This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# LayoutAD: Exploring Semantic-Geometric Misalignment Reasoning for Scene Layout Anomaly Detection

Zhichao Zeng, Jiasheng Zhang, Jiyun Sun, Jiangtao Cui, Xiaotian Qiao*
School of Computer Science and Technology, Xidian University

## Abstract

*Visual anomaly detection is vital for quality control applications by identifying deviations from normal patterns. Previous structural or logical anomaly detection methods mainly focus on pixel-level deviations like texture defects and reconstruction errors, ignoring the object-level structural and contextual inconsistencies. These overlooked layout anomalies remain critical yet underexplored,* e.g.*, factually defective hallucinations appeared in generative text-to-image models. Based on the above observation, in this paper, we introduce scene layout anomaly detection, a new task that predicts an object-level anomaly map from the input image to reveal the semantic plausibility and geometric consistency of each object in the scene. Specifically, we propose LayoutAD, an unsupervised learning framework that constructs semantic and geometric graphs to jointly reason over semantic-geometric misalignment among objects. Under this formulation, we are able to detect diverse layout deviations, including object attribute implausibilities and relationship mismatches. Extensive experiments show that LayoutAD outperforms baselines qualitatively and quantitatively across various scenarios, benefiting scene understanding and generation applications like video anomaly detection and self-corrected image generation.*

## 1. Introduction

Anomaly detection, which aims to identify unexpected pattern deviations from normal data, is of great importance in many applications ranging from industrial inspection [2, 33, 53] to autonomous navigation [43]. While recent years have witnessed an emergent interest and great success in this field, visual anomaly detection is still non-trivial, due to diverse data modalities and complex anomaly types.

Towards this objective, structural [29, 41, 52] or logical anomaly detection [1, 3] have been proposed, focusing on pixel-level deviations with a simple background under industrial or medical settings. For object-level deviations in natural scenes, while there has been a line of research on anomaly segmentation [24, 26], only the basic object-background anomalous relationship has been studied yet. Note that existing anomaly detection methods largely ignore diverse implausible object placements or relational inconsistencies. These overlooked layout anomalies are invisible to pixel-level detectors but critical for downstream applications like text-to-image generation. Take Figure 1 as an example. Vanilla models remain vulnerable to these layout anomalies, failing to correct object attribute (*e.g.*, a dog with 5 legs) or spatial relation (*e.g.*, a car on the lake) hallucinations. The above issues motivate us to raise a fundamental question: *Can machines detect high-level scene anomalies, i.e., what objects should exist, where they belong, and how they interact?*

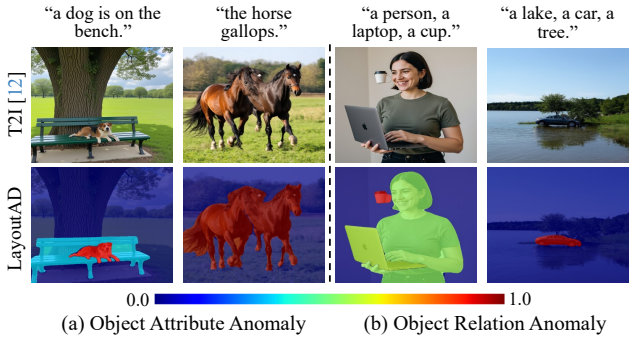


Figure 1. Scene layout anomaly detection. Given an input image (*e.g.*, generated by Stable Diffusion 3 [12]), *LayoutAD* can accurately detect diverse layout anomalies to reveal implausible object attributes (a) or relation inconsistencies (b) in the scene.

In this paper, we take a step towards the above question and introduce the new task of *scene layout anomaly detection*. As shown in Figure 1, given an input image, our goal is to predict an anomaly map that indicates semantic or geometric deviations among objects in the layout space. Note that this new problem is fundamentally different from either visual anomaly detection or hallucination detection.

* Corresponding author: Xiaotian Qiao.

On one hand, visual anomaly detection focuses primarily on low-level, pixel-wise deviations, while largely overlooking object-level structural and contextual inconsistencies. On the other hand, existing hallucination detection methods are inherently prompt-conditioned by measuring misalignment between the generated image and the text prompt. Such a conditional input may be unavailable or irrelevant in real photographs or surveillance scenarios. However, such a task is challenging as it involves intricate reasoning about both the semantic and geometric relationships among objects.

Our key insight is that humans perceive scene anomalies by concurrently reasoning over semantic context (*i.e.*, what the objects are and how they interact) and geometric structure (*i.e.*, where and how they are spatially arranged) [11]. Inspired by this observation, we propose *LayoutAD*, an unsupervised learning framework for detecting scene layout anomalies via semantic-geometric misalignment reasoning. In particular, given an input image, we first construct semantic and geometric graphs to represent the scene layout. We then perform message passing to jointly reason over semantic-geometric misalignment among objects. Finally, object attribute and relationship scores are fused to yield an anomaly score for each object in the scene.

To evaluate the effectiveness of our model, we construct COCOAD, a new benchmark with 1033 samples, and conduct extensive experiments across scenarios. In particular, we extend the COCO dataset [25] with various layout anomaly patterns (*e.g.*, unexpected distributions of object categories, shapes, sizes, positions, and relationships with other objects or background). Experimental results demonstrate that *LayoutAD* can detect a variety of scene anomaly types, outperforming several strong baselines both qualitatively and quantitatively. We further show downstream scene understanding and generation applications enabled by our model, including image anomaly segmentation, video anomaly detection and self-corrected image generation. Our main contributions are summarized as follows:

- We make the first attempt to tackle the problem of scene layout anomaly detection, aiming to identify diverse implausible object attributes or relation inconsistencies.
- We propose *LayoutAD*, an unsupervised learning framework that jointly reasons over semantic and geometric misalignment to estimate the anomaly degree of each object in the scene.
- The extensive experiments show that our method outperforms baselines qualitatively and quantitatively, enabling both scene understanding and generation applications.

