A Proof of the One-directed Adaptive loss function

In this chapter, we will explain our own loss function. First, we analyze why Binary Cross Entropy (BCE) is inadequate for our situation. What we're trying to achieve serves as a clear motivation for a newly constructed loss function. Then, using the properties of a function whose exponent is a positive rational number less than 1, a new loss function is defined. In the last part of this chapter, the derivative of this loss function and the sign of the derivative are mathematically considered, to ensure that the total loss function actually decreases during the learning process. For simplicity in this Appendix, we will use q and p to represent q_t and p_t respectively, without loss of generality. This notation will be used consistently throughout the following proofs and explanations.

A.1 Motivation for Proposing One-directed Adaptive loss

A.1.1 Analysis to Binary Cross Entropy

We will first examine a brief analysis of the BCE. The loss function is constructed as follows:

$$\mathcal{L}_{\text{cluster}} = -\sum p \log q + (1-p)\log(1-q) \tag{1}$$

Before calculating p by equation (4) in main text using one-directed threshold, assume that the threshold is fixed as 0.5 in the loss function. Then, p is determined by the following rule:

$$p = \begin{cases} 0, & 0 \le q < 0.5 \\ 1, & 0.5 \le q \le 1 \end{cases} \tag{2}$$

So the loss function is calculated by different functions depending on which interval the value of q belongs to. In the BCE, the total interval [0, 1] for the available value of q is divided by a threshold, which is 0.5, into two different intervals: [0, 0.5) and [0.5, 1]. To simplify the analysis, let's consider a function where the variable q is on the x-axis and the value inside the logarithm is on the y-axis. Then we can reconstruct the original BCE into:

$$y = \begin{cases} 1-q, & 0 \le q < 0.5 \\ q, & 0.5 \le q \le 1 \end{cases} \tag{3}$$

Figure 1 shows the value inside the logarithm in the BCE loss function. To reduce the total loss, the value inside the logarithm must be increased.

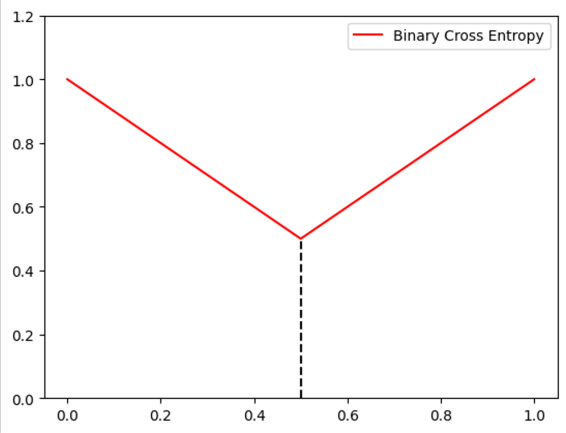

Figure 1: The black dashed line represents the position of the threshold that acts on the value q to classify whether the label is 0 or 1.

Therefore, the closer the value of y is to 1, the smaller the total loss. The distribution of q can therefore be classified into two different labels. One will be located in the neighborhood of 0 and the other will be located in the neighborhood of 1. However, this approach poses a problem in anomaly detection tasks using single clustering, particularly when training only on normal data. The issue arises because the BCE loss function allows normal data to be correctly classified whether it's close to 0 or 1. We typically want normal data to cluster towards one direction - either 0 or 1, not both. The learning process should encourage normal data to converge towards a single value (either 0 or 1), rather than allowing it to be distributed at both extremes.

A.1.2 Desired Goals

What we are aiming for requires two differences from the original loss function. The first one is that the threshold must be learned, and the threshold must increase as it is learned. And second, the distribution of q should only be close to 1, not to 0, during the learning process. If the threshold is denoted by ν, we will take a monotonic function such that the overall graph should approach $y = 1$ as the value of ν increases as a value part of the logarithm of a new loss function.

