UniAD: Integrating Geometric and Semantic Cues for Unified Anomaly Detection

Xiaodong Wang
xdwangjsj@xmut.edu.cn
Xiamen University of Technology
Xiamen, China

Hongmin Hu*
2322071015@stu.xmut.edu.cn
Xiamen University of Technology
Xiamen, China

Fei Yan
fyan@xmut.edu.cn
Xiamen University of Technology
Xiamen, China

Junwen Lu
jwlu@xmut.edu.cn
Xiamen University of Technology
Xiamen, China

Zhiqiang Zeng
2010110707@xmut.edu.cn
Xiamen University of Technology
Xiamen, China

Weidong Hong
otis@truesightai.com
Xiamen Truesight Technology Co., Ltd
Xiamen, China

Zhedong Zheng
zhedongzheng@um.edu.mo
University of Macau
Macau, China

Abstract

Current anomaly detection paradigms face inherent limitations in simultaneously addressing structural anomalies (*e.g.*, geometric distortions) and logical anomalies (*e.g.*, semantic inconsistencies), due to conflicting feature representation requirements between these two anomaly categories. We propose UniAD, a novel dual-branch teacher-student framework that achieves unified anomaly detection through synergistic integration of complementary expertise from heterogeneous vision models without requirements of extra manual annotations. In particular, our framework integrates two frozen expert models as teachers: (1) a structural teacher specializing in geometric-sensitive patterns, and (2) a logical teacher focusing on semantic-aware representations via component relationship modeling. To resolve feature conflicts while preserving complementary information, the student network is equipped with one shared backbone and two independent branches. One branch employs multi-scale feature alignment with the structural teacher while another branch establishing semantic correspondence with the logical teacher through component-aware attention mechanisms. Furthermore, we introduce the text-guided semantic enhancement module as a kind of logical guidance to facilitate the anomaly indicator. Extensive experiments on the challenging MVTec LOCO benchmark validate that the scalability of our model to localize both geometric distortions and semantic inconsistencies. The proposed method outperforms existing single-purpose detectors, yielding 93.7% AUROC for logical anomalies and 93.2% AUROC for structural anomalies.

CCS Concepts

• **Computing methodologies → Anomaly detection**; • **Information systems → Multimedia information systems**.

Keywords

Heterogeneous Networks, Teacher-student Distillation, Anomaly Detection, Dual-branch Architecture, Text-guided Enhancement

ACM Reference Format:
Xiaodong Wang, Hongmin Hu, Fei Yan, Junwen Lu, Zhiqiang Zeng, Weidong Hong, and Zhedong Zheng. 2025. UniAD: Integrating Geometric and Semantic Cues for Unified Anomaly Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755422

1 Introduction

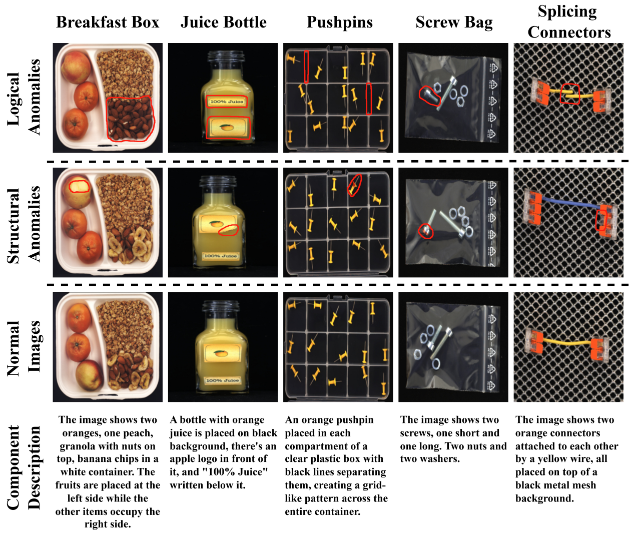
Image anomaly detection stands as a pivotal technology and tool in the field of automation control. Moreover, this technique has wide-ranging applications across various fields, including medical image analysis [18, 47] and industrial inspection [6, 27]. Particularly in the industrial anomaly detection, this technique is of paramount importance. In this context, anomaly detection, also known as defect detection, aims to identify samples that are defective or erroneous. Common defects in samples include scratches, cracks, notches, etc., which are referred to as structural anomalies. Structural anomalies usually involve the physical or architectural structure of system components, indicating an abnormality in the whole structure of object. Effective image anomaly detection techniques can significantly enhance product quality and efficiency, reducing economic losses and waste caused by defective products.

In industrial detection, unsupervised methods do not require labeled data, which is often costly and time-consuming to obtain. This makes unsupervised methods more practical for large-scale applications, especially when labeled data are scarce or unavailable. Most existing anomaly detection models employ unsupervised learning methods to address structural anomaly tasks, with mainstream approaches including image reconstruction-based methods [15, 28] and feature embedding-based methods [7, 25, 37], among others. Through these methods, a variety of models have

[*] Corresponding Author

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755422



The image shows two oranges, one peach, granola with nuts on top, banana chips in a white container. The fruits are placed at the left side while the other items occupy the right side.

A bottle with orange juice is placed on black background, there's an apple logo in front of it, and "100% Juice" written below it.

An orange pushpin placed in each compartment of a clear plastic box with black lines separating them, creating a grid-like pattern across the entire container.

The image shows two screws, one short and one long. Two nuts and two washers.

The image shows two orange connectors attached to each other by a yellow wire, all placed on top of a black metal mesh background.

Figure 1: Here we show both normal and anomalous samples in the MVTec LOCO AD dataset. It include logical anomalies, structural anomalies, and normal images from top to bottom (anomalous regions highlighted in red). In this work, we also harness the textual component description to explicitly modeling semantic cues.

achieved high detection performance in handling structural anomaly tasks [30, 31, 36]. However, these models are all built for structural anomalies and do not perform satisfactorily when it comes to another type of anomaly detection task—logical anomaly detection.

Unlike structural anomalies, logical anomalies do not manifest as obvious notches or defects that are easily identifiable, but rather as logical errors, violating the inherent constraints of normal images. For instance, permissible objects appear in invalid positions, or required objects are entirely absent, as shown in the first row in Fig. 1. Current structural anomaly detection models, which excel in detecting visual irregularities, are less effective in logical anomaly detection. This is because logical anomalies often involve minimal changes in item composition or position that do not alter the appearance of image significantly but break its logical rules. Compared with the structural anomaly detection models, logical anomaly detection models achieve impressive results by paying more attention to the relationships among objects within the image and their logical consistency with the environment, including the existence of objects and their relative positions [17, 29]. Yet, these logical anomaly detection models overly focus on the relationships between components and overlook potential structural defects within the components themselves, resulting in a decline in structural anomaly detection capability.

