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.

# BUSSARD: Normalizing Flows for Bijective Universal Scene-Specific Anomalous Relationship Detection

Melissa Schween Mathis Kruse Bodo Rosenhahn
Institute for Information Processing, L3S - Leibniz University Hannover
schween@tnt.uni-hannover.de

## Abstract

*We propose Bijective Universal Scene-Specific Anomalous Relationship Detection (*BUSSARD*), a normalizing flow-based model for detecting anomalous relations in scene graphs, generated from images. Our work follows a multi-modal approach, embedding object and relationship tokens from scene graphs with a language model to leverage semantic knowledge from the real world. A normalizing flow model is used to learn bijective transformations that map object-relation-object triplets from scene graphs to a simple base distribution (typically Gaussian), allowing anomaly detection through likelihood estimation. We evaluate our approach on the SARD dataset containing office and dining room scenes. Our method achieves around 10% better AUROC results compared to the current state-of-the-art model, while simultaneously being five times faster. Through ablation studies, we demonstrate superior robustness and universality, particularly regarding the use of synonyms, with our model maintaining stable performance while the baseline shows 17.5% deviation. This work demonstrates the strong potential of learning-based methods for relationship anomaly detection in scene graphs. Our code is available at* https://github.com/mschween/BUSSARD.

## 1. Introduction

Detecting anomalies in images is not always straightforward, as the definition of what constitutes an anomaly depends on the setting and context of the scene. For industrial applications, anomalies typically manifest as deformed or missing objects [3, 4]. The automatic detection of anomalies in surveillance footage and similar domains is gaining research attention, addressing complex data with substantial variability [32, 57, 73]. In this area, the focus shifts from identifying individual defects to understanding entire scenes and evaluating context, with the challenge of filtering what is relevant and necessary to consider. Most existing approaches focus on individual scene components, particularly human poses, with limited integration of broader contextual information [21, 42, 43]. However, this human-centric focus neglects important anomaly types, such as objects in inappropriate locations (*e.g.* car on playground) or unusual human-object relationships (*e.g.* child on car). Recognizing this gap, Lai *et al.* [32] introduced the novel task of *Scene-Specific Anomalous Relationship Detection (SARD)* along with a dedicated dataset.

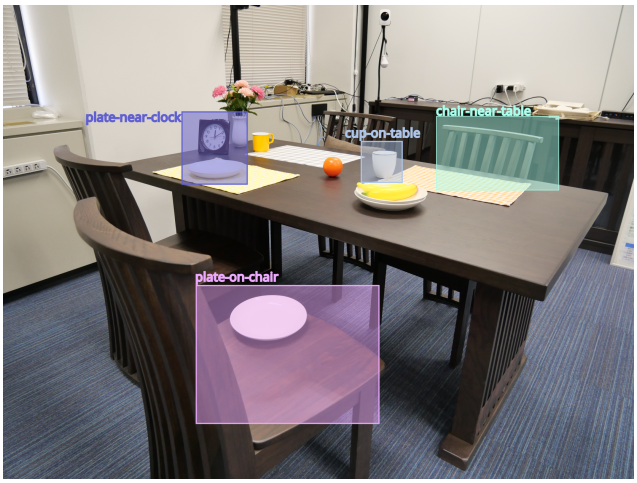


Figure 1. Example image from SARD dataset [32]. A non complete scene graph consists of: 'plate-on-chair', 'plate-near-clock', 'cup-on-table', 'chair-near-table'. The anomaly to detect is 'plate-on-chair'.

Recent progress in scene graph generation (SGG) [24, 46, 65, 69], accelerates the research of applying SGGs for anomaly detection (AD) in complex scenes through structured extraction of visual information. While this combination remains underexplored [9, 16, 41], it offers significant potential by interpreting structured scene representations for more effective anomaly detection. Fig. 1 highlights some object relations that can be extracted from the image using SGG. More examples with complete scene graphs can be found in the appendix (see Sec. 12).

Lai *et al.* [32] propose a counting-based approach to de-

tect anomalous relationships for the SARD task. A key limitation of this method is its vulnerability to the long-tail problem, where data exhibits severe imbalance with a few instances appearing frequently while the majority is rare. This counting-based approach cannot fully compensate for such distributional bias, leading to degraded detection performance. We present, to the best of our knowledge, the first learning-based solution for SARD. Our approach makes three core contributions to efficiently approach the task of SARD:

First, we follow a multimodal approach by converting scene graphs to semantic vector representations using a pretrained language model to capture real-world knowledge about object and relationship semantics. This embedding approach addresses the long-tail distribution of objects and relationships in scene graphs, where many entities appear rarely in training data. By placing semantically similar words in close proximity within the embedding space, the model can generalize to previously unseen objects.

