O	A
BTAD 01	98.5	91.5	61.4	17.0	96.2	92.4	62.8	18.1	**98.8**	95.7	72.8	**49.3**	96.6	**96.5**	**78.4**	47.1	93.0	93.7	71.2	46.8
BTAD 02	68.6	73.4	39.0	23.3	69.3	65.6	28.6	15.8	**84.9**	96.0	55.8	66.1	81.3	94.9	54.0	56.3	81.7	**96.9**	**66.5**	**70.6**
BTAD 03	99.8	96.3	84.3	17.2	99.4	92.4	71.0	5.0	99.5	99.0	**98.8**	45.1	**99.9**	99.2	96.4	51.2	96.9	**99.7**	98.5	**72.3**
BTAD Average	89.0	87.1	61.6	19.2	88.3	83.5	54.1	13.0	**94.4**	**96.9**	75.8	53.5	92.6	**96.9**	76.3	51.5	90.5	96.8	**78.8**	**63.2**
KSDD2	81.1	85.6	67.9	39.1	83.4	86.2	66.1	44.5	**96.0**	**97.6**	**94.7**	43.5	76.5	97.1	88.8	64.1	91.1	96.8	66.5	**86.9**

Table 3: BTAD and KSDD2 performance comparison with SOTA models. "I", "P", "O", and "A" respectively refer to the four metrics of Image auroc, Pixel auroc, PRO, and Pixel AP. The best score is highlighted in bold. The results for DRÆM, SSPCAB, RD, and PatchCore are reported from [36]

Category	**CutPaste[11]**	**DRÆM**	**SSPCAB**	**RD**	**NSA[23]**	**DSR[34]**	**Patchcore**	**Ours**
Carpet	93.9 / 98.3	96.9 / 97.5	93.1 / 92.6	98.7 / 98.9	95.6 / 95.5	**100.0** / 95.5	99.1 / 99.0	99.6 / **99.4**
Grid	**100.0** / 97.5	99.9 / 99.7	99.7 / 99.5	**100.0** / 98.3	99.9 / 99.2	**100.0** / 99.6	97.3 / 98.7	**100.0 / 99.8**
Leather	**100.0** / 99.5	**100.0** / 99.0	98.7 / 96.3	**100.0** / 99.4	99.9 / 99.5	**100.0 / 99.6**	**100.0** / 99.3	**100.0 / 99.6**
Tile	94.6 / 90.5	**100.0** / 99.2	**100.0** / 99.4	99.7 / 95.7	**100.0** / 99.3	**100.0** / 98.2	99.3 / 95.8	**100.0 / 99.5**
Wood	99.1 / 95.5	99.5 / 95.5	98.4 / **96.5**	99.5 / 95.8	97.5 / 90.7	96.3 / 92.5	**99.6** / 95.1	96.8 / 96.2
Average	97.5 / 96.3	99.3 / 98.2	98.0 / 96.9	**99.6** / 97.6	98.6 / 96.8	99.3 / 97.1	99.1 / 97.6	99.3 / **98.9**
Bottle	98.2 / 97.6	98.0 / 99.1	95.6 / **99.2**	**100.0** / 98.8	97.7 / 98.3	**100.0** / 98.9	**100.0** / 98.6	**100.0** / 98.4
Cable	81.2 / 90.0	90.9 / 95.2	92.7 / 95.1	96.1 / 97.2	94.5 / 96.0	93.8 / 96.7	**99.9 / 98.5**	95.7 / 94.4
Capsule	98.2 / 97.4	91.3 / 88.1	96.9 / 90.2	96.1 / 98.7	95.2 / 97.6	98.1 / 95.4	98.0 / 99.0	**99.0 / 99.1**
Hazelnut	98.3 / 97.3	**100.0 / 99.7**	**100.0 / 99.7**	**100.0** / 99.0	94.7 / 97.6	95.6 / 99.2	**100.0** / 98.7	98.9 / 99.5
Metal Nut	99.9 / 93.1	**100.0 / 99.6**	**100.0** / 99.4	**100.0** / 97.3	98.7 / 98.4	98.5 / 93.7	99.9 / 98.3	**100.0** / 98.9
Pill	94.9 / 95.7	97.1 / 97.3	97.4 / 97.2	98.7 / 98.1	**99.2 / 98.5**	97.5 / 93.4	97.5 / 97.6	97.2 / 97.3
Screw	88.7 / 96.7	**98.7** / 99.3	97.8 / 99.0	97.8 / **99.7**	90.2 / 96.5	96.2 / 98.5	98.2 / 99.5	97.5 / 99.3
Toothbrush	99.4 / 98.1	**100.0** / 97.3	97.9 / 97.3	**100.0** / 99.1	**100.0** / 94.9	99.7 / **99.5**	**100.0** / 98.6	99.2 / 99.3
Transistor	96.1 / 93.0	91.7 / 85.2	88.0 / 84.8	95.5 / 92.3	95.1 / 88.0	97.8 / 83.2	**99.9 / 96.5**	96.3 / 92.8
Zipper	99.9 / 99.3	**100.0** / 99.1	**100.0** / 98.4	97.9 / 98.3	99.8 / 94.2	**100.0** / 98.9	99.5 / 98.9	**100.0 / 99.4**
Average	95.5 / 95.8	96.8 / 96.0	96.6 / 96.0	98.2 / 97.9	96.5 / 96.0	97.7 / 95.7	**99.3 / 98.4**	98.4 / 97.8
TotalAverage	96.1 / 96.0	97.6 / 96.7	97.1 / 96.3	98.7 / 97.8	97.2 / 96.3	98.2 / 96.2	**99.2** / 98.1	98.7 / **98.2**