## 2. Related Work

**Structural Anomaly Detection.** Structural anomaly detection traditionally focuses on identifying visual defects in industrial or medical scenarios, where anomalies manifest as texture inconsistencies or structural degradations. Early works, such as RIAD [47], f-AnoGAN [34], and MemAE [15], rely on reconstruction-based paradigms that reconstruct normal patterns and detect deviations via residual maps. DRAEM [46] further combines anomaly-free reconstruction with discriminative anomaly embedding, improving anomaly localization beyond purely reconstructive schemes. However, these methods often reconstruct anomalous content as normal, especially in cluttered or texture-rich conditions. Feature-based approaches, such as PaDiM [8], PatchCore [33], Reverse Distillation [9], and SimpleNet [28], leverage pretrained backbones and detect defects via embedding-space deviations. Recent diffusion-based models [13, 17, 49] have been introduced to improve anomaly localization and normal-pattern restoration. For instance, TransFusion [13] introduces transparency-aware reconstruction, while DiffusionAD [49] leverages norm-guided one-step denoising diffusion to reconstruct normal patterns and enhance anomaly localization.

However, most of these methods are primarily constrained to localizing appearance-based defects of individual objects via pixel-wise prediction. In contrast, our approach focuses on identifying layout anomalies in natural complex scenarios via object-wise prediction.

**Logical Anomaly Detection.** Logical anomaly detection [19, 50] aims to identify violations of predefined logical constraints like object arrangement. SINBAD [7] identifies fine-grained outliers in structured scenes by modeling object-set-level regularities and detecting deviations from them. WinCLIP [18] adapts vision-language models for zero- and few-shot semantic anomaly classification and segmentation by aligning image regions with textual concepts. SALAD [14] introduces a semantics-aware logical anomaly detection framework that explicitly models composition map distributions to capture spatial and semantic relationships between object components. Other works, such as LogicQA [22], leverage pretrained vision-language models to detect mismatches between images and prompts, using external knowledge to guide semantic reasoning.

The above works mainly consider logical anomaly detection with a simple background under industrial or medical settings, ignoring the underlying complex contextual relationships among objects and background. Unlike these works, our goal is to predict an anomaly map indicating the anomaly degree of each object in natural scenarios.

**Scene Anomaly Segmentation.** A variety of scene anomaly segmentation methods have been proposed, primarily categorized into uncertainty estimation [21], reconstruction-based detection [39], and outlier exposure strategies [16, 48]. SynBoost [10] combines semantic segmentation with generative reconstruction, computing a dissimilarity map between input and synthesized images

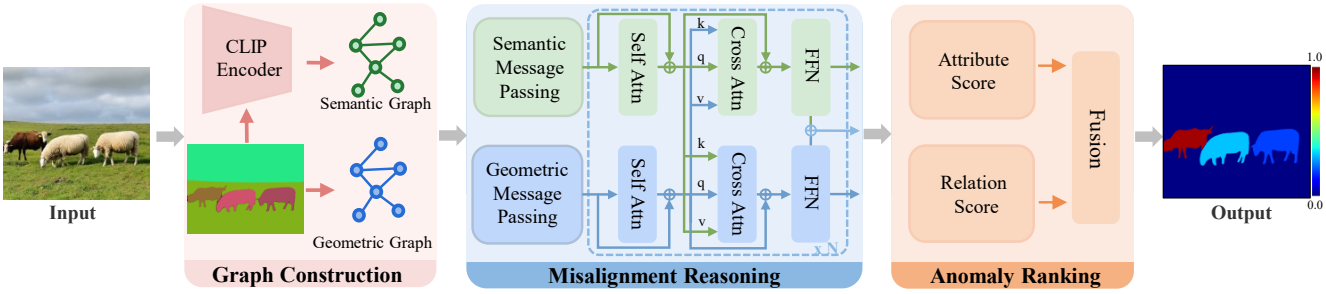


Figure 2. The overall pipeline of *LayoutAD* for scene layout anomaly detection. Given an input image, we first construct semantic and geometric graphs by the Graph Construction Module (GCM), representing the holistic scene layout structures. The Misalignment Reasoning Module (MRM) then performs intra-graph message passing and cross-graph attention to reason over semantic–geometric misalignment among objects. Finally, the Anomaly Ranking Module (ARM) estimates attribute and relation anomaly likelihoods and aggregates them into a unified anomaly score for each object to output the layout anomaly map.

to highlight potential anomalies. PEBAL [37] enhances robustness by introducing an energy-based abstention mechanism to selectively abstain from uncertain regions. SML [21] refines softmax-based scores to suppress noisy responses along object boundaries. More recent advances explore transformer-based and zero-shot architectures for open-set anomaly segmentation. Mask2Anomaly [31] introduces a mask-transformer framework that unifies semantic segmentation and anomaly localization.