A.2 The One-directed Adaptive Loss function Modeling

At first, the total interval $[0, 1]$ in which all possible q values is divided into $[0, \nu)$ and $[\nu, 1]$. Then the value p is determined as follows:

$$p = \begin{cases} 0, & 0 \leq q < \nu \\ 1, & \nu \leq q \leq 1 \end{cases} \tag{4}$$

To avoid the situation where the loss function is not defined, assume that the possible ν is in the range $0 < \nu < 1$. The simplest monotonic function connecting two points $(0, 0)$ and $(1, 1)$ is of the form $y = q^n$. For n which satisfies the inequality $0 < n < 1$, the functions $y = q^n$ are close to $y = 1$ as n decreases. So consider the following function to match the increasing trend of ν with the decreasing trend of n:

$$y = q^{1-\nu} \tag{5}$$

Figure 2 shows the graphs of the above function with different values of ν between 0 and 1. As ν increases, it can be seen that starting from $y = x$ and approaching $y = 1$ rapidly. This effect is more pronounced at lower values of q.

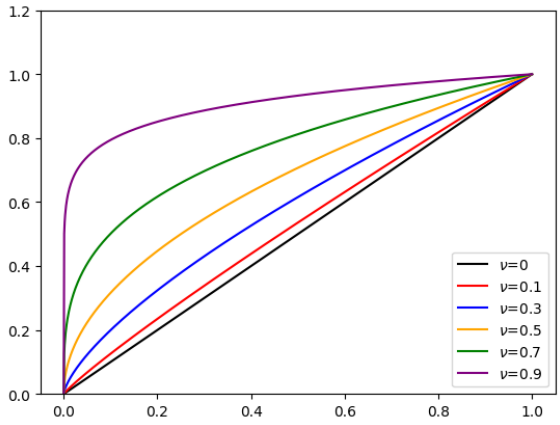

Figure 2: The graph of $y = q^{1-\nu}$ with different values of ν between 0 and 1.

Qualitatively, this function is rapidly increasing to 1 for small q when ν is increasing. So we adopt the function $q^{1-\nu}$ in the interval $[0, \nu)$ as the value inside the logarithm of the loss function. Meanwhile, in the interval $[\nu, 1]$, we define the function as a linear function connecting two points $(\nu, \nu^{1-\nu})$ and $(1, 1)$, ensuring the continuity of the entire function over the interval $[0, 1]$ and reflecting the simplest form.

$$y = \frac{1-\nu^{1-\nu}}{1-\nu}(q-\nu) + \nu^{1-\nu} = \frac{1-\nu^{1-\nu}}{1-\nu}(q-1) + 1 \tag{6}$$

In summary, we adopt the following function as the value inside the logarithm of our new loss function.

$$y = \begin{cases} q^{1-\nu}, & 0 \le q < \nu \\ \frac{1-\nu^{1-\nu}}{1-\nu}(q-1) + 1, & \nu \le q \le 1 \end{cases} \tag{7}$$

Corresponding graphs with different ν are shown in Figure 3. Each colored dashed line indicates the position of the threshold at different values of ν. Before the threshold, the function is concave; after it, the function is linear.

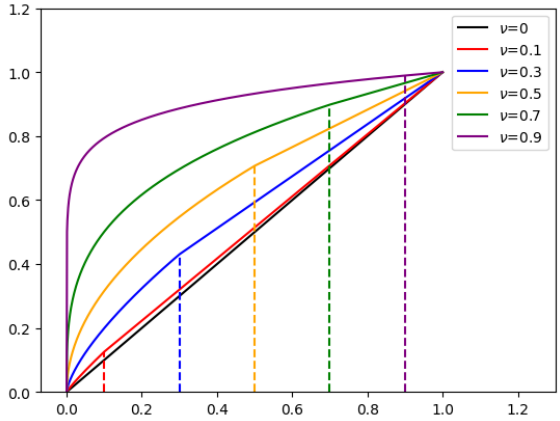

Figure 3: The graph of our new loss function with different values of ν between 0 and 1.