Anomaly detection in industrial scenarios requires distinguishing structural anomalies (*e.g.*, cracks) from logical anomalies (*e.g.*, misplaced components). While structural anomalies demand fine-grained geometric analysis, logical anomalies rely on semantic understanding of component relationships. Existing methods focus on either geometric or semantic features, leading to fragmented solutions. UniAD bridges this gap by integrating geometric cues (*e.g.*, texture, shape distortions) and semantic cues (*e.g.*, object relationships, contextual validity). The structural branch employs a structural teacher to encode multi-level local features, while the logical branch leverages a logical teacher to model global component interactions. A student network integrates both perspectives through dual-branch distillation, aligning geometric patterns and semantic constraints. Additionally, the text guidance enhances semantic consistency via textual descriptions. Our contributions are as follows:

- We design a dual-branch teacher-student architecture that integrates geometric and semantic cues through heterogeneous knowledge distillation. It resolves feature conflicts while preserving complementary information, facilitating a unified framework to simultaneously detect structural and logical anomalies.
- To resolve ambiguities in complex logical scenarios, we introduce a text-guided semantic enhancement module. By leveraging a pretrained vision-language model, we generate textual descriptions of component layouts and encode them into semantic features. During testing, text similarity metrics quantify deviations from language-defined logical rules, providing complementary semantic signals beyond visual features.
- On the challenging MVTec LOCO benchmark, our framework achieves 93.7% AUROC for logical anomalies and 93.2% for structural anomalies, surpassing the single-cue models by a clear margin. Ablation studies confirm that integrating both cues is critical to unified anomaly detection.

2 Related Works

Structural Anomaly Detection. Structural anomaly detection primarily focuses on the physical structure of objects or regions in images. One of the commonly used approaches focuses on the feature embedding utilization. These feature embedding-based methods transform raw data into a low-dimensional feature space, where anomaly detection is performed. Among them, multi-dimensional feature extraction with ResNet [16] stands out as a notable example. For instance, in RD [10], the student model learns multi-dimensional features from the teacher model to encode normal features. SimpleNet [31] trains a discriminator capable of distinguishing between normal and pseudo-anomalous features generated through local feature transformations by learning the differences between local features of normal images and those after transformation, ultimately achieving anomaly detection. In these methods, generating suitable feature representations is critical. Improper feature extraction may lead to image information loss and reduced detection accuracy.

Another type of approach, termed image reconstruction-based method, focuses on reconstructing the input image and detecting anomalies by comparing it with the original. Recently, numerous outstanding models have emerged using VAE (Variational Autoencoders) [4, 8, 32, 38] and GAN (Generative Adversarial Networks) [1, 13] to detect anomalies. However, one major drawback of these methods is that they may misclassify anomalies similar to normal cases in structure, increasing false negatives.

Logical Anomaly Detection. Logical anomaly detection mainly focuses on the logical relationships between objects and their surrounding environment within an image. Recently, segmentation-based methods have become a prominent choice for logical anomaly detection as they can effectively reflect the logical relationships between normal samples. Typically, segmentation-based methods begin by dividing training samples into multiple representative components, followed by conducting comparative analyses on each segmented component. Models such as ComAD [29], PCComAD [36], and PSAD [20] fall into this category. ComAD [29] clusters feature maps generated by the pretrained DINO [9] network to extract representative component features, which are then match against the corresponding features of test samples during the test phase to derive anomaly scores. These methods achieve satisfactory performance in detecting obvious missing or replaced items and positions but are less effective in identifying subtler defects in objects.

Text-guided Anomaly Detection In recent years, there has been a growing emergence of anomaly detection methods [14, 19, 22, 26, 41] guided by textual information. Specifically, text-guided industrial image anomaly detection represents an approach that integrates natural language processing (NLP) with computer vision techniques, aiming to leverage textual descriptions to guide or enhance the identification of anomalies in industrial images. For example, LogicAD [19] uses autoregressive multimodal vision-language models (AVLMs) [23, 24, 35] for logical anomaly detection. More specifically, LogicAD integrates AVLMs with format embeddings and a logical reasoning module, enabling it to detect logical anomalies in images using only textual prompts, without requiring additional visual annotations. Furthermore, it provides explanations for the detected anomalies. However, this approach heavily relies on high-quality textual input; if the textual descriptions are inadequate, the performance of the model may be compromised.

3 Method

Our method, UniAD, a dual-branch teacher-student framework that synergizes geometric cues from structural branch and semantic cues from logical branch. As shown in Fig. 2, UniAD comprises four key components: one structural teacher, one logical teacher, one student model, and one text-guided semantic enhancement module. The student model is equipped with two independent branches to distill the geometric and semantic priors from two teachers. We also introduce the text guidance to store component descriptions for semantic verification. During the test phase, we only keep the student model to extract both structural anomaly score and logical error heatmap with the help of the reference text.

3.1 Heterogeneous Feature Distillation

Most existing feature-distillation-based anomaly detection methods tend to rely on a single perspective, either structural or logical, limiting their ability to detect both types of anomalies. This paper aims to design a model capable of handling structural and logical anomalies simultaneously. Traditional single-teacher-student models have limitations in the amount of knowledge that a single teacher model can impart to the student model. Concretely, we design the dual-teacher-single-student architecture that the student model can aggregate knowledge from multiple sources. Our method allows the student model to learn essential cues from the teacher models, each cue tailored to different anomaly detection tasks. Therefore, the key lies in selecting suitable teacher models and integrating them with the student model to efficiently learn their knowledge about geometric and semantic cues.

Selection of the Teacher Models. To comprehensively detect both structural and logical anomalies, we select two distinct teacher models with complementary strengths, driven by the specific nature of these two anomaly types. Specifically, the structural teacher model, dedicated to handling structural anomalies, is structured to produce feature maps containing information at different levels. By integrating these multi-level features, the model can develop detailed and fine-grained local representations as ours geometric cues, similar to models like [31, 36] that use local features to identify structural anomalies. On the other hand, for the logical anomaly detection, we deliberately choose a logical teacher model, capable of capturing holistic global information and emphasizing relationships between components within an image, which represent semantic cues. This choice is inspired by recent studies [29], which underscore the importance of component-level features for identifying logical anomalies. The complementary strengths of these two types of teacher models provide comprehensive learning resources for the student model, then the complementary strengths of dual-teacher is demonstrated by the ablation experiment in Table 6.