While a variety of anomaly types may exist in a scene, only object-background anomalous relationships have been considered, leaving other scene anomaly types unexplored. Moreover, all these works adopt pixel-wise anomaly scoring followed by a thresholding strategy, often resulting in fragmented detection results and severe artifacts. In contrast, we introduce the new problem of scene layout anomaly detection to identify diverse layout anomalies via object-wise semantic-geometric misalignment reasoning.

# 3. Method

## 3.1. Overview

The goal of scene layout anomaly detection is to develop a model $\mathcal{D}$ that identifies object-level irregularities based on both semantic and geometric context within a scene. Given an input image, we first extract object instances using a pre-trained segmentation model [6], yielding an object set $\mathcal{O}$ to represent the scene layout. Based on $\mathcal{O}$, the output scene layout anomaly map $\hat{M}$ is obtained as follows:

$$\hat{M} = \mathcal{D}(\mathcal{O}). \tag{1}$$

As shown in Figure 2, our framework comprises three main components: a graph construction module (GCM), a misalignment reasoning module (MRM), and an anomaly ranking module (ARM). In particular, GCM builds two complementary graphs from the object set, *i.e.*, a semantic graph $G_{sem} = (V_{sem}, E_{sem})$, and a geometric graph $G_{geo} = (V_{geo}, E_{geo})$. These graphs serve as the structured representation of object appearance and spatial configuration. The constructed graphs are then fed into MRM, which jointly models semantic and geometric relationships among objects and identifies potential misalignment between these two modalities. Finally, ARM systematically evaluates both object attribute and relation anomaly scores and fuses them to obtain the anomaly degree of each object, which is then visualized as the output anomaly map.

## 3.2. Graph Construction Module

Given the object set $\mathcal{O}$ extracted from the input image, GCM transforms the scene into two complementary graphs that describe its semantic context and geometric structure. For each object $o_i \in \mathcal{O}$, we compute a semantic feature $s_i \in \mathbb{R}^{d_s}$, where $d_s$ denotes the dimension based on the concatenation of the CLIP-based [30] appearance embedding and the category-level text representation. In parallel, we derive a geometric feature $g_i \in \mathbb{R}^{d_g}$, where $d_g$ corresponds to normalized spatial descriptors such as centroid position, shape, size, and aspect ratio derived from the object mask. These features are utilized to construct the node representations of the semantic graph $G_{sem}$ and the geometric graph $G_{geo}$, respectively.

To model object relationships in the scene, we adopt a hybrid k-nearest-neighbor and distance-threshold strategy, ensuring both local context and long-range interactions can be preserved. In particular, semantic edge features are obtained by measuring appearance and category-embedding similarity between objects, while geometric edge features are derived from relative spatial cues such as displacement, distance, size ratio, and overlap.

## 3.3. Misalignment Reasoning Module

MRM aims to capture both intra-graph structure and cross-graph semantic-geometric alignment, enabling the model

to identify layout inconsistencies. Given the semantic and geometric graphs constructed by GCM, we begin with modality-specific message passing. Each graph is processed with GATv2 [5] layers, which iteratively update node and edge features by attending to their neighbors, respectively. This yields semantic and geometric representations $Z_{sem}$ and $Z_{geo}$ that capture the initial scene context. To thoroughly learn the semantic and geometric scene context, these embeddings are processed by a cross-graph transformer [38], which consists of self-attention and cross-attention layers. The semantic representation $\hat{Z}_{sem}$ is updated by a set of queries ($Q_{sem}$), keys ($K_{sem}$), and values ($V_{sem}$) in each layer:

$$\hat{Z}_{sem} = \text{Attn}(Q_{sem}, K_{sem}, V_{sem}). \tag{2}$$

The geometric representation $\hat{Z}_{geo}$ is updated similarly. These layers aggregate long-range dependencies within each modality.

To further consider semantic-to-geometric alignment, the semantic tokens or geometric tokens act as queries, while geometric tokens or semantic tokens provide keys and values in a bidirectional way as follows:

$$\begin{aligned} \hat{Z}_{sem} &= \text{Attn}(Q_{sem}, K_{geo}, V_{geo}), \\ \hat{Z}_{geo} &= \text{Attn}(Q_{geo}, K_{sem}, V_{sem}). \end{aligned} \tag{3}$$

These bidirectional interactions allow the model to detect semantic–geometric inconsistencies, such as objects whose appearance contradicts their spatial placement, or objects that are placed plausibly but mismatch the semantic context.

In addition, we introduce an edge-aware relational bias into the attention logits. For any pair of tokens $i$ and $j$, the attention score $\ell_{ij}$ is:

$$\ell_{ij} = \frac{Q_i K_j^\top}{\sqrt{d}} + b_{ij}, \tag{4}$$

where $b_{ij}$ encodes the semantic or geometric relation between the corresponding objects, ensuring that attention respects the layout structure captured in the graphs.

Finally, the aligned semantic and geometric representations $\hat{Z}_{sem}$ and $\hat{Z}_{geo}$ are obtained after several layers of intra- and inter-graph reasoning. We further derive a scene-level global feature by applying a global aggregation operator to the semantic and geometric representations:

$$z_{global} = \mathcal{G}(\hat{Z}_{sem},\, \hat{Z}_{geo}), \tag{5}$$

where $\mathcal{G}(\cdot)$ denotes a learnable aggregation and integration function that summarizes multi-object semantic and geometric cues into a holistic scene representation. The refined semantic and geometric graph representations, together with the global feature, are passed to the anomaly ranking module for scene layout anomaly detection.