Thus, the final loss function can be expressed as follows:

$$\mathcal{L}_{\text{cluster}} = -\sum p \log\left(\frac{1-\nu^{1-\nu}}{1-\nu}(q-1) + 1\right) + (1-p)\log\left(q^{1-\nu}\right) \tag{8}$$

A.3 Derivative of loss function

In order to mathematically confirm that the new loss function really decreases when q and ν are increasing, to simplify the derivative procedure, let us define f_1 and f_2 as:

$$f_1 \equiv \frac{1-\nu^{1-\nu}}{1-\nu}(q-\nu) + \nu^{1-\nu} = \frac{1-\nu^{1-\nu}}{1-\nu}(q-1) + 1, \qquad f_2 \equiv q^{1-\nu} \tag{9}$$

Since both f_1 and f_2 satisfy the conditions for a valid logarithm argument, f_1 and f_2 are positive in the entire interval $[0, 1]$. The derivative of total loss $\mathcal{L}_{\text{cluster}}$ with respect to q and ν can be expressed as:

$$\frac{\partial \mathcal{L}_{\text{cluster}}}{\partial q} = -p\frac{1}{f_1}\frac{\partial f_1}{\partial q} - (1-p)\frac{1}{f_2}\frac{\partial f_2}{\partial q}, \qquad \frac{\partial \mathcal{L}_{\text{cluster}}}{\partial \nu} = -p\frac{1}{f_1}\frac{\partial f_1}{\partial \nu} - (1-p)\frac{1}{f_2}\frac{\partial f_2}{\partial \nu}. \tag{10}$$

A.3.1 $\partial \mathcal{L}_{\text{CLUSTER}}/\partial q$

Since both f_1 and f_2 are positive, we need to verify the signs of $\partial f_1/\partial q$ and $\partial f_2/\partial q$. Let's consider the derivative of f_1 with respect to q first:

$$\frac{\partial f_1}{\partial q} = \frac{1-\nu^{1-\nu}}{1-\nu} \tag{11}$$

The condition $0 < \nu < 1$ implies $0 < \nu^{1-\nu} < 1$. Therefore, both the denominator and the numerator are positive, ensuring that $\partial f_1/\partial q > 0$ is satisfied. Meanwhile, the derivative of f_2 with respect to q can be written as:

$$\frac{\partial f_2}{\partial q} = (1-\nu)q^{-\nu} = \frac{1-\nu}{q^{\nu}} \tag{12}$$

Similarly, because $0 < \nu < 1$ and $0 < q < 1$, both the denominator and the numerator are also positive, so $\partial f_2/\partial q > 0$ is satisfied. Thus, we can determine the sign of the derivative of our new loss function with respect to q:

$$\frac{\partial \mathcal{L}_{\text{cluster}}}{\partial q} < 0 \tag{13}$$

This means that the total loss $\mathcal{L}_{\text{cluster}}$ decreases as q increases.

A.3.2 $\partial \mathcal{L}_{\text{CLUSTER}}/\partial \nu$

This part is very similar to proving the sign of $\partial \mathcal{L}_{\text{cluster}}/\partial q$, but it requires a more technical procedure. The derivative of total loss $\mathcal{L}_{\text{cluster}}$ with respect to ν can be written as follows:

$$\frac{\partial \mathcal{L}_{\text{cluster}}}{\partial \nu} = -p\frac{1}{f_1}\frac{\partial f_1}{\partial \nu} - (1-p)\frac{1}{f_2}\frac{\partial f_2}{\partial \nu} \tag{14}$$