Design of the Student Model. After selecting the teacher models, the next step is to design the student model. This necessitates taking into account the heterogeneity between teacher and student models, which can include differences in semantics, dimensions, and feature representations. To bridge these gaps, the student model must be designed with strong feature extraction capability and adaptability. In this paper, we incorporate a dual-branch architecture to enhance the student model's ability to process and integrate heterogeneous knowledge from the teacher models. Concretely, the dual-branch are 1) Structural Branch: the student model extracts multi-scale features to capture local information via learning from the structural teacher. 2) Logical Branch: the student model extracts component-level feature via learning from the logical teacher and generates segmentation maps. We use a shared student model and two separate teachers in two aspects: 1) Decouple structural and logical cues. Task-specific branches ensure that structural and logical cues are effectively decoupled. 2) Computation cost. This shared structure effectively halves the number of parameters and cuts down the inference time. Besides, considering the importance of fine-grained feature encoding and multi-level feature consistency between the student and teacher models, the student model employs an identical backbone as the structural teacher. Table 4 shows how different backbone networks affect performance.

3.2 Structural Branch for Geometric Cue

Structural Feature Distillation. To effectively extract geometric cues through the structural branch, we design a multi-scale feature alignment strategy inspired by the knowledge distillation methods [39]. Specifically, we select multi-level features generated by the structural teacher model, including the low-level multi-dimensional features g_{res}, high-level features g_{str}, for the student model to learn, as shown in Fig. 2. The goal is to align the output of student model more closely with that of the structural teacher model by learning and integrating multi-level features, enhancing its ability to obtain fine-grained geometric cues and detect structural anomalies. This choice avoids learning biases that can result from relying on a single type of feature. The experimental results of selecting different

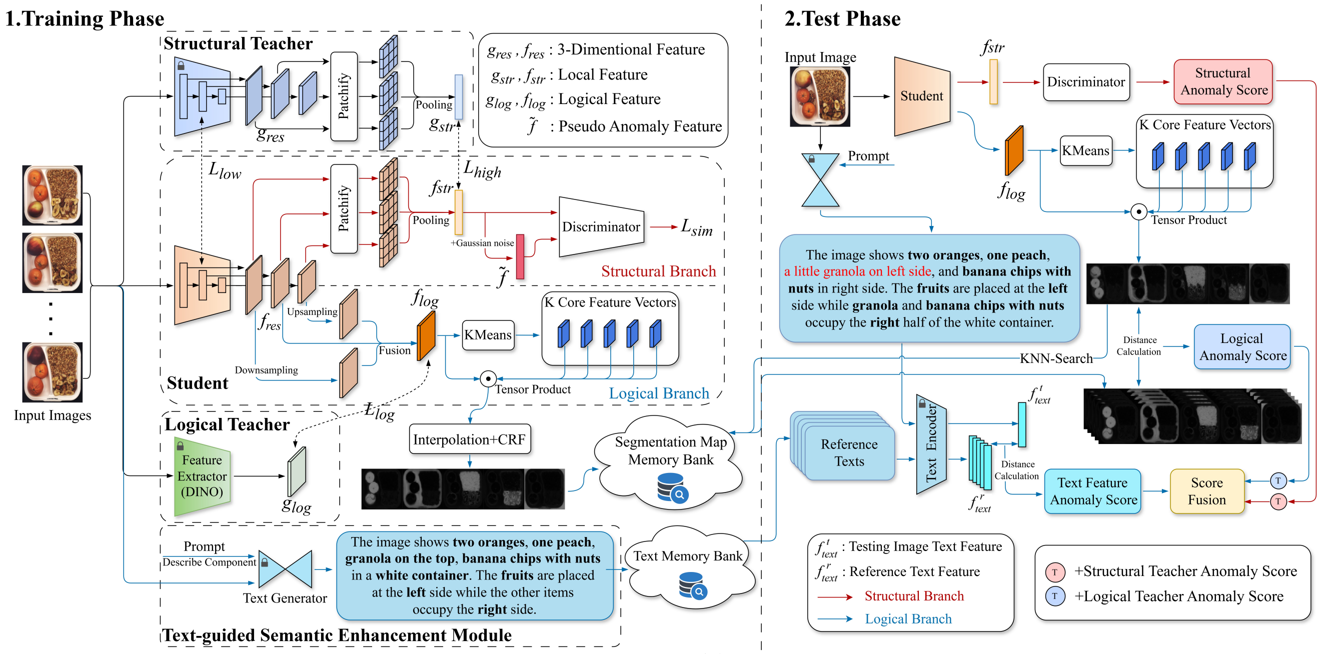

1.Training Phase
Structural Teacher
Patchify
Pooling
g_{res}
g_{str}
g_{res} , f_{res} : 3-Dimentional Feature
g_{str} , f_{str} : Local Feature
g_{log} , f_{log} : Logical Feature
$\tilde{f}$: Pseudo Anomaly Feature
L_{low}
L_{high}
f_{str}
Patchify
Pooling
+Gaussian noise
$\tilde{f}$
Discriminator
L_{sim}
Structural Branch
Input Images
f_{res}
Upsampling
Downsampling
Fusion
f_{log}
KMeans
K Core Feature Vectors
Tensor Product
Logical Branch
Student
L_{log}
Logical Teacher
Feature Extractor (DINO)
g_{log}
Interpolation+CRF
Segmentation Map Memory Bank
Prompt
Describe Component
Text Generator
The image shows two oranges, one peach, granola on the top, banana chips with nuts in a white container. The fruits are placed at the left side while the other items occupy the right side.
Text Memory Bank
Text-guided Semantic Enhancement Module
2.Test Phase
Input Image
Student
f_{str}
Discriminator
Structural Anomaly Score
Prompt
f_{log}
KMeans
K Core Feature Vectors
Tensor Product
The image shows two oranges, one peach, a little granola on left side, and banana chips with nuts in right side. The fruits are placed at the left side while granola and banana chips with nuts occupy the right half of the white container.
KNN-Search
Distance Calculation
Logical Anomaly Score
Reference Texts
Text Encoder
f^t_{text}
f^r_{text}
Distance Calculation
Text Feature Anomaly Score
Score Fusion
T
T
f^t_{text} : Testing Image Text Feature
f^r_{text} : Reference Text Feature
Structural Branch
Logical Branch
T +Structural Teacher Anomaly Score
T +Logical Teacher Anomaly Score

Figure 2: Overview of the training and test phases of our method. (1) During the training phase, the student learns geometric cues from the structural teacher via multi-level alignment (L_{high}, L_{low}) along the red arrows, while acquiring semantic cues from the logical teacher and text-guided semantic enhancement module along the blue arrows. (2) During testing, anomaly scores from both branches are fused to achieve unified detection.

features for distillation are shown in the ablation study in Sec. 4.3. The generation process of features g_{res} and g_{str} is outlined below.