Second, we leverage normalizing flows for anomaly detection. Normalizing flows learn bijective transformations that map data to a standard normal distribution, making them well-suited for anomaly detection, as data points that deviate significantly from the normal distribution after transformation can be assigned high anomaly scores based on their log-likelihood.

Finally, we address the dimensionality challenge inherent to normalizing flows. Since normalizing flows require bijective mappings, they cannot perform dimensionality reduction. However, language embeddings are typically high-dimensional, which can lead to training instability and computational inefficiency. To resolve this, we employ an autoencoder that compresses the concatenated triplet embeddings into a lower-dimensional latent space before feeding them to the normalizing flow.

We call our approach **B**ijective **U**niversal **S**cene-**S**pecific **A**nomalous **R**elationship **D**etection (*BUSSARD*). The 'bijective' refers to the invertible transformations in normalizing flows, while 'universal' reflects the model's robustness to vocabulary variations, a key advantage demonstrated in our experiments. *BUSSARD* achieves state-of-the-art performance with an AUROC 10% higher than the counting-based SARD baseline.

Our work comprises the following contributions:

- We present the first learning-based approach for SARD, employing normalizing flows to learn distributions over scene graph relationships.
- We demonstrate superior robustness and universality through semantic embeddings, effectively mitigating the long-tail problem that affects counting-based approaches.
- We achieve state-of-the-art performance on the SARD dataset, outperforming the baseline by 10% in AUROC.

## 2. Related Work

We review related works from the fields of scene graph generation and anomaly detection.

### 2.1. Scene Graph Generation (SGG)

In scene graph generation (SGG) structured information is extracted from images, which is stored in a graph consisting of multiple triplets. These triplets are composed of two objects, and a relationship describing the connection between the objects. A simple example can be seen in Fig. 1, with triplets such as 'plate-on-chair' or 'plate-near-clock'.

The major approaches for generating scene graphs may be separated into *one-stage* and *two-stage methods*. In two-stage generation, objects are detected first, and then the relations between them are predicted in a second independent step [34, 40, 59]. The one-stage approach differs in that both objects and relations are predicted simultaneously. Models such as RelTR, OED, PGSG, and EGTR fall under this group [11, 24, 36, 61]. EGTR operates by extracting relation graphs from the multi-head self-attention layers of the DETR [6] decoder, treating attention queries and keys as subject and object entities [24]. Overall, numerous diverse approaches are currently emerging to improve SGG, showing the interest in this field [26, 46, 65, 69]. Besides the image-based SGG there is also a subfield that concentrates on generating scene graphs for videos, called Dynamic Scene Graph Generation (DSGG) [8, 66, 70, 71]. Here scene graphs are connected over time to track how they change. This enables the recognition of dynamic relations like 'person-throwing-ball', instead of 'person-holding-ball'.

**Word Embeddings.** Scene graphs consist of textual triplets, making pretrained word embeddings a natural choice for encoding semantic information. Word embedding models such as word2vec [39] and GloVe [7, 47] map words to continuous vector spaces where semantic similarity is reflected by proximity. This property is particularly valuable for addressing the long-tail distribution in scene graphs: semantically similar but lexically different terms (*e.g.* 'person' vs. 'human', 'sitting on' vs. 'seated on') are embedded nearby, enabling the model to generalize across rare or unseen vocabulary.

### 2.2. Anomaly Detection (AD)

The goal of anomaly detection (AD) is to identify unusual, anomalous samples of data that fall outside of what may be considered normal data. However, a prominent challenge is the severe imbalance in available training data, which contains far more normal instances than anomalous ones. For this reason, most algorithms are semi-supervised, where the models are trained exclusively on normal data and only encounter anomalous data during testing and inference. Algorithms that are widely used for AD include

memory banks, student-teacher networks and normalizing flows, while Foundation Models are also gaining more popularity.

*Memory banks* are used to store a learned representation of normal data and use it as a reference for comparison during inference [15, 52, 64].