Table 1: MVTecAD performance comparison with SOTA models, given by Image AUROC / Pixel AUROC. The best score per row is highlighted in bold. The results for DRÆM, SSPCAB, RD and PatchCore are reported from [36]

4.3 Ablation Experiment

4.3.1 Comparison of SSPCBMA and SSPCAB

We evaluate the performance of the proposed SSPCBMA method on the MVTecAD dataset. The primary aim of these experiments is to compare the performance of three conditions: without any attention; SSPCAB[21], and our novel SSPCBMA. The experimental results are presented in Table 4, demonstrating that our SSPCBMA module achieved the highest score on PRO, and performed competitively with the SOTA method on Pixel AP.

	Image AUC		Pixel AUC		Pixel AP		PRO	
	Texture	Object	Texture	Object	Texture	Object	Texture	Object
w/*o* Attention	97.7	99.0	97.7	97.9	**70.6**	77.3	93.4	96.8
SSPCAB	98.1	99.0	97.6	**98.9**	69.7	**79.2**	93.4	97.3
SSPCBMA(Ours)	**98.4**	**99.3**	**97.8**	**98.9**	70.5	79.0	**93.8**	**98.6**

Table 4: MVTecAD performance on different Attention Modules. The best score per row is highlighted in bold. The results for SSPCAB are reported from [36]

4.3.2 Comparison of Multi-Synthesis dynamic Weighting(MSdW) and Multi-Loss dynamic Weighting

The performance of our proposed Multi-Synthesis dynamic Weighting (MSdW) approach is evaluated on the MVTecAD dataset, which includes three distinct loss weight configurations: static weights featuring equal or hand-tuned weighting; Multi-Loss dynamic Weighting method [21]; and our novel MSdW method. The experimental findings are presented in Table 5, and the results indicate that our proposed method outperforms the other methods in all indicators. This is attributed to our ability to consider the impact of synthetic data and leverage various self-supervised synthesis methods during the training process.