For a given image x_i, we select m outputs of different layers as the multi-dimensional features. For the teacher model and student model, suppose the lth output is $g^l_{res}(x_i)$ or $f^l_{res}(x_i) \in \mathbb{R}^{H_l \times W_l \times C_l}$ (l=1, 2, . . . , m), where H_l, W_l, and C_l are the height, width, and channel size of the output. To learn knowledge from the structural teacher, the student undergoes the steps in Fig. 2 as follows:

1) Feature Patchifying and Pooling: For each entry $g^{l,i}_{res}(x_i)^{(h,w)} \in \mathbb{R}^{C_l}$ at location (h, w), its patchified feature set $\mathcal{N}^{(h,w)}_p$ with a patch-size p is constructed as

$$\mathcal{N}^{(h,w)}_p = \{(h', y')|h' \in [h - \lfloor p/2 \rfloor, ..., h + \lfloor p/2 \rfloor], \\ y' \in [w - \lfloor p/2 \rfloor, ..., w + \lfloor p/2 \rfloor]\}. \quad (1)$$

An adaptive average pooling aggregation function F_{agg} is applied to the patchified features in $\mathcal{N}_p$ to obtain aggregated features $z^{l,i}_t$ for each layer, *i.e.*, $z^{l,i}_t = F_{agg}\left(\left\{(g^{l,i}_{res}(x_i))^{(h',y')} \mid (h', y') \in \mathcal{N}^{(h,w)}_p\right\}\right)$.

2) Bilinear Interpolation: As different aggregated feature $z^{l,i}_t$ above may have diverse dimensions, this can pose challenges to their integration. For better feature alignments, we adopt a bilinear interpolation $BI(\cdot)$ to transform these features to a uniform size (H_0, W_0). Then, we concatenate them along the channel dimension to form the final high-level feature g^i_{str}.

$$g^i_{str} = F_{cat}(BI(z^{l,i}_t, (H_0, W_0))|l \in \{1, \dots, m\}). \quad (2)$$

After obtaining g_{str} for the teacher model, we generate the corresponding feature f_{str} for the student model using the same feature processing approach. Then, we define two feature distillation losses L_{high} and L_{low} to ensure that the student model learns both low-level multi-dimensional features and high-level features. The structural student network is trained with MSE loss.

$$L_{low} = \frac{1}{N} \sum_{l \in \{1,\dots,m\}} \sum_{x_i \in X_{train}} (g^l_{res}(x_i) - f^l_{res}(x_i))^2, \quad (3)$$

$$L_{high} = \frac{1}{N} \sum_{x_i \in X_{train}} (g^i_{str} - f^i_{str})^2, \quad (4)$$

where g_{res} and f_{res} are the multi-dimensional features of the structural teacher model and the student model, g^i_{str} and f^i_{str} represent the high-level features of the ith input image of the structural teacher model and the student model. N denotes the total number of samples in the training set.

The total loss for structural feature distillation is given by

$$L_{str} = L_{high} + L_{low}. \quad (5)$$

After obtaining the high-level structural features f_{str}, inspired by [31], we utilize a feature adapter-a fully connected layer-to transform f_{str} into the target domain feature f_{str}, thereby reducing domain bias. Subsequently, we construct pseudo-anomalous counterparts $\tilde{f}$ by applying Gaussian noise to f_{str}. To further enhance the structural anomaly detection capability in distinguishing structural anomalies, especially those with complex structures, we integrate a structural anomaly discriminator D based on a 2-layer MLP, estimating normality as $D(f_{str}) \in \mathbb{R}$. This discriminator takes both the target domain features f_{str} and their pseudo-anomalous counterparts $\tilde{f}$ as its inputs, trained with a truncated l_1 loss as

$$l^{h,w}_i = \max(0, th^+ - D(f^{h,w}_{str})) + \max(0, -th^- + D(\tilde{f}^{h,w}_{str})), \quad (6)$$

where th^+ and th^- are truncation terms preventing overfitting. They are set to 0.5 and -0.5 by default. The training objective is L_{sim}, aiming to boost the sensitivity of model in detecting anomalies.

$$L_{sim} = \min \sum_{x^i \in \mathcal{X}_{train}} \sum_{h,w} \frac{l_i^{h,w}}{H_0 \times W_0}. \tag{7}$$

In the inference stage, the discriminator directly outputs the structural anomaly score.

3.3 Logical Branch for Semantic Cue

Logical Feature Distillation. To capture semantic cues through the logical branch, we propose a component-aware feature alignment strategy via multi-scale feature fusion. With structural feature distillation above, the student model can accurately capture fine-grained structural anomalies. However, they may sometimes lack the necessary semantic depth. The logical models, on the other hand, excel in producing features that are semantically rich, especially in their logical components. Given this contrast, it becomes imperative for the student model to acquire the component-level features from the logical teacher. Nevertheless, due to differences in network design, the features generated by the student model may not match the dimensions of the component-level features extracted by the logical teacher model. To address this issue, inspired by [12], we impose the multi-scale feature fusion techniques for dimension matching. Concretely, for the multi-dimensional feature outputs $f_{res}^l \in \mathbb{R}^{H_l \times W_l \times C_l} (l = 1, \dots, m)$ of the student model, we employ a transformation function to synchronize these various features with the corresponding features of the logical teacher.

$$\phi(f_{res}^l) = \begin{cases} \text{Downsample}(f_{res}^l) & \text{if } D_s^l > D_t^l \\ \text{Upsample}(f_{res}^l) & \text{if } D_s^l < D_t^l \\ f_{res}^l & \text{if } D_s^l = D_t^l \end{cases}, \tag{8}$$

where D_s^l and D_t^l denote the feature dimensions of layer l in the student model and the teacher model, respectively.

Finally, we obtain m dimensionally consistent features, concatenated to form a new feature f_{log} as

$$f_{log} = F_{cat}(\phi(f_{res}^1), \phi(f_{res}^2), \dots, \phi(f_{res}^m)). \tag{9}$$

After feature alignments, we define the following logical distillation loss L_{log} using MSE to train the student network.

$$L_{log} = \frac{1}{N} \sum_{x_i \in \mathcal{X}_{train}} (g_{log}(x_i) - f_{log}(x_i))^2, \tag{10}$$

Through logical feature distillation, the student model learns semantic cues from the teacher while bridging structural differences. To facilitate the performance of model in detecting logical anomaly, representative component-level features are further extracted from the learned features of student model. Specifically, for all channels of f_{log}, KMeans clustering is applied to f_{log} to obtain K clusters, as shown in Fig. 2. The cosine similarity between each cluster and f_{log} is then calculated to generate an initial segmentation map. Subsequently, the segmentation map is resized to the original image dimension via interpolation and refined with a fully connected Gaussian Conditional Random Field (CRF) [21] for post-processing and is stored in a memory bank for future testing.