Since both f_1 and f_2 are positive, we need to verify the signs of $\partial f_1/\partial \nu$ and $\partial f_2/\partial \nu$. Let's consider the derivative of f_1 with respect to ν first:

$$\begin{aligned}
\frac{\partial f_1}{\partial \nu} &= \frac{(q-1)}{(1-\nu)^2}\left[-\nu^{1-\nu}\left(\frac{1-\nu}{\nu} - \log\nu\right)(1-\nu) + (1-\nu^{1-\nu})\right] \\
&= \frac{(q-1)}{(1-\nu)^2}\left[1 + \nu^{1-\nu}\left(-\frac{(1-\nu)^2}{\nu} + (1-\nu)\log\nu - 1\right)\right] \\
&= \frac{(q-1)}{(1-\nu)^2\nu^{\nu}}\left\{\nu^{\nu} + \nu - \nu^2 - 1 + \nu(1-\nu)\log\nu\right\}
\end{aligned} \tag{15}$$

We have a condition for c and q, which is $0 < \nu < 1$ and $0 < q < 1$. The outermost factor satisfies the following inequality:

$$\frac{(q-1)}{(1-\nu)^2\nu^\nu} < 0 \tag{16}$$

Let us define g_1, g_2, g_3 as:

$$\begin{cases} g_1 = \nu^\nu + \nu \\ g_2 = \nu^2 + 1 \\ g_3 = \nu(1-\nu)\log\nu \end{cases} \tag{17}$$

To express the formula inside the braces as $g_1 - g_2 + g_3$, we will confirm the sign of each function for $\nu \in (0, 1)$, thereby justifying the sign of the formula inside the braces. g_3 satisfies $g_3 < 0$ because of two inequalities:

$$\log\nu < 0, \qquad \nu(1-\nu) > 0 \tag{18}$$

From the limit $\lim_{\nu\to 0+} \nu^\nu = 1$, we can obtain the values of g_1 and g_2 at $\nu = 1$ and the left-side limit values of g_1 and g_2:

$$\begin{cases} g_1(0+) = g_2(0+) = 1 \\ g_1(1) = g_2(1) = 2 \end{cases} \tag{19}$$

The derivative of g_1 with respect to ν is:

$$\frac{\partial g_1}{\partial \nu} = \nu^\nu(1 + \log\nu) + 1 \tag{20}$$

Here, the first term $\nu^\nu(1 + \log\nu)$ is negative when $\nu \in (0, e^{-1})$, while it is positive due to the factor $(1 + \log\nu)$ when $\nu \in (e^{-1}, 1)$. Consequently, the function $g_1 - \nu$ decreases in the interval $(0, e^{-1})$ and increases in the interval $(e^{-1}, 1)$. Additionally, the first term $\nu^\nu(1 + \log\nu)$ diverges to $-\infty$ as ν approaches 0 from the positive side. While the interval of increase or decrease might differ by adding the constant 1 to the first term, the overall trend of g_1 remains the same even when considering $g_1 - \nu$. The derivative of g_2 with respect to ν is:

$$\frac{\partial g_2}{\partial \nu} = 2\nu \tag{21}$$

This quantity is always positive if $\nu \in (0, 1)$, so the function g_2 increases in the interval $(0, 1)$. Therefore, in the interval $(0, 1)$, the function g_1 is always smaller than the function g_2; $g_1 - g_2 < 0$. This means that the formula $g_1 - g_2 + g_3$ satisfies the following inequality where $\nu \in (0, 1)$:

$$g_1 - g_2 + g_3 < 0 \tag{22}$$

The graph of $g_1 - g_2 + g_3$ represents negative values in the interval $(0, 1)$, as shown in Figure 4.