### 3.4. Anomaly Ranking Module

ARM aims to evaluate how well each object and each pairwise relation conforms to the distribution of normal scene layouts. Given the refined semantic and geometric graph representations together with the global feature, ARM performs probability modeling, allowing object-level and relation-level plausibility to be assessed under the holistic scene context.

To measure semantic–geometric consistency in the scene, ARM models the likelihood of one modality conditioned on the other under the global scene context. Denoting $\hat{z}_i^{sem}$ and $\hat{z}_i^{geo}$ as the updated semantic and geometric node embeddings of object $i$ in the scene, the object attribute anomaly score $s_i^{attr}$ is computed with a mixture-density–based [4] normality estimator as follows:

$$\begin{aligned} s_i^{attr} = &-\lambda_1 \log p(\hat{z}_i^{sem} \mid \hat{z}_i^{geo},\, z_{global}) \\ &- \lambda_2 \log p(\hat{z}_i^{geo} \mid \hat{z}_i^{sem},\, z_{global}), \end{aligned} \tag{6}$$

where $\lambda_1$ and $\lambda_2$ are learnable weights.

In addition to object attribute anomalies, we also consider relational inconsistencies in the scene. Let $\hat{r}_{ij}$ denote the updated geometric relation feature between objects $i$ and $j$. We evaluate the plausibility of each pairwise relation as follows:

$$s_{ij}^{rel} = -\log p\big(\hat{r}_{ij} \mid \hat{z}_i^{sem},\, \hat{z}_j^{sem},\, z_{global}\big). \tag{7}$$

To obtain the object-wise relational anomaly score, the relation likelihoods around each object are aggregated:

$$s_i^{rel} = \log \sum_{j:(i,j)\in E} \exp(s_{ij}^{rel}). \tag{8}$$

Finally, the overall anomaly score $s_i$ for each object combines both the attribute anomaly score $s_i^{attr}$ and the relation anomaly score $s_i^{rel}$:

$$s_i = (1-\alpha) s_i^{attr} + \alpha s_i^{rel}, \tag{9}$$

where $\alpha$ balances object-centric and relation-centric cues.

### 3.5. Training

Our model is trained using two complementary likelihood-based objectives, including an attribute-level loss that models semantic–geometric consistency, and a relation-level loss that models spatial plausibility between objects.

For each object $i$, ARM predicts the conditional density of one modality given the other, both under the global scene context. The training objective is:

$$\begin{aligned} \mathcal{L}_i^{attr} = &-\lambda_1 \log p(\hat{z}_i^{sem} \mid \hat{z}_i^{geo},\, z_{global}) \\ &- \lambda_2 \log p(\hat{z}_i^{geo} \mid \hat{z}_i^{sem},\, z_{global}), \end{aligned} \tag{10}$$

where $\lambda_1$ and $\lambda_2$ are learnable weights. $p(x|h)$ is modeled as a $K$-component Gaussian mixture:

$$p(x|h) = \sum_{k=1}^{K} \pi_k(h) \mathcal{N}\big(x \mid \mu_k(h), \mathrm{diag}(\sigma_k^2(h))\big) \quad (11)$$

where the parameters are predicted from the conditional input $h$. The negative log-likelihood encourages the model to assign higher probability to semantically coherent and geometrically plausible object attributes.

In addition to object attribute anomalies, layout plausibility also hinges on object relation anomalies. For each connected pair $(i, j)$, ARM models the density of their geometric relation feature $\hat{r}_{ij}$ conditioned on the semantic context of both objects and the global feature:

$$\mathcal{L}_{ij}^{rel} = -\log p(\hat{r}_{ij} | \hat{z}_i^{sem}, \hat{z}_j^{sem}, z_{global}). \quad (12)$$

In the end, the training loss is a weighted sum of the object-level and relation-level likelihoods:

$$\mathcal{L}_{total} = \beta_{attr} \sum_i \mathcal{L}_i^{attr} + \beta_{rel} \sum_{(i,j)\in E} \mathcal{L}_{ij}^{rel}, \quad (13)$$

where $\beta_{attr}$ and $\beta_{rel}$ are controllable weights.

# 4. Experiment

## 4.1. Experiment Setup

**Implementation Details.** Our model is trained with a single NVIDIA RTX 4090 GPU. The input images are resized to $640 \times 640$. Training is conducted for 30 epochs using the AdamW optimizer with a learning rate of $1e-4$, and weight decay of $1e-4$. We adopt learning rate scheduling with a 5-epoch warm-up. $\beta_{attr}$ is 3.0, and $\beta_{rel}$ is 1.0.

**Dataset.** To evaluate our approach, we construct COCOAD, a benchmark derived from the COCO2017 dataset [25]. We select images that contain multiple objects with clear spatial arrangements, and use them to generate diverse scene layout anomalies. We then employ the Qwen-Image model [42] in text-guided mode to insert one or more anomalous objects, while preserving the global scene composition, including the camera viewpoint, background, and the original object layout.

We consider anomalies in COCOAD from two aspects. First, the object attribute anomaly indicates inconsistencies of visual or geometric properties within an object. Second, the object relation anomaly indicates violations of inter-object relationships, including physically impossible spatial configurations or semantically incompatible co-occurrences. In the end, the COCOAD benchmark contains 1033 anomaly images, serving as a unified benchmark for assessing intra- and inter-object inconsistency.