Category	**CutPaste[11]**	**DRÆM**	**SSPCAB**	**RD**	**NSA[23]**	**DSR[34]**	**Patchcore**	**Ours**
Carpet	50.4 / -	92.9 / 65.1	86.4 / 48.6	95.4 / 56.5	85.0 / -	- / 78.2	95.5 / 62.2	**99.8 / 80.6**
Grid	91.5 / -	98.3 / 62.8	98.0 / 57.9	94.2 / 15.8	96.8 / -	- / 68.0	94.0 / 24.5	**99.1 / 77.0**
Leather	83.7 / -	97.4 / 72.9	94.0 / 60.7	98.2 / 47.6	98.7 / -	- / 62.5	96.9 / 45.3	**99.2 / 68.3**
Tile	54.4 / -	98.2 / **95.2**	98.1 / 96.1	85.6 / 54.1	95.3 / -	- / 93.9	91.3 / 56.2	**98.3** / 94.3
Wood	64.0 / -	90.3 / 74.6	92.8 / **78.9**	91.4 / 48.3	85.3 / -	- / 68.4	87.1 / 49.3	**96.7** / 74.9
Average	68.8 / -	95.4 / 74.1	93.9 / 68.4	93.0 / 44.5	92.2 / -	- / 74.2	93.0 / 47.5	**98.6 / 79.0**
Bottle	91.2 / -	**96.8** / 88.9	96.3 / **89.4**	96.3 / 78.0	92.9 / -	- / 91.5	95.4 / 76.8	**96.8** / 86.5
Cable	59.8 / -	81.0 / 56.4	80.4 / 52.0	94.1 / 52.6	89.9 / -	- / **70.4**	**96.8** / 67.0	95.2 / 66.8
Capsule	83.5 / -	82.7 / 39.6	92.5 / 46.4	95.5 / 47.2	91.4 / -	- / 53.3	93.4 / 46.0	**95.9 / 57.6**
Hazelnut	81.3 / -	**98.5** / 92.6	98.2 / **93.4**	96.9 / 60.7	93.6 / -	- / 87.3	90.9 / 53.2	96.9 / 87.4
Metal Nut	54.4 / -	97.0 / **97.0**	**97.7** / 94.7	94.9 / 78.6	94.6 / -	- / 67.5	92.6 / 86.6	96.9 / 89.3
Pill	83.1 / -	88.4 / 47.6	89.6 / 48.3	**96.7 / 76.5**	96.0 / -	- / 65.7	94.5 / 75.7	93.6 / 68.8
Screw	72.6 / -	95.0 / **66.5**	95.2 / 61.7	98.5 / 52.1	90.1 / -	- / 52.5	97.5 / 34.7	**99.4** / 54.4
Toothbrush	88.1 / -	85.6 / 45.5	85.5 / 39.3	92.3 / 51.1	90.7 / -	- / **74.2**	**94.0** / 37.9	92.6 / 62.5
Transistor	68.5 / -	70.4 / 39.0	62.5 / 38.1	83.3 / 54.1	75.3 / -	- / 41.1	**92.3 / 66.9**	72.4 / 49.2
Zipper	84.9 / -	96.8 / 77.6	95.2 / 76.4	95.3 / 57.5	89.2 / -	- / 78.5	96.1 / 62.3	**98.9 / 83.0**
Average	76.7 / -	89.2 / 65.1	89.3 / 64.0	**94.4** / 60.8	90.4 / -	- / 68.2	**94.4** / 60.7	93.4 / **70.6**
TotalAverage	74.1 / -	91.3 / 68.1	90.8 / 65.5	93.9 / 55.4	91.0 / -	- / 70.2	93.9 / 56.3	**95.4 / 73.4**

Table 2: MVTecAD performance comparison with SOTA models, given by PRO / Pixel AP. The best score per row is highlighted in bold. The results for DRÆM, SSPCAB, RD and PatchCore are reported from [36]

	Image AUC		Pixel AUC		Pixel AP		PRO	
	Texture	Object	Texture	Object	Texture	Object	Texture	Object
Static Weighting(Fixed)	97.2	98.2	97.1	98.7	67.5	77.6	83.8	96.6
Multi-Loss dynamic Weighting[5]	97.0	98.2	97.1	**98.9**	68.2	77.8	92.7	97.2
Multi-Synthesis dynamic Weighting(Ours)	**98.4**	**99.3**	**97.8**	**98.9**	**70.5**	**79.0**	**93.8**	**98.6**

Table 5: MVTecAD performance on different weighting strategies. The best score per column is highlighted in bold.

Table 6 illustrates that the combination of both MSdW and SSPCBMA components yielded the best results in our study. Specifically, we achieved the highest performance metrics for Image AUROC, PRO, and pixel AP, while Pixel AUROC was also close to SOTA performance.