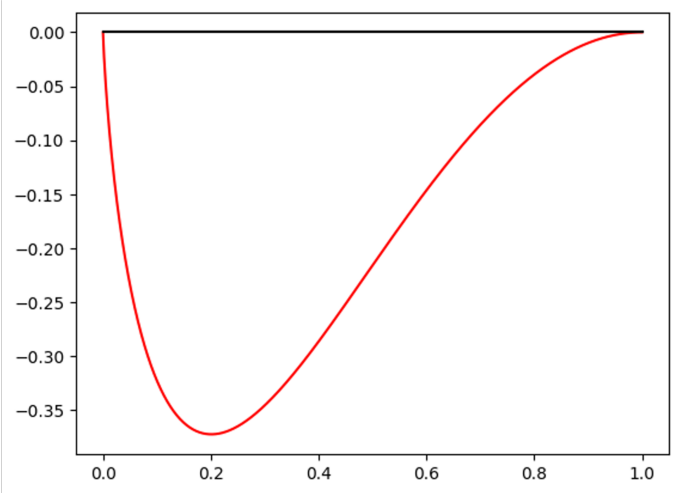

Figure 4: The graph of $g_1 - g_2 + g_3$ in the interval $[0, 1]$. The black line represents the x-axis; values below this line indicate that the function is negative.

Therefore, the sign of the derivative of f_1 with respect to ν is positive, so $\partial f_1/\partial \nu > 0$. On the other hand, for $\partial f_2/\partial \nu$, we have:

$$\frac{\partial f_2}{\partial \nu} = -q^{1-\nu} \log q \tag{23}$$

The value of $q^{1-\nu}$ is between 0 and 1, and $\log q < 0$, so $\partial f_2/\partial \nu > 0$. Thus, we can determine the sign of the derivative of our new loss function with respect to ν:

$$\frac{\partial \mathcal{L}_{\text{cluster}}}{\partial \nu} < 0 \tag{24}$$

This means that the total loss $\mathcal{L}_{\text{cluster}}$ decreases as ν increases.

B MULTI-CLUSTER ($k > 1$) FOR MADCLUSTER

MADCluster employs cosine similarity with a One-directed Adaptive loss function, initially assuming a single cluster ($k = 1$). This design overcomes the trivial solution where the soft assignment of a student's t-distribution always yields a value of 1 when only one cluster is present. Whereas, with several modifications, MADCluster can be extended utilizing student's t-distribution to support multi-cluster based clustering ($k > 1$). The soft assignment q_{tj} and the target distribution p_{tj} represent the assignment of the t-th representation to the j-th cluster and is defined as:

$$q_{tj} = \frac{(1 + |h_t - \hat{c}_j|^2)^{-1}}{\sum_{j=1}(1 + |h_t - \hat{c}_j|^2)^{-1}}, \qquad p_{tj} = \frac{q_{tj}^2 / \sum_{t=1} q_{tj}}{\sum_{j=1}(q_{tj}^2 / \sum_{t=1} q_{tj})} \tag{25}$$

Sequence-wise Clustering loss $\mathcal{L}_{\text{cluster}}$ is calculated using the Kullback-Leibler (KL) divergence instead of the One-directed Adaptive loss. It is defined as follows:

$$\mathcal{L}_{\text{cluster}} = KL(P|Q) = \sum_{j=1}^{K} \sum_{t=1}^{T} p_{tj} \log \frac{p_{tj}}{q_{tj}} \tag{26}$$

And for the Cluster Distance Mapping loss $\mathcal{L}_{\text{distance}}$ we have adopted a simplified notation, omitting some details for clarity, is also defined as follows:

$$\mathcal{L}_{\text{distance}} = \frac{1}{n} \sum_{j=1}^{K} \sum_{t=1}^{T} \|h_t - \hat{c}_j\|^2 + \lambda \Omega(\mathcal{W}) \tag{27}$$

Consequently, during training, we sum two components for each time step t: the KL-divergence values across all clusters for the t-th representation, and the distances from the t-th representation to each cluster center. The anomaly score is also defined as follows:

$$\text{Anomaly Score}(x_t) = \sum_{j=1}^{K} p_{tj} \log \frac{p_{tj}}{q_{tj}} + \left\| h_t - c_j^* \right\|^2 \tag{28}$$

For the multi-cluster case, the anomaly score does not incorporate ν, and therefore ν is not learned. Similar to the single-cluster case, $\text{Anomaly Score}(x_t) \in \mathbb{R}^{T \times 1}$ serves as the point-wise anomaly score for $\mathcal{X}$.