**Compared Methods.** To comprehensively evaluate the effectiveness of our proposed *LayoutAD*, we compare it with representative anomaly detection approaches across three mainstream paradigms, *i.e.*, structural anomaly detection, logical anomaly detection, and anomaly segmentation. Note that these baselines are primarily designed for pixel-level anomaly localization, while *LayoutAD* performs object-centric reasoning in the layout space. For a fair and consistent comparison, we project the object-level anomaly scores predicted by *LayoutAD* onto the pixel space using the corresponding segmentation masks, ensuring all methods are evaluated in the pixel space.

For structural anomaly detection, we evaluate PatchCore [33], SimpleNet [28], UCAD [27], GeneralAD[36], UniAD [45], and DualAnoDiff [20]. These approaches model normal appearance or texture statistics through feature reconstruction or distance modeling, effectively capturing low-level structural deviations. For logical anomaly detection, we include SINBAD [7], and WinCLIP[18], which detect high-level semantic or contextual anomalies using set-based reasoning or foundation-model embeddings. For anomaly segmentation, we choose SynBoost [10] and PixOOD [40], which aim to segment anomalous regions by combining pixel-level reconstruction and out-of-distribution detection. While effective in identifying unseen or foreign objects, these methods primarily focus on OOD anomaly localization rather than relational or layout-level inconsistencies between objects. All methods are retrained on the same subset of COCO and evaluated on the COCOAD benchmark.

**Evaluation Metrics.** Image-level anomaly detection is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC) computed from the final anomaly score of each image. For anomaly localization, the prediction result is assessed using pixel-wise AUROC (P-AUROC) and anomaly-pixel AUROC (A-P-AUROC). Following prior work [33], we report the mean AUROC on COCOAD. For fair comparison across methods, the object-level anomaly scores produced by *LayoutAD* are projected onto the pixel space and lightly smoothed to obtain stable localization maps.

## 4.2. Results

**Qualitative Evaluation.** Figure 3 provides qualitative comparisons between our approach and representative baselines from structural, logical, and segmentation-based anomaly detection paradigms. Structural-anomaly detectors like UniAD [45] primarily respond to local feature or texture deviations, resulting in scattered activations that fail to reflect the actual relational inconsistencies in the scene. Logical anomaly methods like WinCLIP [18] rely heavily on global semantic priors and tend to highlight broad contextual regions, often missing fine-grained mismatches. Segmentation-oriented approaches, such as PixOOD [40] and SynBoost [10], emphasize visually salient regions but

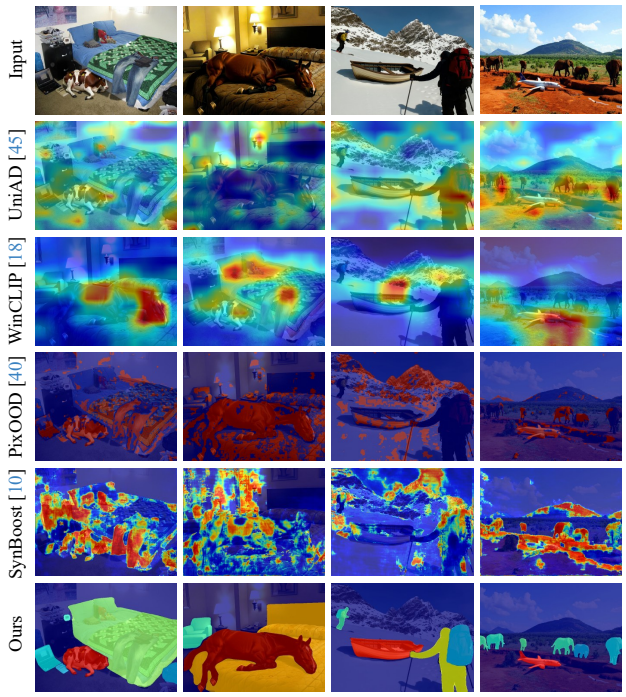


Figure 3. Qualitative comparison between the baselines and our model. Given an input image ($1st$ row), we show the anomaly detection results from different baselines ($2nd$ to $5th$ rows) and our model ($6th$ row).

frequently misinterpret normal areas as anomalous, reflecting their limited ability to reason over holistic scene context. In contrast, our *LayoutAD* precisely localizes both object attribute anomalies and object relation anomalies by explicitly reasoning over semantic–geometric alignment. Take the second column in Figure 3 as an example. *LayoutAD* distinctly activates on the horse region and suppresses background responses, correctly identifying the implausible object–context relation. Similarly, in the third column case, *LayoutAD* accurately highlights the inconsistent spatial configuration between the boat and the surrounding snow terrain, while baseline methods respond diffusely to unrelated areas. These results demonstrate that our model is able to generate semantically interpretable and spatially compact anomaly maps, effectively detecting layout inconsistencies in the scene.

