UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection

Shun Wei[1] Jielin Jiang[1,2,3]* Xiaolong Xu[1,2,3]
[1]School of Software, Nanjing University of Information Science and Technology
[2]State Key Laboratory for Novel Software Technology, Nanjing University
[3]Jiangsu Province Engineering Research Center of Advanced Computing and Intelligent Services
{pangdatangtt,jiangjielin2008,xlxu}@nuist.edu.cn[1]

Abstract

Anomaly detection (AD) is a crucial visual task aimed at recognizing abnormal pattern within samples. However, most existing AD methods suffer from limited generalizability, as they are primarily designed for domain-specific applications, such as industrial scenarios, and often perform poorly when applied to other domains. This challenge largely stems from the inherent discrepancies in features across domains. To bridge this domain gap, we introduce UniNet, a generic unified framework that incorporates effective feature selection and contrastive learning-guided anomaly discrimination. UniNet comprises student-teacher models and a bottleneck, featuring several vital innovations: First, we propose domain-related feature selection, where the student is guided to select and focus on representative features from the teacher with domain-relevant priors, while restoring them effectively. Second, a similarity contrastive loss function is developed to strengthen the correlations among homogeneous features. Meanwhile, a margin loss function is proposed to enforce the separation between the similarities of abnormality and normality, effectively improving the model's ability to discriminate anomalies. Third, we propose a weighted decision mechanism for dynamically evaluating the anomaly score to achieve robust AD. Large-scale experiments on 11 datasets from various domains show that UniNet surpasses existing methods[1].

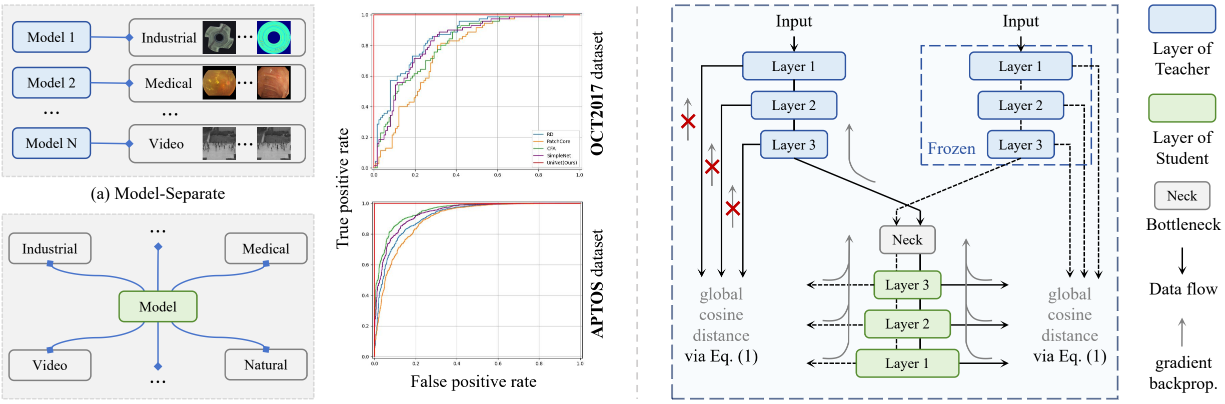
1. Introduction

Visual anomaly detection (AD) has gained significant traction in recent years, with applications spanning across various fields, such as medical image diagnosis [7, 37, 48], industrial defect inspection [9, 25, 33, 50], and video surveillance [1, 39, 44]. Prior AD paradigms typically develop separate models tailored to each domain (see Fig. 1(a)). Despite considerable advancements in domain-specific applications, these approaches often suffer from limited cross-domain applicability. This limitation primarily arises from domain differences and inherent discrepancies in features. For instance, in industrial AD, some self-supervised methods [29, 43, 55] employ external data [10] or data augmentation technologies to synthesize anomalies and learn anomalous feature distribution. However, the anomalies generated in this manner can differ substantially from those encountered in other domains, *e.g.*, medical imaging [11, 26, 45] or video surveillance [30], potentially resulting in insufficient learning of the anomalous distribution. In fact, beyond the stark differences in visual appearance of anomalies–such as defects on industrial products *vs.* polyps on the intestine or anomalous behavior like cyclist in video surveillance–there are also notable differences in the normal features across different domains. This variability further complicates their cross-domain applications. Moreover, another main challenge hindering the effective application of most methods across other domains is their reliance on pre-trained networks–trained on source domains such as ImageNet [13]–for feature extraction. Recent studies [17, 23, 57] have demonstrated that pre-trained features often bear little resemblance to those needed to the target domain owing to inherent biases in these pre-trained networks, adversely affecting performance (see Fig. 1(c)). In light of these challenges, this paper explores the problem of how to develop a unified framework capable of adapting to diverse domains while achieving accurate AD (see Fig. 1(b)).

Recently, ReContrast [17] introduces contrastive learning (CL) elements to optimize its framework for adaptation to different target domains, showing good transfer ability. Nevertheless, two limitations restrict its further development. First, it struggles to capture representative features relevant to the target domain, which impacts its ability to understand domain-related information. Second, it still

*Corresponding author.

[1] *https://github.com/pangdaTangtt/UniNet*



(a) Model-Separate

(b) Unified Model

(c) AUROC curve comparison

(d) Architecture of ReContrast

Figure 1. (a) One-model-one-domain setting. (b) One-model multi-domain setting. (c) AUROC curve comparison of UniNet and competing AD methods (reliance on pre-trained features) on medical datasets. (d) Architecture of ReContrast.

faces a significant challenge in effectively discriminating between abnormality and normality, even after being trained on some anomalous samples, which limits its applicability in supervised settings [2, 6, 24, 46].

In respond to these problems, we propose a novel generic unified framework based on ReContrast, termed UniNet. It consists of student-teacher (S-T) models along with a bottleneck. Concretely, UniNet first develop a lightweight-yet-powerful multi-scale embedding module (MEM) within the bottleneck to better capture the contextual relationships among features provided to the student. We then propose domain-related feature selection, a method that prompts the student to select crucial features from the teacher with prior knowledge to learn domain-related information. To effectively distinguish anomalies, a similarity-contrastive loss is first proposed to strengthen the correlations among homogeneous features. Followed this, a margin loss is developed to enhance the similarity of normal features, ensuring they are separated from anomalous ones with low similarity. Finally, considering the similarity between the outputs of S-T networks, we propose a weighted decision mechanism to adaptively calculate the anomaly score for improved AD performance. Notably, unlike ReContrast [17] that mainly focuses on unsupervised AD, our UniNet can be suited for unsupervised and supervised settings simultaneously. In summary, our contributions are as follows:

- This paper presents UniNet, a generic unified framework that can be oriented towards a wider range of domains, applicable to both unsupervised and supervised settings.
- We design MEM to capture contextual information and propose domain-related feature selection to guide the student in selecting and learning target-oriented representative features from the teacher.
- A similarity-contrastive loss is developed to enhance the relationships among homogeneous features and we then employ a margin loss to enhance the similarity of normal features for better anomaly discrimination. We implement a weighted decision mechanism to achieve superior AD performance during inference.
- Large-scale experiments conducted on 11 datasets from industrial, medical and video domains manifest that UniNet achieves superior results across different metrics.

2. Related work

Unsupervised methods. Unsupervised AD methods rely solely on available anomaly-free samples to learn their distribution due to the scarcity of anomalous data. Consequently, numerous promising methods [17, 25, 33, 39, 41, 47, 50, 58] have been continuously proposed. AST [41] employs an asymmetric S-T framework, minimizing the distance between their outputs to identify anomalies with large deviations. THF [25] proposes a new flow-based method that prevents the overlap of distribution between normal and anomalous features. Other attempts explore the use of memory banks to store additional normal prototypes to effectively detect anomalies, such as PatchCore [40] and MemKD [16]. The aforementioned methods are typically classified as one-class AD methods, as they train a separate model for each class. Recently, some efforts [17, 18, 53] have shifted toward multi-class AD, aiming to use one unified model to detect anomalies across different classes concurrently. UniAD [53] pioneers this approach, solving the problem that a growing number of training categories often leads to increased computational time. MambaAD [18] further advances this idea by exploring state space models, achieving outstanding performance while maintaining low complexity and computational overheads.

Supervised methods. Unlike unsupervised AD methods, supervised AD approaches can train a model on anomalous samples, thereby improving the accuracy of class boundaries [42]. DevNet [38] utilizes some labeled anomalous samples and prior probabilities to enforce that the anomaly scores of the anomalous samples significantly deviate from those of the normal samples in the upper tail. FCDD [34] employs a fully convolutional neural network architecture to map normal samples towards the center of the feature space, effectively distancing anomalous samples from this central region. DRA [14] generalizes to unknown anomalies by learning disentangled anomalous representations for different types of anomalies. Due to the scarcity of available supervised datasets, these methods are primarily trained on the widely used MVTec AD dataset [4], where the labeled anomalous samples are generally derived from the test set. However, they still face challenge in having adequate anomalous samples. Consequently, they often employ anomaly synthesis strategy [29, 43, 55] to generate anomalous samples, but these anomalies hardly conform to the real-world anomaly distribution. To tackle the issue of limited supervised datasets, Baitieva et al. [2] recently introduced a new industrial supervised AD benchmark, which features a wider array of complex anomalies and substantial intra-class variability among anomalous-free images. Given that conventional AD methods struggle with this benchmark, they further incorporated a segmentation-based anomaly detector to enhance AD performance.

In this paper, our work seeks to develop a unified solution for AD across different domains, in contrast to most current mainstream methods, which design separate models for domain-specific AD.