Furthermore, we conducted experiments using multi-cluster with k=1,2,3,4,5,6,7,8,9,10. The experimental results for multi-cluster, which utilize the modified equation, are presented in Table 1.

Table 1: Results of evaluating MADCluster performance on four real-world datasets with multi-cluster ($k = 1$ to 10).

Dataset	**MSL**			**SMAP**			**SMD**			**PSM**		
# Clusters	**F1**	**Aff-P**	**Aff-R**	**F1**	**Aff-P**	**Aff-R**	**F1**	**Aff-P**	**Aff-R**	**F1**	**Aff-P**	**Aff-R**
1	**0.52**	**0.74**	**0.99**	**0.49**	**0.71**	0.96	**0.48**	**0.83**	0.90	**0.67**	**0.68**	**1.00**
2	0.47	0.66	0.98	0.36	0.64	**0.99**	0.17	0.57	0.59	0.50	0.55	**1.00**
3	0.48	0.71	0.98	0.33	0.61	**0.99**	0.24	0.58	**0.95**	0.51	0.55	**1.00**
4	0.47	0.69	0.98	0.35	0.60	**0.99**	0.24	0.61	0.92	0.50	0.55	**1.00**
5	0.47	0.68	**0.99**	0.34	0.62	0.98	0.22	0.59	0.85	0.51	0.56	**1.00**
6	0.49	0.68	0.98	0.34	0.58	**0.99**	0.23	0.57	0.92	0.54	0.57	0.99
7	0.49	0.68	**0.99**	0.39	0.65	**0.99**	0.28	0.61	0.88	0.50	0.55	**1.00**
8	0.47	0.69	0.97	0.37	0.66	0.98	0.30	0.62	0.84	0.50	0.55	**1.00**
9	0.48	0.72	0.97	0.38	0.64	0.98	0.25	0.59	0.92	0.50	0.55	**1.00**
10	0.51	0.71	0.98	0.37	0.65	**0.99**	0.27	0.63	0.81	0.62	0.60	0.69

Overall, across the benchmark datasets—MSL, SMAP, SMD, and PSM—the performance patterns do not consistently improve or degrade with increasing numbers of clusters. In particular, the best performance in terms of F1 and Aff-P is often observed when using a single cluster, while in the SMD and SMAP datasets, Aff-R tends to remain high across multiple clusters, although the difference from the single-cluster case is not significant. Unlike what might be expected, the number of clusters does not show a clear monotonic relationship with detection performance. That is, increasing the number of clusters does not necessarily improve anomaly detection performance. Instead, the most stable and strong performance is frequently achieved with just a single cluster.

This suggests that a single-cluster approach in MADCluster is not only sufficient to model the distribution of normal patterns across diverse time-series datasets but actually be optimal in many cases. The robustness of the MADCluster aligns with the design intent of the proposed One-directed Adaptive loss, which proves most effective when applied in this setting.

C Results after applying MADCluster to baseline models

C.1 Computational Efficiency

Table 2 lists the computational costs and validation accuracy, with all models trained on the MSL dataset. When applying MADCluster, performance significantly improves without substantially impacting structural complexity or efficiency. This integration results in only a slight increase in computational demands, as measured by MACs (KMac units), with a modest increase in parameter size. By maintaining a balance between efficiency and performance, this method enhances the anomaly detection capabilities of existing models without imposing significant changes. This demonstrates the effectiveness and adaptability of MADCluster, indicating its potential to improve existing anomaly detection techniques while balancing computational demands and performance enhancement.

Table 2: Computational Efficiency and metrics(F1, V_ROC, V_PR) Comparison on the MSL Dataset, detailing the number of parameters ('# Params') indicating model size and Multiply-Accumulate Computations ('MACs') reflecting processing speed.