**Quantitative Evaluation.** As shown in Table 1, *LayoutAD* consistently delivers the best overall performance across all three metrics, underscoring the effectiveness of explicit semantic–geometric reasoning for scene layout anomaly detection. Structural anomaly detection methods, such as PatchCore [33], SimpleNet [28], UCAD [27], GeneralAD [36] and UniAD [45], rely primarily on appearance- or feature-level deviations, making them insufficiently sensitive to layout-level irregularities arising from implausible object attributes or spatial relations. Logical anomaly detection approaches, including SINBAD [7] and WinCLIP [18], emphasize semantic consistency but do not explicitly model fine-grained geometric relations, limiting their ability to capture subtle relational inconsistencies. Segmentation-based methods, such as SynBoost [10] and PixOOD [40], perform competitively at the pixel level. However, their performance largely stems from detecting visually unfamiliar regions rather than assessing inter-object plausibility.

| Method | I-AUROC ↑ | P-AUROC ↑ | A-P-AUROC ↑ |
|---|---|---|---|
| PatchCore [33] | 0.539 | 0.571 | 0.565 |
| SimpleNet [28] | 0.551 | 0.571 | 0.515 |
| UCAD [27] | 0.547 | 0.678 | 0.682 |
| UniAD [45] | 0.479 | 0.575 | 0.508 |
| DualAnoDiff [20] | 0.573 | 0.572 | – |
| GeneralAD [36] | 0.543 | 0.565 | 0.314 |
| SynBoost [10] | 0.542 | 0.773 | 0.777 |
| PixOOD [40] | 0.538 | 0.720 | 0.722 |
| SINBAD [7] | 0.449 | – | – |
| WinCLIP [18] | 0.455 | 0.54 | – |
| Ours | **0.586** | **0.871** | **0.883** |

Table 1. Quantitative comparison of the proposed method with the baselines on COCOAD. The best results are highlighted in bold.

In contrast, *LayoutAD* explicitly models object-level semantics, geometry, and their interactions, making it possible to detect a broad range of anomalies that are overlooked by existing methods. The performance improvement demonstrates the benefit of jointly reasoning about object attributes and spatial relationships within complex scenes.

### 4.3. Ablation Study

To understand how each design choice contributes to scene layout anomaly detection, we conduct ablation studies of network components and training objectives by considering the following variants:

- *G+G & S+S*: We restrict both graph branches to use only geometric (*G+G*) or only semantic (*S+S*) features to evaluate the necessity of semantic and geometric interactions.
- *w/o GNN*: We remove the iterative message passing to evaluate the effect of localized structural dependencies.
- *w/o Transformer*: We remove the cross-graph Transformer to evaluate the semantic–geometric alignment.
- *w/o Global*: we remove the global conditional modeling to evaluate global scene context.
- *w/o Attribute / Relation Loss*: We train the model without the object-level or relation-level likelihood terms.

Table 2 shows the results of the ablation studies. Without utilizing both geometric and semantic cues, the performance drops significantly. This indicates that cross-modal information is complementary and essential for cap-

| Method | I-AUROC ↑ | P-AUROC ↑ | A-P-AUROC ↑ |
|---|---|---|---|
| G+G | 0.473 | 0.765 | 0.772 |
| S+S | 0.535 | 0.842 | 0.843 |
| w/o GNN | 0.546 | 0.848 | 0.852 |
| w/o Transformer | 0.509 | 0.794 | 0.815 |
| w/o Global | 0.522 | 0.837 | 0.848 |
| w/o Attribute Loss | 0.527 | 0.857 | 0.864 |
| w/o Relation Loss | 0.530 | 0.851 | 0.866 |
| Ours (full) | **0.586** | **0.871** | **0.883** |

Table 2. Results of the ablation study. The best results are highlighted in bold.

turing context-incompatible anomalies. If the cross-graph Transformer is removed, the performance becomes worse, implying that semantic-geometric alignment is crucial to identifying implausible configurations. Without the GNN or global aggregation, the model struggles to encode fine-grained structural dependencies and resolve ambiguities via holistic context. In addition, with the help of attribute and relation-level losses, our model learns to accurately rank both attribute-centric and relational anomalies, achieving the highest performance.

## 5. Applications

### 5.1. Image Anomaly Segmentation

Image anomaly segmentation aims to identify and localize out-of-distribution regions or unexpected obstacles that deviate from the normal distribution. While existing approaches [10, 40] are effective in detecting pixel-level unfamiliar textures, they tend to be vulnerable to background variations, such as shadows, illumination changes, or complex textures. In contrast, *LayoutAD* reasons explicitly over object-level attributes and relations in the layout space, making it robust to irrelevant background noise.

We evaluate the performance of our model and existing methods on the Road Anomaly dataset [26]. As shown in Figure 4, baseline methods tend to generate broad, noisy heatmaps that spread across multiple irrelevant background regions, while *LayoutAD* produces clean, object-aligned anomaly masks that correspond directly to underlying layout inconsistencies.

### 5.2. Video Anomaly Detection

Video anomaly detection aims to identify events or behaviors that deviate from normal spatiotemporal patterns in videos. Most existing approaches [44, 51] focus on modeling temporal dynamics to detect unusual activities. While effective for motion-based anomalies, these methods may be less effective in scenarios where temporal cues are weak or unavailable. In contrast, *LayoutAD* is able to detect spatial–semantic inconsistencies within a single frame by reasoning plausibility over object attributes and relationships in the scene, making it suitable for identifying contextual anomalies in complex scenarios.

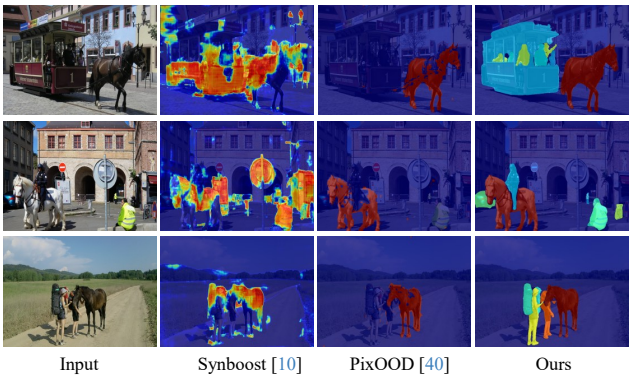


Figure 4. Qualitative results of image anomaly detection. *LayoutAD* identifies layout anomaly in real-world scenes.