3. Preliminaries

The prototype of ReContrast [17] is composed of a pre-trained teacher model, a bottleneck, and a learnable student model. ReContrast aims to optimize the entire framework to the target domain through CL elements. Let $F_T^i, F_S^i \in \mathbb{R}^{C_i \times H_i \times W_i}$ respectively represent the output of i^{th} layer of the teacher and student models, where C_i, H_i, and W_i are the channels, height, width of the corresponding output. To optimize the student model, ReContrast first proposes global cosine distance to better maintain global consistency between feature points and avoid instability during training:

$$\mathcal{L}_g = \sum_{i=1}^{n} d(F_T^i, F_S^i) = \sum_{i=1}^{n} 1 - \frac{\mathcal{F}(F_T^i)^\top}{\|\mathcal{F}(F_T^i)\|} \cdot \frac{\mathcal{F}(F_S^i)}{\|\mathcal{F}(F_S^i)\|}, \quad (1)$$

where n represents the number of layers, $d(:,:)$ is the cosine distance, $\|\cdot\|$ is ℓ_2 norm, and $\mathcal{F}(\cdot)$:$\mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{CHW}$ denotes a flattening operation. Subsequently, ReContrast optimizes the pre-trained teacher model to adapt it to the target domain. However, prior works suggested that this would result in pattern collapse. Inspired by CL for self-supervised learning [8], the stop gradient operation is introduced to mitigate pattern collapse by modifying Eq. (1):

$$\mathcal{L}_g = \sum_{i=1}^{n} 1 - \frac{\mathrm{SG}(\mathcal{F}(F_T^i))^\top}{\|\mathrm{SG}(\mathcal{F}(F_T^i))\|} \cdot \frac{\mathcal{F}(F_S^i)}{\|\mathcal{F}(F_S^i)\|}, \quad (2)$$

where $\mathrm{SG}(\cdot)$ is the stop gradient operation. To prevent "identical shortcut" caused by no contrastive pairs, ReContrast introduces an additional frozen teacher model without any optimization during training. In this way, two teacher models can produce two views from one image, *i.e.*, a target domain view and a source domain view, to achieve image augmentations, similar to a CL paradigm (see Fig. 1(d)).

4. Methodology

4.1. Approach overview

With inspiration from ReContrast [17], this paper proposes UniNet, a generic unified framework for different domains, as illustrated in Fig. 2. The goal of UniNet is to optimize the entire framework towards the target domain, while enabling domain-relevant feature selection and learning, along with effective anomaly discrimination.

To capture the contextual relationships among features, UniNet first develops a lightweight-yet-effective MEM within the bottleneck (Sec. 4.2). Then, we propose domain-related feature selection, guiding the student to select target-oriented features from the teacher with prior knowledge and prompting its learning (Sec. 4.3). Besides, a similarity-contrastive loss is proposed to enhance the correlations among homogeneous features, followed by the development of a margin loss to preserve the distinction between the similarities of normal and anomalous features, thereby enhancing discriminability (Sec. 4.4). Based on the similarity between the outputs of S-T network, a weighted decision mechanism is proposed to achieve robust AD performance during inference (Sec. 4.5).

4.2. Multi-Scale Embedding Module

Motivation. Some prior approaches [12, 25] struggle to capture the contextual relationships among features, impeding enhancement in feature correlations and redundancy reduction. These methods typically employ a set of small kernels to mitigate increased computational overheads, but recent research [15] has demonstrated that fittingly using a few larger kernels can be helpful for vision tasks. Propelled by this insight, we design a simple yet powerful Multi-Scale Embedding Module (MEM) within the bottleneck for feature extraction across various contexts while maintaining low memory consumption, as visualized in Fig. 2.

Module design. Considering the multi-scale features, we first split the input as two parts along the channel dimensions, with two different size of kernels to capture global and local information, thus enriching the contextual relationships among features. These two parts are respectively

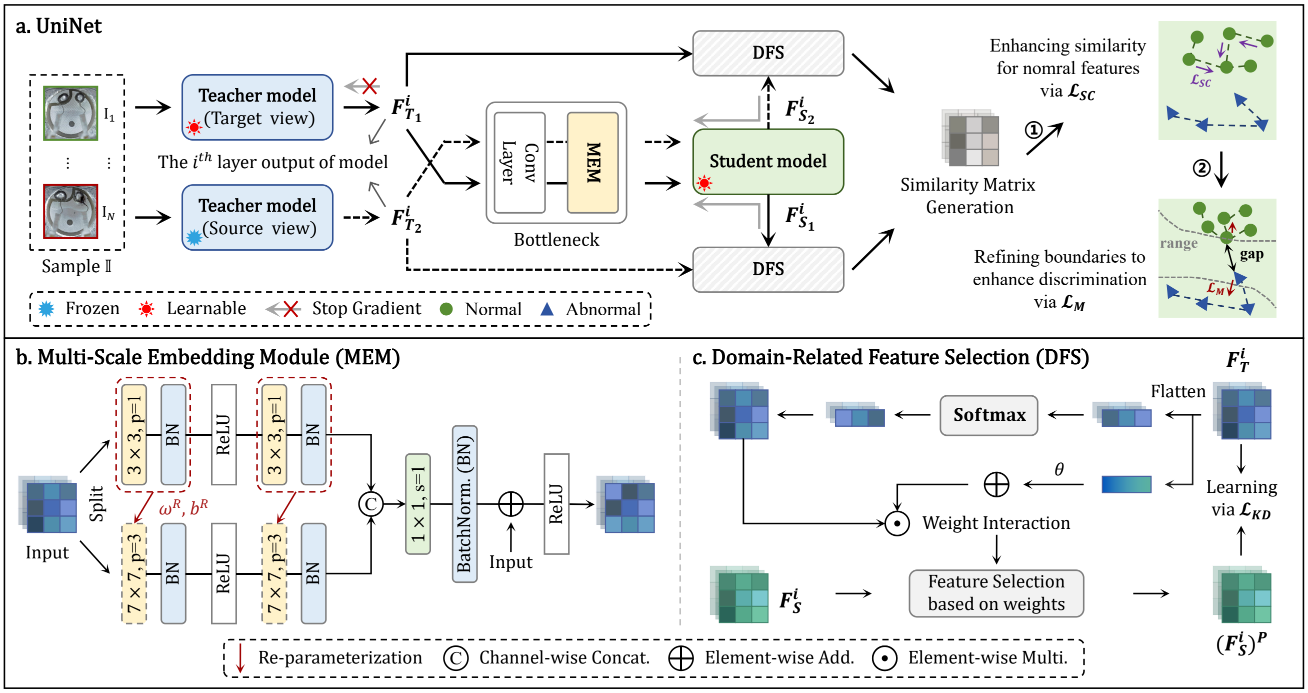


Figure 2. Overall framework of the proposed UniNet. It consists of a pair of teachers models, a bottleneck, and a student model, with several key components: MEM, DFS, Similarity-Contrastive loss $\mathcal{L}_{SC}$ and Margin loss $\mathcal{L}_M$.

fed into a $k \times k$ (where k is 3 or 7) kernel convolution layer, followed by a batch normalization (BN) layer and a ReLU activation (ReLU). Similarly, they are further fed into a $k\times k$ kernel convolution layer and a BN layer to enhance feature extraction. Then, these two parts are concatenated to compress channel dimensions by a 1×1 kernel convolution layer and a BN layer. Finally, to achieve better regularization, a residual connection is conducted before ReLU.

Due to the use of large kernels, the number of parameters and inference time would increase. To mitigate this, we reparameterized the large kernel convolutional layers through the small kernel convolutional layers and BN layers:

$$\omega^R = \omega \cdot \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}}, b^R = b - \frac{\mu \cdot \gamma}{\sqrt{\sigma^2 + \epsilon}}, \tag{3}$$

where ω is the weight of small kernel convolutional layer, while ω^R and b^R are the weight and bias of large kernel convolutional layer after re-parameterization. μ, σ^2, γ, and b denote the mean, variance, weight, and bias of BN layers, respectively. ϵ is a small constant. In this way, 7×7 kernel convolution layers can be viewed as equivalent to 3×3 kernel convolution layers, thus decreasing computational costs.

4.3. Domain-Related Feature Selection

Motivation. Despite its great transfer ability across different domains, ReContrast [17] still suffer from an insufficient capture of feature representations relevant to the target domain, resulting in the loss of crucial information. To solve this challenge, we propose Domain-Related Feature Selection (DFS), which encourages the student to selectively concentrate on target-oriented features from the teacher and well restore them, thereby avoiding the inclusion of unimportant information. Particularly, the student is required to learn only representative information pertaining to the target domain, rather than all available information.

Selection and learning. Introducing information from target-oriented domain into the student is crucial for enabling it to understand and generate the feature representations required for that domain. To achieve it, we utilize the weight to control how much representative information should be selected from the teacher with prior knowledge. Concretely, let the features from the teacher and student models be $F^i_T, F^i_S \in \mathbb{R}^{C_i\times H_i\times W_i}$, and the teacher feature is flattened to $\widehat{F}^i_T \in \mathbb{R}^{C_i\times H_i W_i}$. The weight can be generated as follows:

$$w_i = \frac{\exp(\widehat{F}^i_T - \varrho)}{\sum_{j=1}^{H_i W_i} \exp(\widehat{F}^i_T(:, j) - \varrho)}, \tag{4}$$

where $w_i \in \mathbb{R}^{C_i\times H_i W_i}$ and $\varrho = \max(\widehat{F}^i_T)$. To avoid relying on local information only, we integrate global information to ensure weight interaction, enhancing aware ability.

The global information F_i^g can be obtained as follows:

$$F_i^g = \frac{1}{H_i \cdot W_i} \sum_{h=1}^{H_i} \sum_{w=1}^{W_i} F_T^i(:, h, w), \tag{5}$$

where $F_i^g \in \mathbb{R}^{C_i \times 1 \times 1}$. Let the weight w_i be further reshaped to $(w_i)^R \in \mathbb{R}^{C_i \times H_i \times W_i}$. The interacted weight can be obtained through the fusion of global and local information, and the student can then select domain-related feature information based on the interacted weight to generate domain-related features $(F_{S_i})^P$:

$$(F_S^i)^P = F_S^i \odot \{(w_i)^R \odot (\theta + F_i^g)\}, \tag{6}$$

where θ is a learnable parameter flexibly controlling the domain-related feature selection and $\odot$ represents element-wise multiplication. To ensure that the student effectively learns the prior knowledge, we aim to minimize the distance between its output and that of the teacher during training:

$$\mathcal{L}_{KD} = \sum_{i=1}^{n} d(\mathrm{SG}(F_T^i), (F_S^i)^P). \tag{7}$$

4.4. Comparing similarity and enhancing discrimination

Motivation. Most unsupervised methods [12, 17, 33, 40] may fail to establish class boundaries [38] due to the absence of anomalous samples. However, even when trained with some anomalous samples, these methods still face challenge in effectively discriminating anomalies, particularly unseen ones [14]. To overcome this limitation, we propose a similarity-contrastive loss and a margin loss. The similarity-contrastive loss is first employed to enhance the correlations among normal features, ensuring they remain tightly clustered. Yet, since anomalous features are not expected to exhibit high similarity to normal ones, this encourages a clear gap between them. Thus, we then use the margin loss to enforce a greater separation between their similarity, further enhancing the model's discrimination ability.