Text-guided Semantic Enhancement Module. In the field of industrial image anomaly detection, traditional methods have predominantly relied on pure vision-based models. The introduction of a text generator enables multimodal information fusion, which involves joint modeling of images and text. By employing a text generator to produce semantic descriptions of image content, additional contextual information beyond visual features can be incorporated, thereby enhancing the accuracy of anomaly detection.

Text-guided Semantic Enhancement Module consists of two components: text generation and the calculation of anomaly scores based on text features. The detailed steps for calculating anomaly scores from textual features will be described in Sec. 3.4. Following this, we will delve into the specifics of the text generation process.

To encode semantic relationships beyond visual features, we leverage textual descriptions generated by VLLM (visual-language large model). These descriptions explicitly capture component-level semantics, enabling the model to compare test images against language-defined logical rules. Concretely, for all images x_i from the training set $\mathcal{X}_{train}$ and the test set $\mathcal{X}_{test}$, inputting x_i into the pre-trained model, along with prompts that guide the description of components and their positions, enables the generation of a textual description. This process can be represented as

$$text_i = VLLM((x_i, Prompt)|x_i \in \mathcal{X}_{train} \cup \mathcal{X}_{test}), \tag{11}$$

where $VLLM(\cdot)$ denotes visual-language large model. In practice, we deploy Qwen-VL [2] pre-trained model. Then the generated texts are stored in the memory bank for future testing.

3.4 Synergistic Anomaly Score

Computing Anomaly Score. After training, the student model can generate multi-level features and component-level features for structural anomaly and logical anomaly respectively. Leveraging these distinct features, we can generate corresponding anomaly scores: structural anomaly score S_{str} and logical anomaly score S_{log}. The S_{str}^t and S_{str}^s are directly derived from the teachers’ and students’ output of the structural anomaly discriminator respectively. The S_{log}^t and S_{log}^s, on the other hand, are computed by comparing the component feature segmentation maps of the training images and testing images following [29]. Concretely, we first convert the component-level segmentation maps into three types of features: **1)** Area features are calculated by summing pixels counts within segmented regions. **2)** Color features are extracted by converting RGB images to CILAB space, which includes three components: L (lightness), a (green-red), and b (blue-yellow). For each pixel, the lightness component is ignored, and the ratio b/a is calculated, with the final color feature being the average value across the region. **3)** Quantity features are derived by grouping regions using DBSCAN [11] and calculating their density. Then, by combining the above three features, the logical anomaly score is derived from the average L_2 distance between the test image and its 5-nearest neighbors $\mathcal{N}_k$ $(k = 5)$.

Meanwhile, the text feature anomaly score is calculated. To bridge the gap between language descriptions and anomaly quantification, we employ a pretrained text encoder. Textual descriptions generated by Qwen-VL contain rich logical rules. The text encoder maps these descriptions into dense vectors in a shared semantic

Table 1: Comparison of our model and existing methods on MVTec LOCO AD. Results are given as logical anomalies/structural anomalies. The highest and second-highest scores are highlighted in bold and underlined, respectively.

Metric	Category	LogicAD [19]	SLSG [40]	PatchCore [36]	GCAD [5]	SAM-LAD [34]	EfficientAD-S [3]	DSKD [46]	**Ours**
AUROC	Breakfast Box	93.1/-	-	80.0/75.2	87.0/80.9	96.7/85.2	-	86.4/82.3	95.1/86.3
	Juice Bottle	81.6/-	-	92.3/97.8	100/98.9	98.7/96.5	-	99.1/98.9	99.5/98.5
	Pushpins	98.1/-	-	73.8/81.9	97.5/74.9	97.2/79.2	-	75.8/89.2	96.0/97.6
	Screw Bag	83.8/-	-	55.7/88.6	56.0/70.5	95.2/77.9	-	63.1/88.7	83.4/88.1
	Splicing Connectors	73.4/-	-	75.6/94.9	89.7/78.3	91.4/88.6	-	91.2/92.5	94.4/95.4
	Average	86.0/81.5 83.8	89.6/91.4 90.3	75.5/87.7 81.6	86.0/80.7 83.4	**95.8**/85.5 90.7	90.0	83.1/90.3 86.7	93.7/**93.2** **93.4**
sPRO	Breakfast Box	-	-	-	50.2	81.9/79.1	-	56.8	80.5/82.1
	Juice Bottle	-	-	-	91.0	94.4/93.5	-	86.5	94.6/95.2
	Pushpins	-	-	-	73.9	76.2/74.2	-	82.5	75.1/85.6
	Screw Bag	-	-	-	55.8	86.3/71.6	-	62.7	72.3/78.6
	Splicing Connectors	-	-	-	79.8	89.1/85.2	-	76.7	92.0/90.3
	Average	-	-	-	70.1	**85.6**/80.7 83.2	77.8	73.0	82.9/**86.4** **84.6**

space, enabling quantitative comparison between test images and normal reference images. Specifically, for the text $text_i$ corresponding to a test image $x_i \in \mathcal{X}_{test}$, extracting the text feature $f^{t,i}_{text}$ and then the text features of the nearest five neighbors based on the segmentation map are calculated as reference features $f^{r,k}_{text} \in \mathcal{N}_k$.

$$f^{t,i}_{text} = F_{text}((text_i)),\ f^{r,k}_{text} = F_{text}((text_k)|text_k \in \mathcal{N}_k), \quad (12)$$

where $F_{text}(\cdot)$ uses the Long-CLIP [45] pre-trained model. The text feature anomaly score S_{text} is defined as

$$S_{text} = 1 - \frac{1}{k}\sum_{j=1}^{k} sim(f^{t,i}_{text}, f^{r,k}_{text}), \quad (13)$$

where $sim(\cdot,\cdot)$ represents cosine similarity. The more similar the text features of the test image are to the reference text features, the lower the anomaly score.