We conduct evaluations on the StreetScene [32] dataset using the non-temporal Region-Based Detection Criterion (RBDC) metric. Our approach achieves the best performance with a final score of $25.4\%$, outperforming traditional motion-based methods such as Flow [32] ($11.0\%$), FG [32] ($21.0\%$), and EVAL [35] ($24.3\%$). The results demonstrate that even without temporal modeling, *LayoutAD* can identify video anomalies by leveraging efficient semantic–geometric reasoning, making it a promising and complementary direction for video anomaly detection.

### 5.3. Self-corrected Image Generation

Although modern T2I diffusion models [12, 23] demonstrate impressive visual quality, they frequently generate scenes containing layout-level anomalies that violate commonsense knowledge in the physical world. For example, animals with distorted proportions, objects floating in the air, or implausible spatial relations between entities. As illustrated in Figure 1, current generative models lack explicit mechanisms to reason about object interactions or evaluate the plausibility of the resulting layout.

To address this issue, we adopt *LayoutAD* as a layout-level structural critic to form a closed-loop self-corrected image generation pipeline. Given an image synthesized by a T2I model, we obtain its object masks via an off-the-shelf segmentation model [6]. *LayoutAD* then evaluates the semantic–geometric coherence of the scene and outputs an anomaly map, indicating which objects violate learned layout priors. Based on the predicted anomaly map, we resample a new image from the T2I model based on the identified object attribute or relationship inconsistencies. The automatic self-correction would iteratively updated until no anomaly pattern could be detected.

The results are shown in Figure 5. For the $1st$ row, the

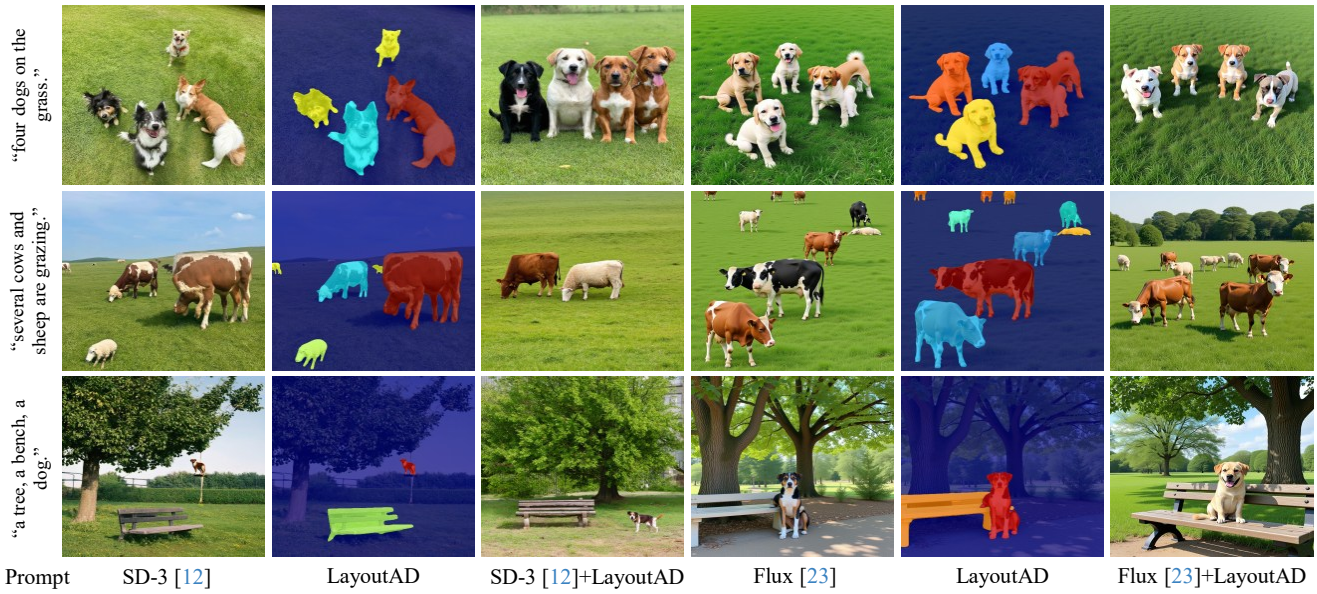


Figure 5. Self-corrected image generation. *LayoutAD* can be applied to identify and correct factually defective hallucinations that appear in generative text-to-image models (*e.g.*, SD-3 [12] and Flux [23]).

initial SD-3 generation for "four dogs on the grass" produces severely distorted objects. *LayoutAD* identifies these as high-scoring object attribute anomalies accurately. Unlike methods that rely on manual inspection or expensive fine-grained constraints, our approach requires no modification or retraining of the underlying T2I model. *LayoutAD* serves as a verifier to ensure layout plausibility during the generation process. Overall, combining *LayoutAD* with diffusion models forms a robust self-corrected generation paradigm that effectively enforces semantic–geometric consistency in the generated image.




Figure 6. Failure cases. Given an input complex scene image (left), when the segmentation results are noisy (middle), our model may fail to identify anomaly objects properly (right). The noisy object masks may mislead the graph construction process, leading to semantically plausible regions being flagged incorrectly.