Mechanism. The learned student feature $(F_S^i)^P$ is flattened and reshaped to $(\widehat{F}_S^i)^P \in \mathbb{R}^{H_i W_i \times C_i}$. The similarity matrix m_i between the outputs of S-T models is obtained:

$$m_i = \frac{(\widehat{F}_S^i)^P \cdot (\widehat{F}_T^i)}{\left\|(\widehat{F}_S^i)^P\right\| \cdot \left\|\widehat{F}_T^i\right\| \cdot \mathcal{T}}, \tag{8}$$

where $\mathcal{T}$ is a temperature parameter which controls the distribution of similarity. With the similarity matrix, the normalized similarity matrix can be then obtained by:

$$\widehat{m}_i = \frac{\exp(m_i)}{\sum_{k=1}^{H_i W_i} \exp(m_i(:, k)) + \epsilon}. \tag{9}$$

Each diagonal element of $\widehat{m}_i$ is the similarity between the features of S-T pairs. To strengthen the relationships among normal features within these pairs, we employ the similarity-contrastive loss to maximize their similarity:

$$\mathcal{L}_{SC} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N_1} \log(\mathrm{diag}((\widehat{m}_i)_j) + \epsilon), \tag{10}$$

where $N_1 (N_1 \leq N)$ is the number of normal samples and $\mathrm{diag}(\cdot)$ is the operation of selecting diagonal elements.

After maximizing feature similarity, we further use the margin loss to improve the separability between features. It ensures higher similarity for normal features, which helps mitigate the issue of class boundaries. When trained with anomalous samples, it further constrains the similarity of anomalous features below a range, thereby improving discriminability. The margin loss is defined as follows:

$$\begin{aligned}\mathcal{L}_M = \frac{1}{n} \sum_{i=1}^{n} \Bigg\{ & \sum_{j=1}^{N_1} \max\left(0, (\tau - \mathrm{diag}\left((\widehat{m}_i)_j\right))\right) + \\ & \sum_{k=1}^{N-N_1} \max(0, (\mathrm{diag}((\widehat{m}_i)_k) - \frac{\tau}{2})) \Bigg\},\end{aligned} \tag{11}$$

where τ is a hyper-parameter controlling the boundary. In this way, homogeneous features can be grouped together.

During training, the overall losses are measured as:

$$\mathcal{L}_U = \lambda \mathcal{L}_{KD} + (1 - \lambda) \mathcal{L}_{SC} + \mathcal{L}_M, \tag{12}$$

where λ is a balancing hyper-parameter. In supervised settings, Eq.(12) can be modified based on its task:

$$\mathcal{L} = \mathcal{L}_U + \sum_{i=1}^{n-1} \mathcal{L}_S(\Phi(F_S^i), l), \tag{13}$$

where $\mathcal{L}_S(:,:)$ denotes a Binary Cross-Entropy loss or a Dice loss, $\Phi(\cdot)$ is a flattening or upsampling operation, and l is a label or a ground-truth mask.

4.5. Anomaly detection

Motivation. Some efforts [22, 38] use top K or top-ranked values from the anomaly map to evaluate the anomaly score. Nevertheless, they rely on a fixed value (*e.g.*, $K = 3\%$) for score calculation, which does not ensure robust AD. To remedy this, we propose a weighted decision mechanism that dynamically calculates image-level anomaly score for each sample, where anomaly score is determined by weight.

Weighted decision mechanism. Formally, we begin with obtaining the pixel-level anomaly map for each sample. With a pair of teacher models and one student model, two anomaly maps can be generated from each layer by $d(\cdot)$ in Eq. (1), with in total of $2n$ anomaly maps. Then,

(a) MVTec AD

Method	I-AUROC	P-AUROC	PRO
RD++ [47]	99.44	98.25	94.99
DMAD [32]	99.50	98.21	–
GLAD [52]	99.30	98.62	95.31
ReConPacth [23]	99.56	98.18	–
RealNet [57]	99.65	99.03	93.07
ReContrast [17]	99.46	98.41	95.20
UniNet(Ours)	99.90	98.81	96.00

(b) BTAD

Method	I-AUROC	P-AUROC
PatchCore [40]	93.13	97.27
RD++ [47]	95.63	97.41
PyramidFlow [28]	95.83	97.70
ReConPacth [23]	95.80	97.47
RealNet [57]	96.07	97.90
ReContrast [17]	95.06	97.50
UniNet(Ours)	97.73	97.70

(c) MVTec 3D-AD

Method	I-AUROC	PRO
PatchCore [40]	81.14	91.03
AST [41]	88.00	–
M3DM [50]	85.03	94.22
Shape-Guided [9]	81.51	93.30
BTF [20]	78.52	87.63
ReContrast [17]	88.63	95.20
UniNet(Ours)	95.76	95.55

Table 1. Quantitative results on industrial datasets, including (a) MVTec AD, (b) BTAD, and (c) MVTec 3D-AD. We report the Image-level AUROC (I-AUROC), Pixel-level AUROC (P-AUROC), and PRO. Best and second-best results are highlighted in red and blue, respectively.

we obtain $2n$ low similarity values by taking the maximum value in each anomaly map. These low similarity values are further transformed into a probability distribution $\mathcal{N}$ via a Softmax activation. The values from $\mathcal{N}$ higher than the average of $\mathcal{N}$ are added to a *set* $\mathcal{P}$ to dynamically calculate the weight. The weight is defined as follows:

$$v_w = \max\{\alpha(\frac{1}{L}\sum_{p_j\in\mathcal{P}}^{L} p_j), \beta\} \tag{14}$$

where $\mathcal{P} = \{p_j\}_{j=1}^{L}$ contains L high probability values and $L \in [n-1, n+1]$. α controls its upper limit and β decides the lower limit. After obtaining weight, $2n$ anomaly maps are upsampled and accumulated to form the final pixel-level anomaly map M_{AS}. Finally, our method adaptively selects the largest v_w values from M_{AS} and averages them to obtain the final image-level anomaly score for AD. More details can be found in Appendix.

For the segmentation task, the anomaly map M_{AS} is employed, where higher values in it indicate the presence of anomalies at that corresponding positions.

5. Experiments

5.1. Experimental setup

Datasets. To demonstrate the superiority of UniNet, extensive experiments were conducted across 11 datasets from various domains, including industrial defect inspection, medical imaging analysis, and video surveillance. In industrial AD, we considered three unsupervised benchmarks (MVTec AD [4], BTAD [36], and MVTec AD-3D [5]) and one recently published supervised benchmark, VAD [2]. For medical diagnosis, we utilized three unsupervised datasets (APTOS [45], OCT2017 [26], and ISIC2018 [11]) alongside three supervised datasets (Kvasir [24], CVC-ClinicDB [6], and CVC-ColonDB [46]). In video surveillance, we considered one popular unsupervised dataset, Ped2 [30]. Further details can be found in Appendix.

Evaluation metrics. Following [2, 17, 37, 50, 51, 57], appropriate metrics are selected for each task. The image-level Area Under the Receiver Operating Characteristic Curve (AUROC) and Average precision (AP) are employed to evaluate AD performance. For anomaly segmentation, both the pixel-level AUROC and Per-Region Overlap (PRO) are used. In medical datasets, F1-score (F1) and accuracy (ACC) are used for AD, while Dice Similarity Coefficient (DSC) and mean Intersection over Union (mIoU) are applied for anomaly segmentation.

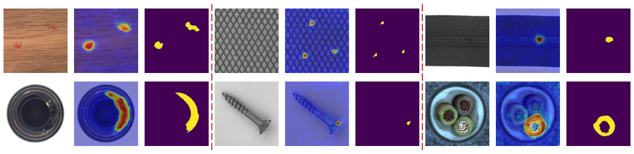


Figure 3. Qualitative results of UniNet on MVTec AD dataset. Each group from left to right: the anomalous images, our segmentation results, and ground-truths.

Implementation details. Following [12, 17], we used the publicly available WideResNet50 as backbone in S-T models. AdamW [35] was employed as the optimizer with a learning rate of 5e-3 and 1e-6 for the learnable student and teacher, respectively. The batch size was 8. All images were resized into 256 × 256. For MVTec 3D-AD dataset, only RGB data were used for training. Hyper-parameters n, $\mathcal{T}$, τ, λ, α, and β were set to 3, 2, 1, 0.7, 0.01, and 0.03, respectively. More details can be found in Appendix.

5.2. Comparison with state-of-the-art methods

We compared UniNet against the state-of-the-art (SOTA) methods on each dataset, selecting ReContrast [17] as baseline model. Further details can be found in Appendix.

5.2.1. Results under the unsupervised setting

MVTec AD. Five leading methods were considered: RealNet [57], ReConPatch [23], GLAD [52], DMAD [32], and RD++ [47]. The comparison results are presented in Table 1(a). UniNet achieves superior performance across multiple metrics, except for the pixel-level AUROC slightly lower than that of RealNet by 0.22%. UniNet yields the significant results in terms of both image-level AUROC and pixel-

(a) Medical anomaly detection

Dataset →	APTOS			OCT2017			ISIC2018		
Method ↓	I-AUROC	F1	ACC	I-AUROC	F1	ACC	I-AUROC	F1	ACC
RD4AD [12]	92.43	90.65	86.44	99.25	97.79	96.70	85.09	74.53	78.76
PatchCore [40]	90.45	90.18	85.57	99.61	98.34	97.50	78.94	68.57	71.50
AE-flow [58]	–	–	–	98.15	96.36	94.42	87.79	80.56	84.97
CFA [27]	94.21	94.39	92.03	98.01	96.40	94.70	81.31	72.31	74.61
SimpleNet [33]	93.42	91.16	87.27	98.50	96.91	95.40	82.17	69.82	73.59
ReContrast [17]	97.51	95.27	93.35	99.60	98.53	97.80	90.15	81.12	86.01
UniNet(Ours)	100.0	99.60	99.44	100.0	99.60	99.40	100.0	100.0	100.0

(b) Video anomaly detection

Ped2	
Method	I-AUROC
zxVAD [1]	96.9
SLM [44]	97.6
PDM-Net [21]	97.7
Ristea et al.[39]	95.4
AnomalyRuler [51]	97.9
ReContrast [17]	95.2
UniNet(Ours)	97.9

Table 2. Quantitative results on three medical datasets (APTOS, OCT2017, and ISIC2018) and one video dataset (VAD). We report the I-AUROC, F1, and ACC. Best and second-best results are highlighted in red and blue, respectively.