*Student-teacher networks* typically use a pretrained network (such as a ResNet [19] trained on ImageNet [12]) as the teacher, which is typically trained on some auxiliary task. The student network is optimized to imitate the teacher using only normal data. During inference, the student may only be able to mimic the teacher on normal data, leading to high errors for anomalous (unseen) patterns. Thereby, the difference between student and teacher serves as a proxy for an anomaly score. A larger difference indicates higher anomaly likelihood, since the student produces more varied results for unseen data. There are various approaches implementing student-teacher networks for AD [5, 23, 54].

*Foundation Models* are trained on massive datasets and can be applied to a wide range of downstream tasks. One such task is AD, as their learned understanding of the world helps to identify what might be unusual. A key advantage is their applicability in *few-shot* or *zero-shot* settings, requiring minimal or no task-specific training. Therefore, several approaches have been proposed to address AD tasks, but they typically require careful pipeline design tailored to the specific problem or they rely on fine-tuned prompts to overcome challenges [25, 33, 72, 75]. Specifically there are multiple approaches using GPT [44] for zero-shot AD [17, 27, 74].

*Normalizing Flows* learn a bijective mapping between a simple base distribution and a complex target distribution. The bijective nature enables their use for applications such as data generation [22, 45, 51], by sampling from the base distribution and applying the reverse mapping to the complex distribution. Additionally, normalizing flows provide exact and tractable log-likelihood evaluation, which is essential for their application to AD, as demonstrated by several existing methods [21, 31, 53].

Anomaly detection finds application across diverse domains, with unique challenges and specialized approaches.

*Industrial applications* represent one of the most mature areas, supported by well-established benchmarks such as MVTec AD and VisA [3, 4, 20, 76]. These datasets have motivated numerous methods advancing detection capabilities [18, 31, 52, 54]. The medical field has similarly developed domain-specific datasets and approaches [2, 60, 67].

*Video-based AD* introduces temporal complexity, requiring methods to track elements over time. This is particularly relevant for surveillance applications, where datasets like ShanghaiTech and UCF-Crime [57, 73] have driven the development of pose-based approaches, sometimes augmented with contextual information [21, 35, 42, 43].

*Graph-based AD* addresses structured relational data, in domains such as attributed networks, social networks and industrial control systems [1, 37, 50, 62]. Recent works have begun incorporating spatial context through context graphs [58] and spatio-temporal graphs [48].

**Anomaly Detection using Scene Graphs.** Recent advances in SGG have sparked increased interest in applying scene graphs to anomaly detection, a relatively underexplored area. Garg *et al*. [16] focus on generating new scene graphs, primarily for image generation, and demonstrate anomaly detection on their own generated outputs. However, their method requires node classification for likelihood estimation, necessitating predefined entity classes for all objects, and evaluates generation quality rather than detecting anomalies in observed scene graphs from real images, making it unsuitable as a baseline for our task. The work of Chen *et al*. showed the potential of using scene graphs for video-based anomaly detection [9]. Complex-VAD a recently published method, also uses scene graphs for anomaly detection in videos and highlights the importance of focusing on interaction anomalies [41]. The authors of SARD highlight the importance of focusing on anomalous relationships and use a counting based approach for detecting those using scene graphs in images [32]. SARD is explained in more detail in the following section.

## 3. Foundations

This section provides a brief overview of the *Scene-Specific Anomalous Relationship Detection (SARD)* task [32], its baseline model, and the evaluation metrics employed. SARD shifts the focus from detecting anomalous objects to identifying context-dependent anomalies by analyzing relationships between objects. This relationship-based approach captures anomalies that arise from unusual interactions within a scene.

### 3.1. Baseline Method

We use the AD-model of Lai *et al*. as baseline [32]. The authors use EGTR [24] for creating scene graphs for each image to extract the relevant information. Subsequently the top 30 triplets with the highest confidence for every image are used for further analysis. Lai *et al*. do not train a neural network to detect anomalous triplets but instead employ an enhanced counting-based approach, where the triplets that appear least often are considered to be an anomaly. For this, scene graphs from multiple images are grouped in subgroups with a specified size. Each subgroup, that can have 11, 21 or 31 elements, consists of one anomalous image with the remainder filled with normal images. The normal images appear in multiple subgroups.