Model	MACs	#Params	F1	V_ROC	V_PR
DeepSVDD	31.81M	311.55K	0.37	0.63	0.28
+ MADCluster	31.81M	311.62K	**0.52**	**0.72**	**0.42**
USAD	427.36M	256.26M	**0.53**	0.71	**0.43**
+ MADCluster	427.36M	256.26M	**0.53**	**0.72**	**0.43**
BeatGAN	10.22G	185.85M	0.49	0.70	0.39
+ MADCluster	10.22G	185.85M	**0.50**	**0.71**	**0.43**
OmniAnomaly	35.44M	350.72K	0.42	0.64	0.31
+ MADCluster	35.44M	350.83K	**0.45**	**0.67**	**0.34**
THOC	69.42M	390.78K	0.50	0.71	0.41
+ MADCluster	69.42M	390.91K	**0.55**	**0.72**	**0.46**
AnomalyTransformer	485.23M	4.86M	0.50	0.70	**0.40**
+ MADCluster	485.23M	4.86M	**0.51**	**0.71**	**0.40**
DCdetector	1.189G	912.18K	0.28	0.52	0.19
+ MADCluster	1.189G	912.30K	**0.41**	**0.64**	**0.32**

C.2 Impact of Clustering and Distance Mapping on Anomaly Detection Performance

In Table 3 we evaluated the performance of the anomaly detection approaches illustrated in maintext Figure 2. This table presents quantitative results of our proposed method, which learns center coordinates and performs single clustering as we hypothesized. DeepSVDD represents only distance mapping, while Clustering denotes the experimental results using self-labeling without distance mapping. MADCluster, our proposed method, applies both distance mapping and clustering.

Table 3: Performance comparison of anomaly detection approaches across four datasets: (1) DeepSVDD (Cluster Distance Mapping), (2) Clustering (Sequence-wise Clustering), and (3) MADCluster (Combined Cluster Distance Mapping and Sequence-wise Clustering)

Dataset	MSL			SMAP			SMD			PSM		
Metric	F1	Aff-P	Aff-R	F1	Aff-P	Aff-R	F1	Aff-P	Aff-R	F1	Aff-P	Aff-R
DeepSVDD	0.37	0.63	0.99	0.40	0.70	**0.99**	0.19	0.56	0.61	0.50	0.55	**1.00**
Clustering	0.30	0.70	0.66	0.39	0.70	0.95	0.20	0.58	0.61	0.19	0.56	0.23
MADCluster	0.52	**0.74**	**0.99**	**0.49**	**0.71**	0.96	**0.48**	**0.83**	**0.90**	**0.65**	**0.66**	0.62

Evaluation of original F1, Aff-P, and Aff-R using Cluster Distance Mapping and Sequence-wise Clustering individually did not reveal a consistently dominant method across datasets. In contrast,

MADCluster, which integrates both approaches, achieved the highest performance in all metrics except Aff-R on the SMAP and PSM datasets.

D Extension of MADCluster to Image Data Domains

To evaluate the applicability of MADCluster beyond time-series data, experiments were conducted on image anomaly detection tasks. In the original implementation, the extracted dynamics through the Base Embedder are represented as `[batch, sequence, hidden_dim]`. For image inputs typically structured as `[batch_size, channels, height, width]`, the spatial dimensions were flattened to form a sequence-like representation of shape `[batch_size, channels, height × width]`. This transformation enables the application of MADCluster sequence-wise clustering and cluster distance mapping mechanisms to spatial data.

Experiments were performed on the MVTec AD dataset, a widely used benchmark for unsupervised image anomaly detection. MADCluster was integrated with three representative models: Reverse Distillation for Anomaly Detection (RD4AD), PyramidFlow, and RealNet. All models were trained for 10 epochs, and performance was evaluated using the Area Under the Receiver Operating Characteristic (AUROC) at both the image and pixel levels. The results are summarized in Table 4 and Table 5, demonstrating that MADCluster can be effectively extended to image domains with minimal architectural modifications.