(a) Industrial anomaly detection

Dataset →	VAD		
Method ↓	I-AUROC	F1	ACC
DevNet [38]	87.00	80.05	84.53
DRA [14]	87.53	80.46	84.87
RD4AD [12]	87.40(90.23)	80.33	84.56
PatchCore [40]	88.52(91.70)	81.43	85.12
EfficientAD [3]	88.01(91.75)	81.55	85.31
ReContrast [17]	84.52(88.31)	74.03	78.23
UniNet(Ours)	99.95	98.60	98.60

(b) Medical anomaly detection

Dataset →	Kvasir		CVC-ClinicDB		CVC-ColonDB	
Method ↓	DSC	mIOU	DSC	mIOU	DSC	mIOU
KDAS [49]	91.3	84.8	92.5	87.2	91.2	83.3
HarDNet-CPS [54]	91.1	85.6	91.7	88.7	91.0	83.6
SAM-EG [48]	91.5	86.2	93.1	87.9	91.5	84.3
MEGANet [7]	91.3	86.3	93.8	89.4	–	–
MADGNet [37]	90.7	85.3	93.9	89.5	–	–
ReContrast [17]	87.3	80.5	90.3	84.0	87.3	79.2
UniNet(Ours)	91.5	85.7	94.2	89.5	91.9	85.6

Table 3. Quantitative results on three medical datasets (Kvasir, CVC-ClinicDB, and CVC-ColonDB) and one industrial dataset (VAD). For the medical datasets, we report the DSC and mIOS metrics, while for the VAD dataset, we provide the I-AUROC, F1, and ACC. The values in parentheses indicate results after using SegAD [2]. Best and second-best results are highlighted in red and blue, respectively.

level PRO compared to SOTA methods, particularly excelling in image-level AUROC with an average of **99.90**%. Our UniNet improves the baseline model by 0.44% and 0.40% AUROC, and 0.80% PRO. Besides, the anomalous regions segmented by UniNet are identical to ground-truths, as shown in Fig. 3. We also evaluated UniNet under the multi-class AD setting, with results provided in Appendix.

BTAD. We compared UniNet with five SOTA methods: ReConPatch [23], RealNet [57], PyramidFlow [28], RD++ [47], and PatchCore [40]. As shown in Table 1(b), UniNet surpasses all methods in terms of image-level AUROC, outperforming the recent two top methods, RealNet and ReConPatch, by a large margin of 1.66% and 1.93%, respectively. Meanwhile, UniNet achieves comparable segmentation performance compared to RealNet. Besides, UniNet improves the baseline by both 2.67% and 0.20% AUROC.

MTVec 3D-AD. We also evaluated UniNet on a more challenging 3D dataset and compared it with SOTA methods, including M3DM [50], AST [41], Shape-Guided [9], BTF [20], and PatchCore [40]. Table 1(c) shows the comparison results. Despite only RGB data used, UniNet still achieves 95.76% image-level AUROC, surpassing other methods and the baseline model by significant gains of 7.76% and 7.13%, respectively. Additionally, UniNet also enhances PRO metric compared to competing methods.

Medical datasets. UniNet was evaluated on three medical datasets: APTOS, OCT2017, and ISIC2018. Following [17], we compared UniNet with five recent methods: SimpleNet [33], CFA [27], AE-flow [58], PatchCore [40], and RD4AD [12]. Detailed results are presented in Table 2(a). UniNet achieves exceptional performance across all three evaluation metrics, with **100.0**% image-level AUROC on all three datasets. UniNet improves other methods and the baseline model by 12.21% AUROC and 9.85% AUROC on the more challenging ISIC2018 dataset, respectively.

Ped2. We compared UniNet with SOTA methods, including [39], AnomalyRuler [51], PDM-Net [21], SLM [44], and zxVAD [1]. As reported in Table 2(b), UniNet also achieves comparable performance to AnomalyRuler in video domain and surpasses the baseline by 2.7% AUROC.

5.2.2. Results under the supervised setting

VAD. UniNet was compared with methods reported in [2], including three unsupervised methods (EfficientAD [3], PatchCore [40], and RD4AD[12]) and two supervised methods (DevNet [38], DRA[14]). The results are summarized in Table 3(a). Even without SegAD, UniNet still achieves **99.95**% AUROC on this new and challenging dataset and improves the competing methods that equip with SegAD by 8.20% AUROC. It is noted that the unsupervised methods struggle without SegAD, as they face difficulty in distinguish anomalies. Additionally, supervised methods also perform poorly, as this dataset contains unseen anomalies, and they are biased by the seen anomalies.

MEM	DFS	$\mathcal{L}_{SC}$	$\mathcal{L}_M$	$\mathcal{M}$	I-AUROC	P-AUROC	PRO
					98.42	98.13	94.89
✓					99.01	98.20	95.06
✓	✓				99.23	98.29	95.13
✓		✓			99.36	98.34	95.40
✓	✓	✓			99.40	98.59	95.30
	✓	✓	✓		99.42	98.64	95.52
✓		✓	✓	✓	99.77	98.76	95.73
✓	✓	✓	✓	✓	**99.90**	**98.81**	**96.00**

Table 4. Ablation study on the key components of UniNet on the MVTec AD dataset. Best results are highlighted in bold.

Dataset	v_m				
	Max	3%	5%	10%	$\mathcal{M}$(Ours)
MVTec 3D-AD	90.30	94.42	94.90	93.68	**95.76**
APTOS	99.53	99.87	99.95	99.98	**100.0**
VAD	99.35	99.36	99.38	99.37	**99.95**

Table 5. Study on the effect of v_m on AD performance. Max=$\frac{100\%}{H \times W}$ denotes using the maximum value in the anomaly map. Best results are highlighted in bold.

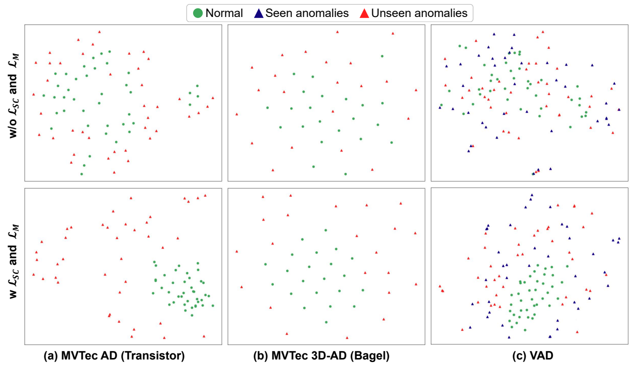
Medical datasets. UniNet was compared with several SOTA supervised methods (MADGNet [37], HarDNet-CPS [54], SAM-EG [48], MEGANet [7], and KDAS [49]) on three datasets, including Kvasir, CVC-ClinicDB, and CVC-ColonDB. Table 3(b) shows the detailed results. UniNet demonstrates promising performance on both the DSC and mIoU evaluation metrics, except for a 0.6% lower mIoU on the Kvasir dataset compared to MEGANet. Notably, although these methods are specifically designed for medical polyp segmentation, UniNet still outperforms them.

5.3. Ablation study

To demonstrate the effectiveness of UniNet, we conducted comprehensive ablation studies on datasets from different domains. More details can be found in Appendix.

Study on key elements. The key components of UniNet include MEM, DFS, similarity-contrastive loss $\mathcal{L}_{SC}$, margin loss $\mathcal{L}_M$, and weighted decision mechanism $\mathcal{M}$. The numerical results on the MVTec AD dataset are presented in Table 4. With MEM, UniNet can capture richer contextual relationships among features, which is help for visual tasks. By incorporating DFS, UniNet obtains domain-related information and thus enhances performance. When only employing $\mathcal{L}_{SC}$, the performance is lower than using both $\mathcal{L}_{SC}$ and $\mathcal{L}_M$ since the similarity for some normal features is inadequately high. Besides, the use of $\mathcal{M}$ substantially contributes to AD performance. Finally, combining all these elements, UniNet yields superior anomaly detection and segmentation performance.

Study on loss functions. To validate the effectiveness of $\mathcal{L}_{SC}$ and $\mathcal{L}_M$, we conducted experiments on two unsupervised datasets (MTVec AD and MTVec 3D-AD) and a supervised dataset (VAD). Without $\mathcal{L}_{SC}$ and $\mathcal{L}_M$, UniNet struggles to improve class boundaries in unsupervised setting. Instead, integrating them to UniNet enhances the correlations among normal features and maintains higher similarity for them, enabling effective anomaly discrimination (see Fig. 4(a) and (b)). In supervised setting, normal features are clustered, while anomalous features are repelled and their similarity are further constrained, which is helpful for the model to distinguish anomalies, including unseen anomalies (see Fig. 4(c)).



(a) MVTec AD (Transistor)
(b) MVTec 3D-AD (Bagel)
(c) VAD

Figure 4. Feature distributions visualized using t-SNE across three datasets.

Study on weighted decision mechanism. During inference, the precise calculation of weight v_w is crucial for the image-level anomaly score. Table 5 investigates the effects of different v_w on AD performance. As reported in Table 5, our weighted decision mechanism $\mathcal{M}$ achieves the best results across three datasets. Notably, compared to recent methods [22, 38] that select the fixed largest v_w (*e.g.*, v_w=3%) values from the anomaly map to calculate the anomaly score, our method dynamically determines the optimal v_w for more accurate AD, achieving a notable improvement of 0.86% on the MVTec 3D-AD dataset.

6. Conclusion

In this paper, we present UniNet, a generic unified anomaly detection framework designed for diverse domains. UniNet comprises student-teacher models and a bottleneck. The key innovations of UniNet are threefold: domain-related feature selection, similarity-contrastive loss and margin loss, and weighted decision mechanism. These components collectively enhance UniNet's ability to effectively select and learn domain-relevant feature information, distinguish abnormality and normality, and achieve robust AD performance during inference. Extensive experiments across 11 datasets from various domains demonstrate UniNet's superiority over SOTA methods.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 62372242.

References

[1] Abhishek Aich, Kuan-Chuan Peng, and Amit K Roy-Chowdhury. Cross-domain video anomaly detection without target domain adaptation. In *WACV*, pages 2579–2591, 2023.

[2] Aimira Baitieva, David Hurych, Victor Besnier, and Olivier Bernard. Supervised anomaly detection for complex industrial images. In *CVPR*, pages 17754–17762, 2024.

[3] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *WACV*, pages 128–138, 2024.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9592–9600, 2019.

[5] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.

[6] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[7] Nhat-Tan Bui, Dinh-Hieu Hoang, Quang-Thuc Nguyen, Minh-Triet Tran, and Ngan Le. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *WACV*, pages 7985–7994, 2024.