Anomaly Score Fusion. The above-mentioned scores are then combined to obtain final anomaly detection results. Concretely, the structural anomaly score S_{str} is derived by summing the scores from the teacher S^t_{str} and student model S^s_{str}. The same approach is applied to obtain the logical anomaly score S_{log}.

$$S_{str} = S^t_{str} + S^s_{str},\ \ S_{log} = S^t_{log} + S^s_{log}. \quad (14)$$

However, since S_{str}, S_{log}, and S_{text} are derived from different feature spaces and distinct methodologies, directly summing them may compromise the anomaly detection performance. To ensure balanced and meaningful fusion, normalization of these scores is necessary. The normalized anomaly score $\tilde{S}$ is calculated as $\tilde{S} = (S - \mu)/\sigma$, where S represents the original anomaly score, μ and σ denote the mean and standard deviation, respectively. The final total anomaly score $\tilde{S}_{total}$ is then obtained by summing the normalized anomaly score $\tilde{S}_{str}$, $\tilde{S}_{log}$, and $\tilde{S}_{text}$.

$$\tilde{S}_{total} = \tilde{S}_{str} + \tilde{S}_{log} + \tilde{S}_{text}. \quad (15)$$

4 Experiment

4.1 Implementation Details

Following the analysis in Sec. 3.1, we adopt WideResNet50 [16] as backbone for structural teacher and student while selecting DINO ViT-S/8 [9] as backbone for logical teacher. In the feature distillation, the dimension of the feature m is set to 3. We train our student model using the Adam optimizer with a learning rate of 0.05 for 500 iterations. For the structural branch, the structural teacher model employs a widely used feature extractor [36] to generate high-level features and the feature dimension from the feature extractor is set to 1536. The dimensions of the input and output features for the FC layer are the same. The $\tilde{f}$ is obtained by adding i.i.d. Gaussian noise $N(0, \sigma^2)$ to each entry of normal features. σ is set to 0.015 by default. The subsequent discriminator composes of a linear layer, a batch normalization layer, a leaky relu (0.2 slope), and a linear layer. The Adam optimizer is used, setting the learning rate for discriminator to 2e-4, and weight decay to 1e-5. Training epochs is set to 160 for each dataset and batchsize is 4. For the logical branch, we resize all the images into 224 × 224 and the cluster number K is set to 5. For the text-guided semantic enhancement module, we used a text prompt-"Describe all components and their spatial relationships in the image" and the size of Qwen-VL is 32B.

Table 2: Comparison of our model and existing methods on VisA and MVTec-AD with AUROC.

Method	VisA	MVTec-AD (Texture)	MVTec-AD (Object)
DSR [44]	91.8	99.2	97.1
PromptAD [26]	89.1	96.6	96.6
HGAD [43]	93.5	**99.8**	97.7
MoEAD [33]	-	99.4	96.9
PatchCore [36]	-	99.0	99.2
Ours	**96.9**	99.6	**99.3**

Table 3: Efficiency comparison of different methods on MVTec LOCO AD.

Method	Params	FLOPs	Latency(ms)	AUROC
PatchCore [36]	-	-	47.1	81.6
DADF [42]	298.5 M	149.3 G	-	83.7
SAM-LAD [34]	-	94.2 G	-	91.8
Ours	225.6 M	89.9 G	19.6	**93.4**

We compare our model with LogicAD [19], SLSG [40], PatchCore [36], GCAD [5], SAM-LAD [34], EfficientAD-S [3], DSKD [46], DSR [44], PromptAD [26], HGAD [43], and MoEAD [33] on MVTec

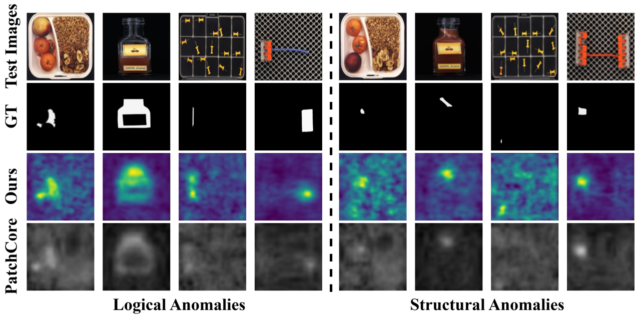


Figure 3: Qualitative analysis on MVTec LOCO AD with logical and structural anomalies.

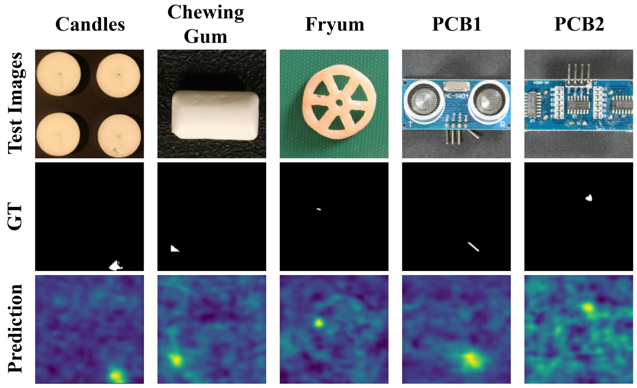


Figure 4: Qualitative analysis on VisA with only structural anomalies.

LOCO AD [5], VisA [48], and MVTec-AD [6] datasets. Consistent with previous studies, we adopt the Area Under the Receiver Operating Characteristic (AUROC) and saturated Per-Region Overlap (sPRO) as the primary quantitative evaluation metric.

4.2 Quantitative and Qualitative Results

The anomaly detection results on MVTec LOCO AD are shown in Table 1, where the image-level anomaly scores are calculated using (15). Across both structural and logical anomaly detection, our method achieves the highest overall performance. Our model shows significantly higher **average** anomaly detection performance with AUROC, *i.e.,* **surpassing** the second-best method by **2.7%**. Similarly, in terms of performance under the sPRO metric, our method outperforms the second-best method by 1.4%, showcasing its effectiveness and precision across various anomaly types. Moreover, as shown in Table 3, our model not only achieves superior performance but also demonstrates a significant advantage in terms of efficiency.

For models designed to detect logical anomalies, such as [5], our model not only outperforms it in detecting structural anomalies but also surpasses it in identifying logical anomalies. Similarly, our model shows superior performance compared to models like [36], which exhibits stronger capabilities in detecting structural anomalies than logical ones. Moreover, our method presents superior performance when compared to [19], which focuses solely on logical anomaly metrics in text-guided approaches.

Representative samples and the corresponding anomaly localization results are visualized in Fig. 3. We observe that our method effectively detects logical anomalies. For instance, in the "juice bottle" category, the juice color is quite similar to the label, making it difficult to distinguish between them. Despite this, our model accurately identifies the incorrectly assigned label position. Similarly, in the "pushpins" category, the grids where the pushpins are placed are extremely easily confused with the background. However, our model still effectively detects violations of the logical constraintwhere each grid should contain only one pushpin. These results underscore the ability of our model for logical anomaly detection, even in challenging scenarios. For structural anomalies, in many cases, the defects are extremely small and difficult to detect. However, our model accurately pinpoints these smaller and localized defects, such as missing labels, contaminated pushpins, and broken screws. These results highlight the capability of our model to detect and locate a wide range of structural anomalies.