Table 4: Image-level AUROC (%) on the MVTec AD dataset with and without MADCluster.

Class Name	RealNet	+MADCluster	RD4AD	+MADCluster	PyramidFlow	+MADCluster
Bottle	0.918	**0.961**	0.991	**0.998**	0.778	**0.993**
Cable	0.631	**0.669**	0.945	**0.946**	0.638	**0.695**
Capsule	0.694	**0.698**	0.868	**0.870**	0.870	**0.916**
Carpet	0.969	**0.977**	0.996	**0.997**	0.938	**0.964**
Grid	0.872	**0.875**	0.921	**0.945**	0.794	**0.824**
Hazelnut	0.972	**0.994**	**1.000**	**1.000**	0.930	**0.935**
Leather	0.806	**0.830**	**1.000**	**1.000**	0.993	**0.999**
Metal Nut	0.670	**0.688**	0.995	**0.996**	0.735	**0.742**
Pill	0.823	**0.844**	0.936	**0.956**	0.810	**0.834**
Screw	0.552	**0.572**	0.829	**0.848**	0.595	**0.752**
Tile	0.972	**0.981**	0.993	**0.994**	0.994	**0.995**
Toothbrush	0.553	**0.644**	0.997	**1.000**	0.944	**0.947**
Transistor	0.659	**0.660**	0.967	**0.970**	0.908	**0.936**
Wood	0.959	**0.966**	0.990	**0.993**	0.991	**0.996**
Zipper	0.882	**0.901**	0.871	**0.889**	**0.938**	**0.938**

Table 5: Pixel-level AUROC (%) on the MVTec AD dataset with and without MADCluster.

Class Name	RealNet	+MADCluster	RD4AD	+MADCluster	PyramidFlow	+MADCluster
Bottle	0.949	**0.963**	0.982	**0.986**	0.960	**0.974**
Cable	0.631	**0.897**	**0.977**	**0.977**	0.895	**0.912**
Capsule	0.901	**0.927**	0.981	**0.982**	0.977	**0.980**
Carpet	0.970	**0.984**	**0.992**	**0.992**	0.964	**0.978**
Grid	0.873	**0.894**	0.942	**0.963**	0.941	**0.948**
Hazelnut	0.925	**0.956**	**0.991**	**0.991**	0.964	**0.973**
Leather	0.968	**0.971**	**0.994**	**0.994**	0.985	**0.987**
Metal Nut	0.754	**0.770**	0.969	**0.974**	0.938	**0.959**
Pill	0.942	**0.943**	0.967	**0.968**	0.943	**0.956**
Screw	0.929	**0.946**	0.985	**0.986**	0.898	**0.903**
Tile	0.930	**0.937**	**0.953**	**0.953**	0.962	**0.973**
Toothbrush	0.918	**0.924**	0.987	**0.988**	0.975	**0.977**
Transistor	0.704	**0.722**	**0.890**	**0.890**	0.965	**0.972**
Wood	0.930	**0.932**	**0.955**	**0.955**	0.957	**0.960**
Zipper	0.951	**0.962**	0.968	**0.970**	**0.968**	**0.968**

E DATASET

We summarize the four adopted benchmark datasets for evaluation in Table 6. These datasets include multivariate time series scenarios with different types and anomaly ratios. MSL, SWaT, SMD and PSM are multivariate time series datasets.

Table 6: Statistics and details of the benchmark datasets used. AR (anomaly ratio) represents the abnormal proportion of the whole dataset.

Benchmarks	Applications	Dim	Win	#Train	#Test	AR (Truth)
MSL	Space	55	100	58,317	73,729	0.105
SMAP	Space	25	100	135,183	427,617	0.128
SMD	Server	38	100	708,405	708,420	0.042
PSM	Server	25	100	132,481	87,841	0.278