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[9] Yu-Min Chu, Chieh Liu, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Shape-guided dual-memory learning for 3d anomaly detection. In *ICML*, pages 6185–6194, 2023.

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[11] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[12] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, pages 9737–9746, 2022.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[14] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, pages 7388–7398, 2022.

[15] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, pages 11963–11975, 2022.

[16] Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Annan Shu, Guannan Jiang, and Lizhuang Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *ICCV*, pages 16401–16409, 2023.

[17] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. In *NIPS*, pages 10721–10740, 2023.

[18] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2404.06564*, 2024.

[19] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *AAAI*, pages 8472–8480, 2024.

[20] Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In *CVPR*, pages 2968–2977, 2023.

[21] Chao Huang, Jie Wen, Chengliang Liu, and Yabo Liu. Long short-term dynamic prototype alignment learning for video anomaly detection. In *IJCAI*, page 866–874, 2024.

[22] Yiming Huang, Guole Liu, Yaoru Luo, and Ge Yang. Adfa: Attention-augmented differentiable top-k feature adaptation for unsupervised medical anomaly detection. In *ICIP*, pages 206–210, 2023.

[23] Jeeho Hyun, Sangyun Kim, Giyoung Jeon, Seung Hwan Kim, Kyunghoon Bae, and Byung Jun Kang. Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In *WACV*, pages 2052–2061, 2024.

[24] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462, 2020.

[25] Jielin Jiang, Shun Wei, Xiaolong Xu, Yan Cui, and Xiying Liu. Unsupervised anomaly detection and localization based on two-hierarchy normalizing flow. *IEEE Trans. on Instrumentation and Measurement*, 73, 2024.

[26] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[27] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022.

[28] Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. In *CVPR*, pages 14143–14152, 2023.

[29] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, page 9664–9674, 2021.

[30] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):18–32, 2013.

[31] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, pages 6536–6545, 2018.

[32] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *CVPR*, pages 12147–12156, 2023.

[33] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *CVPR*, pages 20402–20411, 2023.

[34] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020.

[35] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

[36] 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 *ISIE*, pages 01–06, 2021.

[37] Ju-Hyeon Nam, Nur Suriza Syazwany, Su Jung Kim, and Sang-Chul Lee. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *CVPR*, pages 11480–11491, 2024.

[38] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021.

[39] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *CVPR*, pages 15984–15995, 2024.

[40] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022.

[41] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *WACV*, pages 2592–2602, 2023.

[42] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[43] Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *ECCV*, pages 474–489, 2022.

[44] Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. In *ICCV*, pages 10330–10340, 2023.

[45] Asia Pacific Tele-Ophthalmology Society. Aptos 2019 blindness detection. 2019.

[46] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. medical imaging*, 35(2):630–644, 2015.

[47] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *CVPR*, pages 24511–24520, 2023.

[48] Quoc-Huy Trinh, Hai-Dang Nguyen, Bao-Tram Nguyen Ngoc, Debesh Jha, Ulas Bagci, and Minh-Triet Tran. Sam-eg: Segment anything model with egde guidance framework for efficient polyp segmentation. *arXiv preprint arXiv:2406.14819*, 2024.

[49] Quoc-Huy Trinh, Minh-Van Nguyen, and Phuoc-Thao Vo Thi. Kdas: Knowledge distillation via attention supervision framework for polyp segmentation. In *ICME*, pages 1–6, 2024.

[50] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *CVPR*, pages 8032–8041, 2023.

[51] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *ECCV*, 2024.

[52] Hang Yao, Ming Liu, Haolin Wang, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. *arXiv preprint arXiv:2406.07487*, 2024.

[53] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *NIPS*, 35:4571–4584, 2022.

[54] Tong Yu and Qingxiang Wu. Hardnet-cps: colorectal polyp segmentation based on harmonic densely united network. *Biomedical Signal Processing and Control*, 85: 104953, 2023.

[55] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8330–8339, 2021.

[56] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *CVPR*, pages 3914–3923, 2023.

[57] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *CVPR*, pages 16699–16708, 2024.

[58] Yuzhong Zhao, Qiaoqiao Ding, and Xiaoqun Zhang. Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection. In *ICLR*, 2023.

[59] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, pages 392–408. Springer, 2022.

UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection

Supplementary Material

7. Overview

The supplementary material is organized as follows: Appendix 8 provides additional details on the datasets and the implementation of UniNet. Appendix 9 presents further experimental results on the MVTec AD, BTAD, VisA, and MVTec 3D-AD datasets, as well as the complexity analysis of UniNet. Appendix 10 presents supplementary ablation study results. Appendix 11 includes additional visualization results across various datasets. Appendix 12 discusses the limitations and potential directions for future work.

8. Experimental setup

8.1. Datasets

MVTec AD [4] is a widely used dataset for industrial anomaly detection, comprising 15 object and texture categories with a total of over 5,000 images. The dataset contains various types of anomalies, such as scratches and crackes. Each category includes a training set consisting solely of normal images and a test set of containing both normal and abnormal images.

BTAD [36] is a real-world industrial dataset consisting of 3 different types of industrial products. The dataset contains 2830 images, with 400, 1,000, and 399 training images in class 1, 2, and 3, respectively.

VisA [59] is a publicly available dataset for visual anomaly detection, comprising 12 categories and a total of 10,821 high-resolution images from diverse domains such as electronics, food, and industrial parts. The dataset includes both normal and anomalous samples, with detailed annotations for the anomalies.

MVTec 3D-AD [5] is a multi-modal dataset that includes two different modality: RGB images and Point Clouds. The dataset consists of 10 real-world categories with a total of 4147 high-resolution images.

VAD [2] is a newly introduced supervised benchmark designed to encompass a wider array of complex anomalies and substantial intra-class variability in anomalous-free images. The dataset contains 5,000 object images, with 165 unseen anomalous images reserved for testing.

APTOS [45] is a collection of color fundus images from the 2019 APTOS blindness detection challenge. Each image is associated with a label (ranging from 0 to 4) that indicates the severity of diabetes retinopathy, with grade 0 representing normal images.

OCT2017 [26] is a dataset of optical coherence tomography images, with one class labeled as normal and three other classes labeled as abnormal. The dataset contains over 20,000 images, with 1000 images used for testing.

Algorithm 1: Weighted Decision Mechanism for anomaly detection

```
Input: Pixel-level anomaly map $M_{AS}$ for each
       sample and outputs $F_S^i$ and $F_T^i$ from the S-T
       models
Output: Anomaly score $S_{AD}$ for each sample
Function Main
    // Low similarity value $v_l$
       generation
    for $i = 1, 2, \ldots, n$ do
        minimize vector-wise cosine similarity
         between $\{F_S^i, F_T^i\}$ by $d(\cdot)$ and obtain $v_{l_i}$
         by $max(\cdot)$
    // Transform into the
       probability distribution
    $v_p = \text{Softmax}(v_l)$
    // Weight $v_w$ generation
    for $i = 1, 2, \ldots, 2n$ do
        if $v_{p_i} > \frac{1}{2n} \sum_{i=1}^{2n} v_{p_i}$ then
            Incorporate large values into $\mathcal{P}$
    Compute $v_w$ using $\mathcal{P}$ based on Eq. (14)
    // Evaluate anomaly score $S_{AD}$
    $S_{AD} = \frac{1}{v_w} \sum_{1}^{K=v_w} \text{top}K(M_{AS})$
    return $S_{AD}$
```

ISIC2018 [11] is a collection of skin disease images from Task 3 of the ISIC2018 challenge. The dataset includes seven classes, with *nevus* labeled as the normal class and the remaining classes representing various types of anomalies. Following [17], 6705 normal images from training set are used, while the validation set of 193 images serves as the test set.

Kvasir [24], **CVC-ClinicDB** [6], and **CVC-ColonDB** [46] are three polyp segmentation datasets, containing a total of 1,000, 612, and 379 images, respectively, sourced from diverse imaging clinics and centers. Each image is accompanied by a corresponding pixel-level mask.

Ped2 [30] is a dataset designed for video anomaly detection, consisting of 2.6K frames for training and 2.0K frames for testing. The anomalies in the dataset include cycling, skateboarding, *etc*.

(a) MVTec AD

Category	**UniNet(Ours)**	ReContrast [17]	RealNet [57]	ReConPatch [23]	GLAD [52]	RD++ [47]	DMAD [32]
Carpet	100.0 / 99.2 / 97.5	99.8 / 99.3 / 97.9	99.8 / 99.2 / 96.4	99.6 / 98.8 / –	99.0 / 98.5 / –	100.0 / 99.2 / 97.7	– / – / –
Grid	99.5 / 99.4 / 98.0	100.0 / 99.2 / 97.8	100.0 / 99.5 / 97.3	100.0 / 99.0 / –	100.0 / 99.6 / –	100.0 / 99.3 / 97.7	– / – / –
Leather	100.0 / 99.5 / 98.3	100.0 / 99.5 / 99.2	100.0 / 99.8 / 96.2	100.0 / 96.0 / –	100.0 / 99.8 / –	100.0 / 99.4 / 99.2	– / – / –
Tile	99.5 / 97.3 / 90.9	99.8 / 96.3 / 93.6	100.0 / 99.4 / 97.7	99.8 / 98.9 / –	100.0 / 98.7 / –	99.7 / 96.6 / 92.4	– / – / –
Wood	100.0 / 99.2 / 98.1	99.0 / 95.9 / 92.5	99.2 / 98.2 / 90.5	99.7 / 98.9 / –	99.4 / 98.4 / –	99.3 / 95.8 / 93.3	– / – / –
Bottle	100.0 / 98.9 / 96.6	100.0 / 99.0 / 97.1	100.0 / 99.3 / 95.6	100.0 / 98.2 / –	100.0 / 98.9 / –	100.0 / 98.8 / 97.0	– / – / –
Cable	100.0 / 98.5 / 93.7	99.8 / 98.9 / 95.6	99.2 / 98.1 / 93.9	99.8 / 99.3 / –	99.9 / 98.1 / –	99.2 / 98.4 / 93.9	– / – / –
Capsule	100.0 / 99.0 / 94.8	97.7 / 98.4 / 95.4	99.6 / 99.3 / 84.5	98.8 / 97.6 / –	99.5 / 98.5 / –	99.0 / 98.8 / 96.4	– / – / –
Hazelnut	100.0 / 99.0 / 96.8	100.0 / 99.1 / 95.9	100.0 / 99.7 / 93.1	100.0 / 98.9 / –	100.0 / 98.5 / –	100.0 / 99.2 / 96.3	– / – / –
Metal nut	100.0 / 98.7 / 96.5	100.0 / 98.7 / 94.4	99.8 / 98.6 / 94.4	100.0 / 95.8 / –	100.0 / 98.8 / –	100.0 / 98.1 / 93.0	– / – / –
Pill	100.0 / 98.5 / 96.9	98.6 / 99.1 / 97.7	99.1 / 99.0 / 91.0	97.5 / 95.4 / –	98.1 / 97.9 / –	98.4 / 98.3 / 97.0	– / – / –
Screw	100.0 / 99.5 / 97.6	98.0 / 99.6 / 98.6	99.4 / 99.5 / 87.9	98.5 / 98.8 / –	96.9 / 99.1 / –	98.9 / 99.7 / 98.6	– / – / –
Toothbrush	100.0 / 99.1 / 93.4	100.0 / 99.2 / 95.0	100.0 / 98.7 / 91.6	100.0 / 98.9 / –	100.0 / 99.4 / –	100.0 / 99.1 / 94.2	– / – / –
Transistor	100.0 / 97.7 / 94.7	99.7 / 95.4 / 82.3	99.8 / 98.0 / 92.9	100.0 / 99.6 / –	98.3 / 96.2 / –	98.5 / 94.3 / 81.8	– / – / –
Zipper	99.5 / 98.7 / 95.9	99.5 / 98.1 / 94.9	99.6 / 99.2 / 93.4	99.8 / 98.6 / –	98.5 / 97.9 / –	98.6 / 98.8 / 96.3	– / – / –
Mean	99.90 / 98.81 / 96.00	99.46 / 98.41 / 95.20	99.65 / 99.03 / 93.07	99.56 / 98.18 / –	99.30 / 98.62 / 95.31	99.44 / 98.25 / 94.99	99.50 / 98.21 / –