Scene graphs of each subgroup are grouped separately and the triplets that appear the least, are considered to be

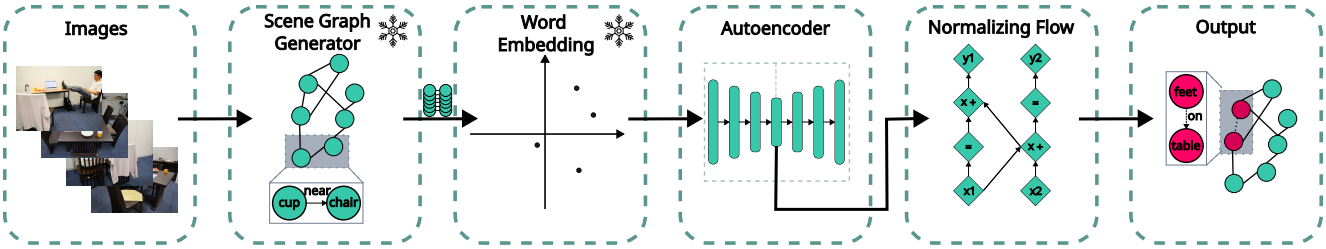


Figure 2. The components of *BUSSARD*. The images are parsed using a pretrained scene graph generator. Each triplet is then encoded using a pretrained word embedding model. The embeddings of the triplets are each concatenated and the dimension is reduced using an autoencoder. In the end, a normalizing flow is used to predict the likelihood of the triplets being anomalous.

anomalies. For enhancing the counting-based approach, they include the concepts *rarity adjustment*, *noisy normal relationship reduction* and *soft counting*. Rarity adjustment is used to balance the confidence score bias of the SGG model, as those models tend to assign anomalous relationships lower confidence scores which can lead to an overshadowing through the normal relationships. Noisy normal relationship reduction is added to reduce the likelihood of classifying infrequent, but normal relationships as anomalies. The last enhancement employs soft counting to improve the robustness of object relationship representations, as it enables the consideration of multiple likely relationships for each object pair. These enhancements increase the performance of their model, however improving the biased confidence score, or reducing the influence of rare normal relationships does not solve the problem of unbalanced data that is typically found in AD. Similarly, efforts to increase robustness through considering multiple relationships do not fully enable accurate performance in an open-world setting.

### 3.2. Metrics

A common metric to evaluate the performance of anomaly detection frameworks is Area Under the Receiver Operating Characteristic Curve (*AUROC*). AUROC is commonly used as it is scale- and threshold-invariant, which helps to measure the quality of prediction rankings, irrespective of the classification threshold chosen. Conceptually, AUROC corresponds to the area under the curve of the true positive rate (TPR) as a function of the false positive rate (FPR) with

$$\text{AUROC} = \int_0^1 \text{TPR}(u)\, d(\text{FPR}(u)), \tag{1}$$

where both rates are evaluated at varying thresholds $u$.

The authors of SARD introduce an additional metric, *AUC-Recall@k*, to assess ranking quality in the top positions [32]. This metric combines the intuition of *Recall@k*, measuring how many anomalies appear among the top-$k$ ranked samples, with the concept of integrating over the recall curve. Following the protocol of SARD, 30 triplets with the highest confidence from the SGG are used. Let $\mathcal{T}$ denote the set of all used triplets extracted from the dataset (or from a particular image subgroup, similar to SARD). For each triplet $T \in \mathcal{T}$, let $a \in \mathbb{R}$ denote its anomaly score, and define the ground truth anomaly indicator

$$y = \begin{cases} 1 & \text{if } T \text{ is anomalous} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Let $N$ be the total number of anomalous triplets (assume $N > 0$), and let $r(T)$ denote the rank of triplet $T$ when all triplets are sorted in descending order by their anomaly scores. Then, Recall@$k$ is defined as

$$\text{Recall@}k = \frac{1}{N} \sum_{T \in \mathcal{T}} \mathbf{1}\{r(T) \leq k\} \cdot y, \tag{3}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. To aggregate Recall@$k$ across a range of $k$ values (following the SARD protocol), we define AUC-Recall@$k$ as the normalized discrete area under the Recall@$k$ curve for integer $k \in k_{\min}, \ldots, k_{\max}$

$$\text{AUC-Recall@}k = \frac{1}{k_{\max} - k_{\min} + 1} \sum_{k=k_{\min}}^{k_{\max}} \text{Recall@}k. \tag{4}$$

In our experiments, we set $k_{\min} = 1$ and $k_{\max} = 100$, following the setting of SARD. A higher AUC-Recall@$k$ indicates that anomalous triplets are ranked closer to the top of the list. Since Lai *et al*. evaluate ranking quality within contextual subgroups, AUC-Recall@$k$ is also computed separately for each image subgroup.