ConvNeXUnetV2		Performance			
MSdW	SSPCBMA	I	P	O	A
		97.7	**98.9**	94.8	70.7
✓		98.1	97.8	94.6	73
	✓	97.5	97.6	88.1	70.9
✓	✓	**98.7**	98.2	**95.4**	**73.4**

Table 6: Components capability on MVTecAD. "I", "P", "O", and "A" respectively refer to the four metrics of Image auroc, Pixel auroc, PRO, and Pixel AP. The best score per column is highlighted in bold.

5 Conclusions

This paper presents a novel anomaly detection and segmentation approach, Masked Attention ConvNeXtUNetV2 with Multi-Synthesis Weighting (MACoW). Our methodology harnesses various anomaly simulation strategies applied to anomaly-free samples to generate synthetic anomaly images for model training. Through comprehensive experimentation, we have demonstrated the efficacy of the introduced SSPCBMA attention mechanism within the reconstructive subnetwork, resulting in improved feature extraction capabilities for anomaly detection. Furthermore, our study showcases the benefits of integrating multiple self-supervised learning approaches, which effectively regularize our reconstructive subnetwork and subsequently enhance the overall performance in anomaly detection and segmentation tasks. The MACoW model delivers competitive results across benchmark datasets, including MVTec, BTAD, and KSDD2. Of particular note is the versatility of our approach, as the Multi-Synthesis Dynamic Weighting algorithm enables easy integration of new synthesis methods on image datasets. This adaptability allows our model to leverage the advantages of emerging self-supervised synthesis techniques, positioning it for continued anomaly detection and segmentation research advancements.

References

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

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[3] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

[4] Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. A survey on unsupervised industrial anomaly detection algorithms. *arXiv preprint arXiv:2204.11161*, 2022.

[5] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, June 2022.

[6] Rick Groenendijk, Sezer Karaoglu, Theo Gevers, and Thomas Mensink. Multi-loss weighting with coefficient of variations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1469–1478, 2021.

[7] Yi Hao, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognition*, 121:108232, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Chaoqin Huang, Fei Ye, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection, 2020.

[10] Jielin Jiang, Jiale Zhu, Muhammad Bilal, Yan Cui, Neeraj Kumar, Ruihan Dou, Feng Su, and Xiaolong Xu. Masked swin transformer unet for industrial anomaly detection. *IEEE Transactions on Industrial Informatics*, 19(2):2200–2209, 2022.

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

[12] Yufei Liang, Jiangning Zhang, Shiwei Zhao, Runze Wu, Yong Liu, and Shuwen Pan. Omni-frequency channel-selection representations for unsupervised anomaly detection. *arXiv preprint arXiv:2203.00259*, 2022.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[14] Tongkun Liu, Bing Li, Zhuo Zhao, Xiao Du, Bingke Jiang, and Leqi Geng. Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection. *arXiv preprint arXiv:2210.14485*, 2022.

[15] Neelu Madan, Nicolae-Catalin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *arXiv preprint arXiv:2209.12148*, 2022.

[16] Emilie Mathian, Huidong Liu, Lynnette Fernandez-Cuesta, Dimitris Samaras, Matthieu Foll, and Liming Chen. Haloae: An halonet based local transformer auto-encoder for anomaly detection and localization. *arXiv preprint arXiv:2208.03486*, 2022.

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[18] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021.

[19] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

[20] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6726–6733. IEEE, 2021.

[21] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13576–13586, 2022.

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

[23] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. 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*, pages 474–489. Springer, 2022.

[24] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023.

[25] Yong Shi, Jie Yang, and Zhiquan Qi. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 424:9–22, 2021.

[26] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 2022.

[27] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images, 2020.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[29] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[30] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023.

[31] Xuan Xia, Xizhou Pan, Nan Li, Xing He, Lin Ma, Xiaoguang Zhang, and Ning Ding. Gan-based anomaly detection: a review. *Neurocomputing*, 2022.

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

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

[34] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr – a dual subspace re-projection network for surface anomaly detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, page 539–554, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19820-5. doi: 10.1007/978-3-031-19821-2_31. URL https://doi.org/10.1007/978-3-031-19821-2_31.

[36] Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16281–16291, 2023.

[37] Jiachi Zhang, Xiaolei Shen, Tianqi Zhuo, and Hong Zhou. Brain tumor segmentation based on refined fully convolutional neural networks with a hierarchical dice loss. *arXiv preprint arXiv:1712.09093*, 2017.

[38] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders, 2019.