(b) BTAD

Category	**UniNet(Ours)**	ReContrast [17]	RealNet [57]	ReConPatch [23]	PyramidFlow [28]	RD++ [47]	PatchCore [40]
Class 01	100.0 / 97.2 / 81.7	100.0 / 97.0 / 78.6	100.0 / 98.2 / –	99.7 / 96.8 / –	100.0 / 97.4 / –	96.8 / 96.2 / 73.2	98.0 / 96.9 / 64.9
Class 02	93.2 / 96.3 / 60.1	89.5 / 96.2 / 57.0	88.6 / 96.3 / –	87.7 / 96.6 / –	88.2 / 97.6 / –	90.1 / 96.4 / 71.3	81.6 / 95.8 / 47.3
Class 03	100.0 / 99.6 / 98.2	95.7 / 99.3 / 96.5	96.1 / 97.9 / –	100.0 / 99.0 / –	99.3 / 98.1 / –	100.0 / 99.6 / 87.4	99.8 / 99.1 / 67.7
Mean	97.73 / 97.70 / 80.01	95.06 / 97.50 / 77.40	96.07 / 97.90 / –	95.80 / 97.47 / –	95.83 / 97.70 / –	95.63 / 97.41 / 77.30	93.13 / 97.27 / 59.97

(c) MVTec 3D-AD

Category	**UniNet(Ours)**	ReContrast [17]	BTF [20]	Shape-Guided [9]	M3DM [50]	AST [41]	PatchCore [40]
Bagel	100.0 / 95.2	99.1 / –	85.4 / 89.8	91.1 / 94.6	94.4 / 95.2	94.7 / –	91.2 / 89.9
Cable Gland	99.6 / 98.1	95.3 / –	84.0 / 94.8	93.6 / 97.2	91.8 / 97.2	92.8 / –	90.2 / 95.3
Carrot	100.0 / 97.3	92.7 / –	82.4 / 92.7	88.3 / 96.0	89.6 / 97.3	85.1 / –	88.5 / 95.7
Cookie	73.3 / 90.3	69.6 / –	68.7 / 87.2	66.2 / 91.4	74.9 / 89.1	82.5 / –	70.9 / 91.8
Dowel	100.0 / 98.4	97.5 / –	97.4 / 92.7	97.4 / 95.8	95.9 / 93.2	98.1 / –	95.2 / 93.0
Foam	92.8 / 85.4	82.5 / –	71.6 / 55.5	77.2 / 77.6	76.7 / 84.3	95.1 / –	73.3 / 71.9
Peach	98.9 / 98.1	95.0 / –	71.3 / 90.2	78.5 / 93.7	91.9 / 97.0	89.5 / –	72.7 / 92.0
Potato	98.6 / 95.8	67.9 / –	59.3 / 93.1	64.1 / 94.9	64.8 / 95.6	61.3 / –	56.2 / 93.7
Rope	99.9 / 99.0	98.8 / –	92.0 / 90.3	88.4 / 95.6	93.8 / 96.8	99.2 / –	96.2 / 93.8
Tire	94.5 / 97.9	87.9 / –	72.4 / 89.9	70.6 / 95.7	76.7 / 96.6	82.1 / –	76.8 / 92.9
Mean	95.76 / 95.55	88.63 / 95.20	78.52 / 87.63	81.51 / 93.30	85.03 / 94.22	88.00 / –	81.14 / 91.03

(d) The standard deviation

Dataset	△
MVTec AD	0.03 0.02 0.04
BTAD	0.01 0.05 0.03
MVTec 3D-AD	0.99 – 0.33

Table 6. Quantitative results across three industrial datasets. We report I-AUROC / P-AUROC / PRO on the (a) MVTec AD and (b) BTAD datasets. For the (c) MVTec 3D-AD dataset, P-AUROC is not presented. (d) Each cell, from top to bottom, represents the standard deviation of I-AUROC, P-AUROC, and PRO. Best and second-best results are highlighted in red and blue, respectively.

8.2. Implementation details

UniNet was trained on a computer with NVIDIA GeForce RTX 3090. Following [12, 17], we used the publicly available WideResNet50 pre-trained on ImageNet [13] as S-T models. AdamW [35] was employed as the optimizer with weight decay=1e-5, and the learning rate of 5e-3 and 1e-6 for the learnable student and teacher, respectively. The batch size was 8. Hyper-parameters n, $\mathcal{T}$, τ, λ, α, and β were set to 3, 2, 1, 0.7, 0.01, and 0.03, respectively.

All images were resized into 256 × 256 without data augmentation, except for the MVTec 3D-AD dataset and three polyp segmentation datasets. For the MVTec 3D-AD dataset, only RGB data were used for training and images were first center-cropped before resizing them. Following [7, 49], we adopted a multi-scale $\{0.75, 1, 1.25\}$ training strategy for three polyp segmentation datasets. For a fair comparison, we followed prior works that selected a specific proportion of images from the Kvasir and CVC-ClinicDB datasets for training, while the remaining images were used for testing. For the CVC-ColonDB dataset, we directly employed its training and test sets for training and evaluation.

The procedure for the Weighted Decision Mechanism is outlined in Algorithm 1. The Weighted Decision Mechanism was not applied to the three polyp segmentation datasets, as only segmentation evaluation metrics were considered. For these three polyp datasets, segmentation accuracy was evaluated by comparing the upsampled output of the student model with its pixel-level ground-truth. For the Ped2 dataset, we employed the frame-ped strategy [31],

(a) MVTec AD

Category	**UniNet(Ours)**	ReContrast [17]	MambaAD [18]	DiAD [19]	DeSTSeg [56]	SimpleNet [33]	UniAD [53]
Carpet	99.4 / 99.8	98.3 / –	99.8 / 99.9	99.4 / 99.9	95.9 / 98.8	95.7 / 98.7	99.8 / 99.9
Grid	99.1 / 99.7	98.9 / –	100.0 / 100.0	98.5 / 99.8	97.9 / 99.2	97.6 / 99.2	98.2 / 99.5
Leather	100.0 / 100.0	100.0 / –	100.0 / 100.0	99.8 / 99.7	99.2 / 99.8	100.0 / 100.0	100.0 / 100.0
Tile	97.8 / 99.2	99.5 / –	98.2 / 99.3	96.8 / 99.9	97.0 / 98.9	99.3 / 99.8	99.3 / 99.8
Wood	100.0 / 100.0	99.7 / –	98.8 / 99.6	99.7 / 100.0	99.9 / 100.0	98.4 / 99.5	98.6 / 99.6
Bottle	100.0 / 100.0	100.0 / –	100.0 / 100.0	99.7 / 96.5	98.7 / 99.6	100.0 / 100.0	99.7 / 100.0
Cable	94.9 / 97.1	95.6 / –	98.8 / 99.2	94.8 / 98.8	89.5 / 94.6	97.5 / 98.5	95.2 / 95.9
Capsule	96.3 / 99.2	97.3 / –	94.4 / 98.7	89.0 / 97.5	82.8 / 95.9	90.7 / 97.9	86.9 / 97.8
Hazelnut	100.0 / 100.0	100.0 / –	100.0 / 100.0	99.5 / 99.7	98.8 / 99.2	99.9 / 99.9	99.8 / 100.0
Metal nut	100.0 / 100.0	100.0 / –	99.9 / 100.0	99.1 / 96.0	92.9 / 98.4	96.9 / 99.3	99.2 / 99.9
Pill	98.3 / 99.6	96.3 / –	97.0 / 99.5	95.7 / 98.5	77.1 / 94.4	88.2 / 97.7	93.7 / 98.7
Screw	100.0 / 100.0	97.2 / –	94.7 / 97.9	90.7 / 99.7	69.9 / 88.4	76.7 / 90.6	87.5 / 96.5
Toothbrush	100.0 / 100.0	96.7 / –	98.3 / 99.3	99.7 / 99.9	71.7 / 89.3	89.7 / 95.7	94.2 / 97.4
Transistor	100.0 / 100.0	94.5 / –	100.0 / 100.0	99.8 / 99.6	78.2 / 79.5	99.2 / 98.7	99.8 / 98.0
Zipper	100.0 / 100.0	99.4 / –	99.3 / 99.8	95.1 / 99.1	88.4 / 96.3	99.0 / 99.7	95.8 / 99.5
Mean	99.05 / 99.64	98.23/ 99.40	98.61 / 99.55	97.15 / 98.97	89.19 / 95.49	95.25 / 98.36	96.51 / 98.83