Table 4: AUROC performance of different backbones for student model on MVTec LOCO AD in logical anomaly (LA) and structural anomaly (SA) detection.

Backbone	LA	SA	Avg
DINO	74.2	85.4	79.8
WideResNet50	**93.7**	**93.2**	**93.4**

The detection results on the ViSA and MVTec-AD datasets are shown in Table 2. On these datasets, which contains only structural anomalies, our model demonstrates commendable performance compared to models [36, 44] specifically designed to address structural anomalies. Furthermore, our method also exhibits notable advantages over [26], a text-guided method, in detecting structural anomalies. In the context of detecting anomalies within this dataset, the logical branch, which excels at identifying logical inconsistencies, plays a diminished role due to the absence of logical anomalies in the dataset. Consequently, the structural branch of our model, which is adept at detecting structural anomalies, becomes more pivotal. As illustrated in Fig. 4, our model can effectively identify structural defects within samples, such as missing corners or notches in objects, or contaminated areas on PCBs.

4.3 Ablation Studies and Further Discussion

Ablation Study on Different Student Backbones. We test the student model with different backbones, and the results are shown in Table 4. We find that among the tested architectures, WideResNet50 manifests superior performance compared to the DINO backbone. This is because, when handling structural branches, it is crucial to focus on the precise capture of geometric cues. However, DINO, as a network suited for extracting semantic cues, tends to perform suboptimally in this regard. On the other hand, WideResNet50, by stacking multiple convolutional layers, excels at extracting geometric cues from images, making it more effective in addressing structural anomalies. Furthermore, its deep architecture provides multi-level feature representations, which enable flexible learning of feature representations from other networks when combined with multi-scale feature fusion techniques. This enhances its capability to detect logical anomalies.

Ablation Study on Different Feature Distillations. We configure the student model to learn different features, and the corresponding results are shown in Table 5. We observe that selecting

Table 5: AUROC performance of different feature distillation on MVTec LOCO AD.

Low-level Feature	High-level Feature	LA	SA	Avg
✓		92.7	92.9	92.8
	✓	92.2	92.4	92.3
✓	✓	**93.7**	**93.2**	**93.4**

Table 6: AUROC performance of different teachers on MVTec LOCO AD.

Structural Teacher	Logical Teacher	LA	SA	Avg
✓		82.3	92.1	87.2
	✓	90.6	75.3	83.0
✓	✓	**93.7**	**93.2**	**93.4**

Table 7: AUROC performance of different VLLMs on MVTec LOCO AD. SB and LB denote structural and logical branches respectively.

VLLM	LA	SA	Avg
Without	93.0	**93.2**	93.1
GPT-4o (only LB)	93.5	93.1	93.3
Qwen-VL (SB & LB)	93.2	93.0	93.1
Qwen-VL (only LB)	**93.7**	**93.2**	**93.4**

Table 8: AUROC performance on MVTec LOCO AD with Dual-branch methods. SFD, LFD, and TSE represent Structural Feature Distillation, Logical Feature Distillation, and Text-guided Semantic Enhancement module, respectively.

Structural Branch	Logical Branch		LA	SA	Avg
SFD	LFD	TSE			
✓			83.4	92.3	87.9
	✓		88.8	75.5	82.2
	✓	✓	91.2	75.6	83.4
✓	✓		93.0	**93.2**	93.1
✓	✓	✓	**93.7**	**93.2**	**93.4**

Table 9: AUROC performance of different anomaly score fusions on MVTec LOCO AD.

S^t_{str}	S^t_{log}	S^s_{str}	S^s_{log}	LA	SA	Avg
✓				82.5	83.9	83.2
	✓			92.2	73.9	83.1
		✓		89.0	83.3	86.2
			✓	92.5	73.5	83.0
✓	✓			92.9	92.6	92.8
		✓	✓	93.1	85.8	89.5
✓	✓	✓	✓	**93.7**	**93.2**	**93.4**

both low-level multi-dimensional features and high-level features for feature distillation yields the best performance, outperforming single-feature approaches. Combining both feature types helps the model more effective capture of geometric cues, thereby improving its capability in detecting structural anomalies. If only single feature is selected for learning, such as high-level features, the student model may fail to learn finer-grained features, leading to incomplete learning and reduced detection performance.

Ablation Study on Different Teachers. The experimental results in Table 6 illustrate the impact of the different teacher models on the overall architecture. The results from experiments using either a single structural teacher or a single logical teacher validate their respective specificity in addressing different types of anomalies. However, when both teacher models are incorporated simultaneously, optimization across all anomaly types is achieved, thereby validating the effectiveness of the dual-teacher design.

Ablation Study on Different VLLMs. The experimental results in Table 7 indicate that using descriptions from different large models consistently improves the average anomaly detection performance, demonstrating that our method is VLLM-agnostic. The third and fourth lines of Table 7 show that applying text guidance to the structural branch slightly degrades the detection performance by 0.3% compared to the case without text guidance. The reason may be that the structural anomaly often exhibits phisical changes, which are not easily described in text without specific technical vocabulary or detailed context. In contrast, the logical anomalies often involve violations of predefined rules, which contains rich semantic information. Text guidance can exploit these rules with additional semantic context, making it crucial for the logical branch.

Ablation Study on Different Branches. The experimental results in Table 8 validate the effectiveness of different branches in handling various types of anomalies. When only the structural branch is employed, the performance in detecting structural anomalies surpasses that of detecting logical anomalies compared to experiments using only the logical branch. Conversely, when relying solely on the logical branch, the detection of logical anomalies achieves superior results. Additionally, the experiments in the second and third rows, as well as the fourth and fifth rows, highlight the efficacy of the text-guided semantic enhancement module. Finally, the fifth-row experiment, which integrates the geometric and semantic cues of both branches, achieves the optimal overall performance, thereby validating the effectiveness of the dual-branch design.

Ablation Study on Different Anomaly Score Fusions. During testing, different score fusion methods can affect the final detection performance. We fuse anomaly scores in different combinations and present the results in Table 9. We find that when all anomaly scores are fused (shown in the last row), the detection metrics reach their highest levels. This indicates that combining scores of students and teachers can yield optimal detection performance. This shows that different anomaly scores complement each other, enhancing the overall detection capability of model when combined.