## 6. Conclusion

In this paper, we take a step towards the new problem of scene layout anomaly detection. To this end, we propose *LayoutAD*, an unsupervised learning framework that constructs semantic and geometric graphs to jointly reason over semantic-geometric misalignment among objects. Under this formulation, we are able to detect diverse layout deviations, including object attribute implausibilities and relationship mismatches. Extensive experiments on the COCOAD benchmark demonstrate that our method outperforms existing baselines across multiple metrics. Furthermore, we showcase its potential in applications such as image anomaly segmentation, video anomaly detection and self-corrected image generation.

Though impressive results have been achieved by our model, as the first attempt for scene layout anomaly detection, *LayoutAD* still has some limitations. First, our model relies on accurate object masks and category labels to construct the layout graph. As illustrated in Figure 6, the noisy object masks may mislead the graph construction process, leading to semantically plausible regions being flagged incorrectly. One possible solution is to utilize the power of the multimodal large language model to improve the scene understanding abilities. Second, our current self-corrected image generation pipeline performs anomaly removal through iterative resampling. As future work, we plan to explore the precise integration of *LayoutAD* with T2I models to support controllable image correction with less computational costs.

**Acknowledgments:** This work was supported in part by National Natural Science Foundation of China (No.62302356, No.62372352), and CCF-ALIMAMA TECH Kangaroo Fund (NO.CCF-ALIMAMA OF 2025007).

## References


[1] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *WACV*, 2024. 1

[2] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 1

[3] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *IJCV*, 2022. 1

[4] Christopher M Bishop. Mixture density networks. 1994. 4

[5] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021. 4

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 7

[7] Niv Cohen, Issar Tzachor, and Yedid Hoshen. Set features for fine-grained anomaly detection. *arXiv preprint arXiv:2302.12245*, 2023. 2, 5, 6

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021. 2

[9] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022. 2

[10] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *CVPR*, 2021. 2, 5, 6, 7

[11] Robert Egly, Jon Driver, and Robert D Rafal. Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 1994. 2

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 7, 8

[13] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Transfusion–a transparency-based diffusion model for anomaly detection. In *ECCV*, 2024. 2

[14] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Salad–semantics-aware logical anomaly detection. In *ICCV*, 2025. 2

[15] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019. 2

[16] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2

[17] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *AAAI*, 2024. 2

[18] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023. 2, 5, 6

[19] Er Jin, Qihui Feng, Yongli Mou, Gerhard Lakemeyer, Stefan Decker, Oliver Simons, and Johannes Stegmaier. Logicad: Explainable anomaly detection via vlm-based text feature extraction. In *AAAI*, 2025. 2

[20] Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, et al. Dual-interrelated diffusion model for few-shot anomaly image generation. In *CVPR*, 2025. 5, 6

[21] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *ICCV*, 2021. 2, 3

[22] Yejin Kwon, Daeun Moon, Youngje Oh, and Hyunsoo Yoon. Logicqa: Logical anomaly detection with vision language model generated questions. *arXiv preprint arXiv:2503.20252*, 2025. 2

[23] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 7, 8

[24] Jingtao Li, Xinyu Wang, Hengwei Zhao, Shaoyu Wang, and Yanfei Zhong. Anomaly segmentation for high-resolution remote sensing images based on pixel descriptors. In *AAAI*, 2023. 1

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[26] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *ICCV*, 2019. 1, 7

[27] Jiaqi Liu, Kai Wu, Qiang Nie, Ying Chen, Bin-Bin Gao, Yong Liu, Jinbao Wang, Chengjie Wang, and Feng Zheng. Unsupervised continual anomaly detection with contrastively-learned prompt. In *AAAI*, 2024. 5, 6

[28] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *CVPR*, 2023. 2, 5, 6

[29] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *CVPR*, 2025. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[31] Shyam Nandan Rai, Fabio Cermelli, Barbara Caputo, and Carlo Masone. Mask2anomaly: Mask transformer for universal open-set segmentation. *IEEE TPAMI*, 2024. 3

[32] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *WACV*, 2020. 7

[33] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022. 1, 2, 5, 6

[34] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2019. 2

[35] Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *CVPR*, 2023. 7

[36] Luc PJ Sträter, Mohammadreza Salehi, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Generalad: Anomaly detection across domains by attending to distorted features. In *ECCV*, 2024. 5, 6

[37] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *ECCV*, 2022. 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[39] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In *ICCV*, 2021. 2

[40] Tomáš Vojíř, Jan Šochman, and Jiří Matas. Pixood: Pixel-level out-of-distribution detection. In *ECCV*, 2024. 5, 6, 7

[41] Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. Decad: Decoupling anomalies in latent space for multi-class unsupervised anomaly detection. In *ICCV*, 2025. 1

[42] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 5

[43] Zhenhua Xu, Yan Bai, Yujia Zhang, Zhuoling Li, Fei Xia, Kwan-Yee K Wong, Jianqiang Wang, and Hengshuang Zhao. Drivegpt4-v2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *CVPR*, 2025. 1

[44] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, 2023. 7

[45] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *NeurIPS*, 2022. 5, 6

[46] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021. 2

[47] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 2021. 2

[48] Dan Zhang, Kaspar Sakmann, William Beluch, Robin Hutmacher, and Yumeng Li. Anomaly-aware semantic segmentation via style-aligned ood augmentation. In *ICCV*, 2023. 2

[49] Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE TPAMI*, 2025. 2

[50] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *CVPR*, 2025. 2

[51] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *CVPR*, 2024. 7

[52] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *CVPR*, 2024. 1

[53] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, 2024. 1