(b) VisA

Category	**UniNet(Ours)**	ReContrast [17]	MambaAD [18]	DiAD [19]	DeSTSeg [56]	SimpleNet [33]	UniAD [53]
pcb1	100.0 / 100.0	96.5 / –	95.4 / 93.0	88.1 / 88.7	87.6 / 83.1	91.6 / 91.9	92.8 / 92.7
pcb2	99.8 / 99.8	96.8 / –	94.2 / 93.7	91.4 / 91.4	86.5 / 85.8	92.4 / 93.3	87.8 / 87.7
pcb3	92.0 / 93.6	96.8 / –	93.7 / 94.1	86.2 / 87.6	93.7 / 95.1	89.1 / 91.1	78.6 / 78.6
pcb4	100.0 / 100.0	99.9 / –	99.9 / 99.9	99.6 / 99.5	97.8 / 97.8	97.0 / 97.0	98.8 / 98.8
macaroni1	100.0 / 100.0	97.6 / –	91.6 / 89.8	85.7 / 85.2	76.6 / 69.0	85.9 / 82.5	79.9 /79.8
macaroni2	100.0 / 100.0	89.5 / –	81.6 / 78.0	62.5 / 57.4	68.9 / 62.1	68.3 / 54.3	71.6 / 71.6
capsules	99.9 / 100.0	77.7 / –	91.8 / 95.0	58.2 / 69.0	87.1 / 93.0	74.1 / 82.8	55.6 / 55.6
candle	100.0 / 100.0	96.3 / –	96.8 / 96.9	92.8 / 92.0	94.9 / 94.8	84.1 / 73.3	94.1 / 94.0
cashew	96.3 / 98.0	94.5 / –	94.5 / 97.3	91.5 / 95.7	92.0 / 96.1	88.2 / 91.3	92.8 / 92.8
chewinggum	100.0 / 100.0	98.6 / –	97.7 / 98.9	95.1 / 99.5	95.8 / 98.3	96.4 / 98.2	96.3 / 96.2
fryum	99.2 / 99.6	97.3 / –	95.2 / 97.7	89.8 / 95.0	92.1 / 96.1	88.4 / 93.0	83.0 / 83.0
pipe_fryum	99.5 / 99.8	99.3 / –	98.7 / 99.3	96.2 / 98.1	94.1 / 97.1	90.8 / 95.5	94.7 / 94.7
Mean	98.9 / 99.2	95.1/ 96.4	94.3 / 94.5	86.8 / 88.3	88.9 / 89.0	87.2 / 87.0	91.5 / 90.8

Table 7. Quantitative results across two industrial datasets. I-AUROC and Image-level AP are reported for the multi-class anomaly detection. Best and second-best results are highlighted in red and blue, respectively.

which detects anomalies by measuring the discrepancy between the student-generated frame and its corresponding ground-truth.

9. More experimental results

9.1. Results on the industrial datasets

Traditional methods develop separate models for each category, known as the *one-class* anomaly detection setting. Recent efforts [17, 18, 53] have attempted to design a unified model that can handle multiple categories, *i.e.*, the *multi-class* anomaly detection setting. Experimental results for both settings are reported as follows.

Results under the one-class setting. In addition to the overall average results across all categories from the MVTec AD, BTAD, and MVTec 3D-AD datasets, the average results for each individual category from these three datasets are also presented in Table 6. As reported in Table 6(a), UniNet achieves **100.0**% anomaly detection performance across all categories of the MVTec AD dataset, with the exception for the grid, tile, and zipper categories. It also shows comparable segmentation performance. Moreover, UniNet demonstrates notable anomaly detection and segmentation performance across most categories in the other two datasets, as illustrated in Table 6(b) and (c). Particularly, on the MVTec 3D-AD dataset, UniNet achieves the best and significant results across all categories, except for *Cookie* category. Finally, the standard deviations of the three evaluation metrics across three datasets are presented in Table 6(d).

Results under the multi-class setting. Table 7 shows the multi-class anomaly detection on the MVTec AD and VisA [59] datasets. Following [17–19], both I-AUROC and Image-level AP are reported. UniNet was compared with state-of-the-art methods reported in [18]: MambaAD [18], DiAD [19], DeSTSeg [56], SimpleNet [33], and UniAD [53]. As shown in Table 7(a), UniNet similarly shows strong performance across most categories, achieving the highest average I-AUROC and Image-level AP, with a perfect score of **100.0**%. UniNet outperforms the other methods and the baseline model by 0.44% and 0.82% in I-AUROC, as well as by 0.09% and 0.24% in Image-level AP.

Moreover, as reported in Table 7(b), UniNet also achieves impressive anomaly detection performance on the more challenging VisA dataset, obtaining the best results in every category except for the *pcb3* category. UniNet markedly surpasses leading methods, with improvement of **3.8%** in I-AUROC and **2.8%** in Image-level AP, respectively.

9.2. Complexity analysis

Table 8 investigates the complexity of UniNet and the baseline model, ReContrast [17]. By utilizing the same backbone as the baseline model, UniNet achieves a comparable model size to ReContrast, while offering a higher inference speed with an improvement of 6.77 FPS. Additionally, UniNet outperforms ReContrast in terms of I-AUROC, P-AUROC and PRO, with increases of 0.44%, 0.40%, and 0.80%, respectively.

Method	Model Size (GB) ↓	Speed (FPS) ↑	Infer. Time (s) ↓	Metrics
ReContrast	**0.141**	9.46	15.6	99.46 / 98.41 / 95.20
UniNet	0.150	**16.23**	**8.2**	**99.90 / 98.81 / 96.00**

Table 8. Complexity analysis between UniNet and the baseline model on the MVTec AD dataset. Metrics are I-AUROC / P-AUROC / PRO. Best results are highlighted in bold.

10. Supplementary ablation studies

10.1. Study on Multi-Scale Embedding Module

To validate the effectiveness of MEM within the bottleneck, we studied the effect on the kernel size k in MEM. The results are presented in Table 9. Both detection and segmentation performance steadily improve as the large kernel size increases and the best results can be obtained when using a combination of (3, 7). Notably, larger kernels lead to a higher number of model size and decreased inference speed. As shown in Table 9, the model size of the bottleneck (*e.g.*, 0.179 GB and 0.363 GB) can significantly surpass that of the entire framework (see Table 8) prior to re-parameterization. Similarly, as the kernel size increases, the inference speed of UniNet progressively decreases. However, re-parameterization results in a smaller model size and improved inference speed.

10.2. Study on Domain-Related Feature Selection

To demonstrate that introducing domain-related information into the student aids in improving its feature representations, we investigated the impact of different selection strategies on three datasets, as shown in Table 10. Without selecting representative features from the teacher, the student faces challenges in understanding target-oriented feature information, especially on more structurally complex datasets (*e.g.*, VAD), which negatively affects performance. Conversely, our method effectively guides the student to select and learn the most crucial features, yielding promising results. However, selecting all available information from the teacher may not be beneficial, as it could include unimportant details.

k	Model Size (GB) ↓	Speed (FPS) ↑	Metrics
(3, 3)	**0.077**+0.00%	**16.04**+0.00%	99.82 / 98.16 / 95.81
(3, 5)	0.118-34.75%	15.86+2.28%	99.88 / 98.20 / 95.87
(3, 7)	0.179-57.02%	15.33+5.55%	**99.90 / 98.81 / 96.00**
(3, 11)	0.363-78.80%	12.13+25.27%	99.87 / 98.16 / 95.75

Table 9. Study on kernel size k in MEM on MVTec AD dataset, with only the model size of bottleneck reported. Metrics are I-AUROC / P-AUROC / PRO. The gains after re-parameterization are highlighted in green, with the best results indicated in bold.

Method	Dataset		
	MVTec AD	APTOS	VAD
F_S	99.77 / 98.76 / 95.73	99.99 / **99.63 / 99.50**	99.25 / 95.88 / 95.90
$(F_S)^P$	**99.90 / 98.81 / 96.00**	**100.0** / 99.60 / 99.44	**99.95 / 98.60 / 98.60**
$(F_S)^A$	99.81 / 98.75 / 95.80	99.99 / 99.55 / 99.37	99.87 / 98.20 / 98.20

Table 10. Study on different selection strategies for DFS. For the MVTec AD, the evaluation metrics include I-AUROC / P-AUROC / PRO. For the APTOS and VAD datasets, three metrics are reported: I-AUROC / FI / ACC. "F_S", "$(F_S)^P$", and "$(F_S)^A$" refer to no feature selection, selecting representative features, and selecting all available features, respectively. Best results are highlighted in bold.

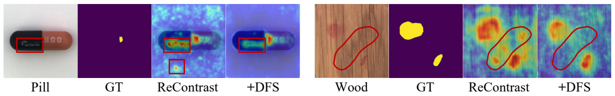


Figure 5. Segmentation results w/o and w DFS on the MVTec AD dataset.

We also investigated the impacts of DFS on the performance of ReContrast, as illustrated in Fig. 5. Without DFS, ReContrast fails to sufficiently learn vital domain-related features, leading to the loss of subtle details–such as the label on a pill being mistakenly identified as an anomaly. By incorporating DFS, ReContrast mitigate this issue by selecting key features for learning.

10.3. Additional study on key elements

In addition to industrial datasets, ablation studies on the key components of UniNet on medical and video domains are listed in Table 11.

10.4. Hyper-parameter sensitivity analysis

The main hyper-parameters include temperature coefficient $\mathcal{T}$ and $\{\alpha, \beta\}$ (controlling the upper and lower limits of the weight in $\mathcal{M}$). As shown in Table 12, we evaluated different

MEM	DFS	$\mathcal{L}_{SC}$	$\mathcal{L}_{M}$	$\mathcal{M}$	ATPOS	Ped2
					95.17	95.01
✓					95.79	95.30
✓	✓				96.52	95.63
✓		✓			96.24	95.35
✓	✓	✓			97.89	96.09
	✓	✓	✓		97.80	96.20
✓		✓	✓	✓	99.55	97.40
✓	✓	✓	✓	✓	**100.0**	**97.91**

Table 11. Ablation studies on the key elements of UniNet on medical and video datasets, with I-AUROC listed.

$\mathcal{T}$	α	β	ATPOS	Ped2
0.1	0.01	0.03	99.88	97.72
0.1	0.1	0.1	99.85	97.66
0.5	0.05	0.05	99.90	**97.94**
0.5	0.01	0.05	99.90	97.84
1	0.1	0.03	99.60	97.68
1	0.01	0.1	99.72	97.60
2	0.01	0.03	**100.0**	97.91
2	0.1	0.05	99.97	97.80

Table 12. Hyper-parameter analysis on medical and video datasets, with I-AUROC reported.

combinations of $\mathcal{T}$ (0.1, 0.5, 1, 2), α (0.01, 0.05, 0.1), and β (0.03, 0.05, 0.1).