5 Conclusion

In this paper, we propose a new image anomaly detection framework, UniAD, which unifies both structural and logical anomaly detection tasks through dual-branch teacher-student network. Geometric cues localize structural defects via multi-level feature alignment, while semantic cues enforce logical constraints through component-aware segmentation maps and text-guided reasoning. The synergistic integration of both teachers enables the student to simultaneously resolve both types of anomalies. Experimental results showcase competitive accuracy improvements in both structural and logical anomaly detection tasks.

Acknowledgments

We acknowledge the support from the Natural Science Foundation of Xiamen (3502Z202473071), the National Natural Science Foundation of Fujian Province (Grant Nos.2023J011428, 2022Y0077), the University of Macau (SRG2024-00002-FST), the Macau Science and Technology Development Fund (FDCT/0043/2025/RIA1), and the Industry-University-Research Collaborative Innovation Project of Fujian Province (Grant No.2024H6035).

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. 2019. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 622–637.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).

[3] Kilian Batzner, Lars Heckler, and Rebecca König. 2024. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 128–138.

[4] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer, 161–169.

[5] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. 2022. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision* 130, 4 (2022), 947–969.

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

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

[8] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. 2018. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011* (2018).

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[10] Hanqiu Deng and Xingyu Li. 2022. Anomaly Detection via Reverse Distillation From One-Class Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9737–9746.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. Density-based spatial clustering of applications with noise. In *Int. Conf. knowledge discovery and data mining*, Vol. 240.

[12] Yuan Gong, Sameer Khurana, Andrew Rouditchenko, and James Glass. 2022. Cmkd: Cnn/transformer-based cross-model knowledge distillation for audio classification. *arXiv preprint arXiv:2203.06760* (2022).

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[14] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. 2024. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2041–2049.

[15] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 8472–8480.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[17] Yu-Hsuan Hsieh and Shang-Hong Lai. 2024. CSAD: Unsupervised Component Segmentation for Logical Anomaly Detection. arXiv:2408.15628 [cs.CV] https://arxiv.org/abs/2408.15628

[18] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11375–11385.

[19] Er Jin, Qihui Feng, Yongli Mou, Stefan Decker, Gerhard Lakemeyer, Oliver Simons, and Johannes Stegmaier. 2025. LogicAD: Explainable Anomaly Detection via VLM-based Text Feature Extraction. *arXiv preprint arXiv:2501.01767* (2025).

[20] Soopil Kim, Sion An, Philip Chikontwe, Myeongkyun Kang, Ehsan Adeli, Kilian M Pohl, and Sang Hyun Park. 2024. Few Shot Part Segmentation Reveals Compositional Logic for Industrial Anomaly Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8591–8599. doi:10.1609/aaai.v38i8.28703

[21] John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, Vol. 1. Williamstown, MA, 3.

[22] Mingyu Lee and Jongwon Choi. 2024. Text-Guided Variational Image Generation for Industrial Anomaly Detection and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26519–26528.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

[25] Wujin Li, Jiawei Zhan, Jinbao Wang, Bizhong Xia, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Feng Zheng. 2022. Towards continual adaptation in industrial anomaly detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2871–2880.

[26] Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. 2024. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1093–1102.

[27] Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. 2023. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070* (2023).

[28] Tongkun Liu, Bing Li, Xiao Du, Bingke Jiang, Leqi Geng, Feiyang Wang, and Zhuo Zhao. 2023. Fair: Frequency-aware image restoration for industrial visual anomaly detection. *arXiv preprint arXiv:2309.07068* (2023).

[29] Tongkun Liu, Bing Li, Xiao Du, Bingke Jiang, Xiao Jin, Liuyi Jin, and Zhuo Zhao. 2023. Component-aware anomaly detection framework for adjustable and logical industrial visual inspection. *Advanced Engineering Informatics* 58 (2023), 102161. doi:10.1016/j.aei.2023.102161

[30] Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. 2024. Dual-modeling decouple distillation for unsupervised anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5035–5044.

[31] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. 2023. Simplenet: A Simple Network for Image Anomaly Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20402–20411.

[32] Wei Luo, Haiming Yao, and Wenyong Yu. 2023. Normal reference attention and defective feature perception network for surface defect detection. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–14.

[33] Shiyuan Meng, Wenchao Meng, Qihang Zhou, Shizhong Li, Weiye Hou, and Shibo He. 2024. MoEAD: A Parameter-Efficient Model for Multi-class Anomaly Detection. In *European Conference on Computer Vision*. Springer, 345–361.

[34] Yun Peng, Xiao Lin, Nachuan Ma, Jiayuan Du, Chuangwei Liu, Chengju Liu, and Qijun Chen. 2025. Sam-lad: Segment anything model meets zero-shot logic anomaly detection. *Knowledge-Based Systems* 314 (2025), 113176.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[36] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14318–14328.

[37] Long Tian, Hongyi Zhao, Ruiying Lu, Rongrong Wang, Yujie Wu, Liming Wang, Xiongpeng He, and Xiyang Liu. 2024. FOCT: Few-shot Industrial Anomaly Detection with Foreground-aware Online Conditional Transport. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6241–6249.

[38] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. 2020. Attention guided anomaly localization in images. In *European Conference on Computer Vision*. Springer, 485–503.

[39] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. 2021. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257* (2021).

[40] Minghui Yang, Jing Liu, Zhiwei Yang, and Zhaoyang Wu. 2024. SLSG: Industrial image anomaly detection with improved feature embeddings and one-class classification. *Pattern Recognition* 156 (2024), 110862. doi:10.1016/j.patcog.2024.110862

[41] Shuyu Yang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2025. Beyond Walking: A Large-Scale Image-Text Benchmark for Text-based Person Anomaly Search. In *ICCV*.

[42] Haiming Yao, Wei Luo, and Wenyong Yu. 2023. Visual anomaly detection via dual-attention transformer and discriminative flow. *arXiv preprint arXiv:2303.17882* (2023).

[43] Xincheng Yao, Ruoqi Li, Zefeng Qian, Lu Wang, and Chongyang Zhang. 2024. Hierarchical gaussian mixture normalizing flow modeling for unified anomaly detection. In *European Conference on Computer Vision*. Springer, 92–108.

[44] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. 2022. Dsr–a dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*. Springer, 539–554.

[45] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-CLIP: Unlocking the Long-Text Capability of CLIP. *arXiv preprint arXiv:2403.15378* (2024).

[46] Jie Zhang, Masanori Suganuma, and Takayuki Okatani. 2024. Contextual affinity distillation for image anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 149–158.

[47] Kang Zhou, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Zaiwang Gu, Jiang Liu, and Shenghua Gao. 2020. Encoding structure-texture relation with p-net for anomaly detection in retinal images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 360–377.

[48] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. 2022. SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation. *arXiv preprint arXiv:2207.14315* (2022).