11. Qualitative Results

To clearly validate the superior segmentation performance of UniNet, comprehensive visualization results are presented across three industrial datasets and three medical datasets.

11.1. Visualization on industrial datasets

As illustrated in Fig. 6, UniNet effectively segments both local and global anomalies across texture and object categories, while maintaining lower anomaly scores in regions devoid of anomalies.

Results on the BTAD and MTVec 3D-AD datasets are respectively shown in Fig. 7(a) and (b). For the BTAD dataset, despite the anomalies closely resembling normal areas, UniNet exhibits exceptional segmentation performance, effectively detecting even the smallest anomalies. For the MVTec 3D-AD dataset, using only the RGB modality, UniNet still achieves promising segmentation results, as shown in Fig. 7(b). However, due to lack of multi-modal information, UniNet may fail to maintain lower anomaly scores in some normal regions, such as the *Bagel*, *Cookie*, and *Potato* categories. This is because the chocolates in the *Cookie* category resemble anomalies, such as holes. As a result, relying on a single modality alone makes it challenge to achieve more accurate segmentation. We will explore combining other modalities with the RGB modality later.

11.2. Visualization on medical datasets

In addition to industrial datasets, results on three polyp datasets are visualized in Fig. 7(c). Despite the variability in images collected from the intestinal environments of different patients, UniNet also demonstrates superior segmentation performance in polyps. As illustrated in Fig 7(c), the segmented results perfectly match the ground-truths, demonstrating that UniNet is highly resistant to both over-segmentation and under-segmentation.

12. Discussion

12.1. Limitation

Similar to ReContrast [17] and other unsupervised AD methods [12, 27, 33], UniNet also experiences training instability for certain categories, with performance fluctuating when overtraining occurs or random seeds are changed, particularly in anomaly segmentation performance. However, thank to weighted decision mechanism $\mathcal{M}$, anomaly detection performance can hardly be influenced, ensuring robust anomaly detection results. Besides, although UniNet has achieved promising results on multimodal datasets like MVTec 3D-AD, relying solely on 2D data limits its potential for better anomaly detection performance.

12.2. Future work

We will apply UniNet to other tasks, such as multimodal anomaly detection or 3D medical image segmentation, by incorporating other modalities like text or point cloud to achieve superior performance. Also, the optimization of loss functions and the model will be investigated to ensure more stable training.

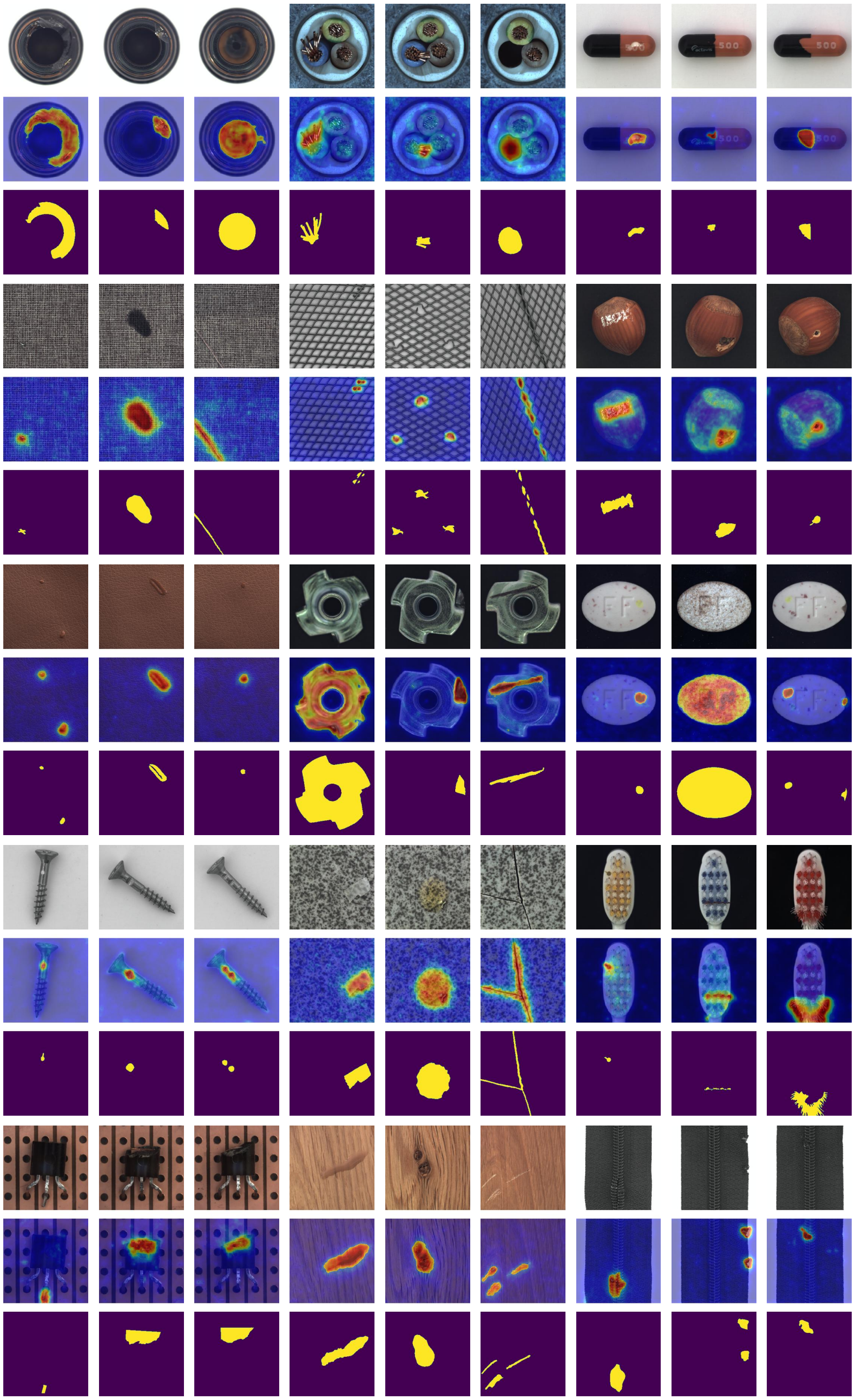


Figure 6. Visualization of UniNet on the MVTec AD dataset. Each group, from top to bottom, displays the anomalous images, our segmentation results, and ground-truths, respectively.

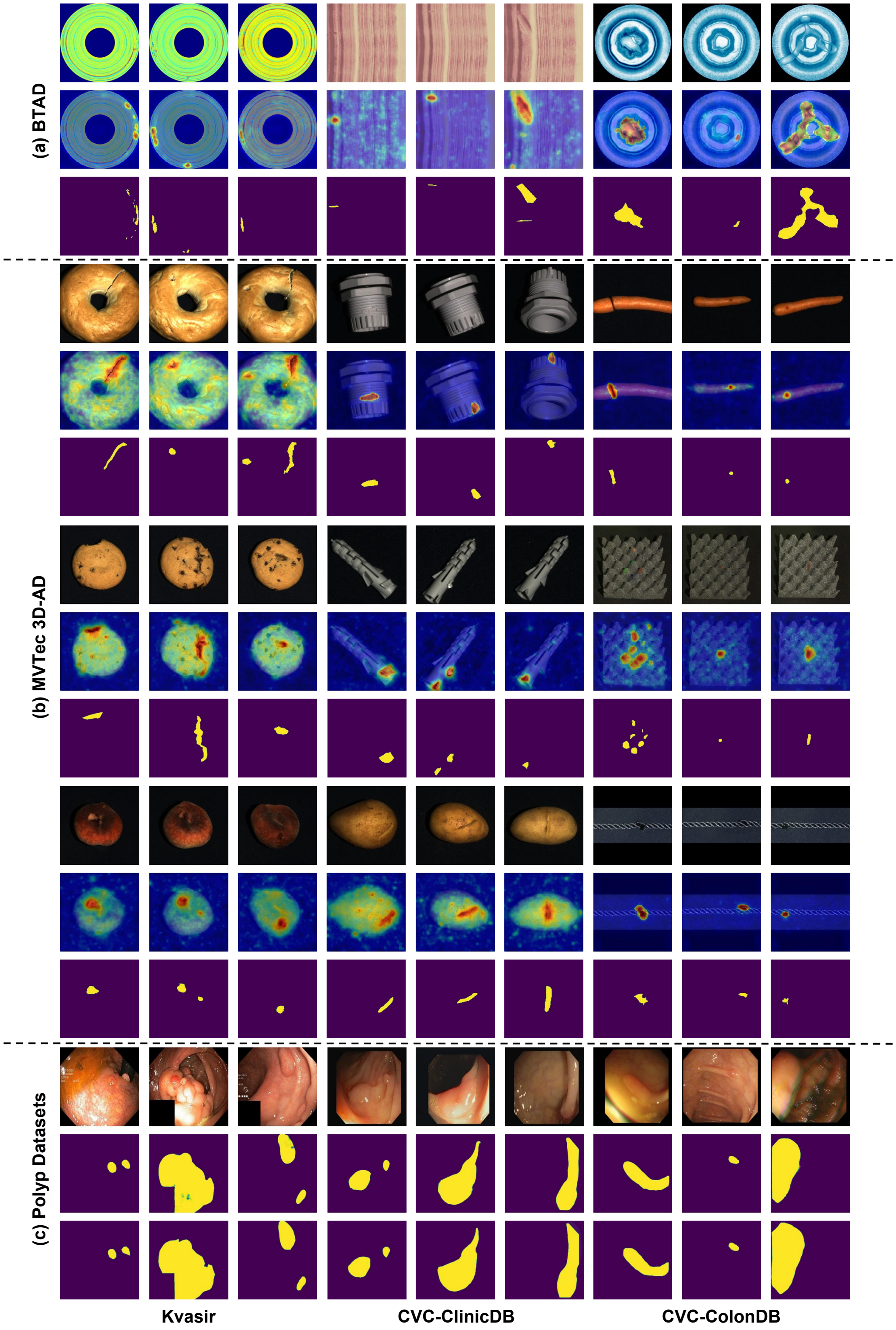


Figure 7. Visualization of UniNet on the BTAD, MVTec 3D-AD, and three polyp datasets (Kvasir, CVC-ClinicDB, and CVC-ColonDB). For the MVTec 3D-AD dataset, all categories are listed in order: *Bagel, Cable gland, Carrot, Cookie, Dowel, Foam, Peach, Potato,* and *Rope*. Each group, from top to bottom, displays the anomalous images, our segmentation results, and ground-truths, respectively.