AUROC thus captures global discrimination ability across all thresholds, while AUC-Recall@$k$ focuses on the ranking performance of the most anomalous triplets within the top-$k$ positions.

## 4. Method

An overview of *BUSSARD*'s architecture is presented in Fig. 2. The model comprises four sequential steps. First,

a pretrained SGG model extracts scene graphs from images, providing triplets for analysis. Second, a pretrained word embedding model converts the object and relationship tokens to semantic vector representations, producing rich embeddings. Afterwards, an autoencoder compresses these features to lower-dimensional latent vectors. This compression is essential for the normalizing flow, as these models require matching input and output dimensions due to their bijective nature and exhibit training instability with high-dimensional inputs. The compressed latent vectors are then passed to the normalizing flow, which assigns anomaly scores to triplets. The SGG model and word embedding models are kept frozen to preserve their learned general semantic knowledge from pretraining.

### 4.1. Scene Graph Generator

The SGG model takes images $I$ as input and generates scene graphs $G = (V, E)$, where $V$ represents nodes and $E$ represents edges. We consider subsets of the scene graphs in the form of triplets $T_{i,j} = (o_i, p_{i,j}, o_j)$ which consist of two nodes (objects) and one directed edge (relationship), where $o_i, o_j \in V$ and $p_{i,j} \in E$. Thus the triplet $T_{i,j}$ describes the relationship $p_{i,j}$ from object $o_i$ to object $o_j$.

### 4.2. Word Embedding

We employ the pretrained word embedding model GloVe [7] to map object and relationship tokens to vector representations. This leverages learned semantic knowledge from large-scale pretraining, encoding meaningful relationships between words in the embedding space. Each object is mapped to a vector $\mathbf{v}_{o_i} \in \mathbb{R}^d$ and each relationship $p_{i,j}$ is mapped to a vector $\mathbf{v}_{p_{i,j}} \in \mathbb{R}^d$. Afterwards, the vectors are concatenated to form a unified triplet representation with

$$\mathbf{t}_{i,j} = \text{concat}(\mathbf{v}_{p_{i,j}}, \mathbf{v}_{o_i}, \mathbf{v}_{o_j}) \in \mathbb{R}^{3d}. \tag{5}$$

In the following, we drop the indices $i$ and $j$ for simplicity, referring simply to triplets $\mathbf{t}$ when the context is clear.

### 4.3. Autoencoder

An autoencoder (AE) compresses concatenated triplet vectors into a lower-dimensional latent space. It consists of an encoder $f_{\text{enc}} : \mathbb{R}^{3d} \to \mathbb{R}^{d_z}$ and a decoder $f_{\text{dec}} : \mathbb{R}^{d_z} \to \mathbb{R}^{3d}$, mapping each triplet $\mathbf{t}$ to a latent vector $\mathbf{z} = f_{\text{enc}}(\mathbf{t})$ and reconstructing it as $\hat{\mathbf{t}} = f_{\text{dec}}(\mathbf{z})$, where $d_z < 3d$. The AE is trained on normal data by minimizing the mean squared error (MSE) reconstruction loss

$$\mathcal{L}_{\text{AE}} = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \|\mathbf{t} - \hat{\mathbf{t}}\|^2. \tag{6}$$

After training the AE in a first stage, only the encoder with frozen weights is used to generate latent representations for the normalizing flow to enable an efficient computation with dense information.

### 4.4. Normalizing Flow

The latent vectors are passed to a normalizing flow to compute anomaly scores for all triplets. Let $\mathbf{u} \in \mathbb{R}^{d_z}$ denote samples from a simple base distribution, which is commonly a standard Gaussian $\mathbf{u} \sim \mathcal{N}(0, I)$ with density $p_{\mathbf{u}}(\mathbf{u})$ and let $\mathbf{z} \in \mathbb{R}^{d_z}$ denote samples from the target distribution. The normalizing flow $f_{\text{flow}} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ models an invertible transformation that maps the target distribution to the base distribution with $\mathbf{u} = f_{\text{flow}}(\mathbf{z})$.

The density $p_{\mathbf{z}}(\mathbf{z})$ can be estimated using the change of variables formula as

$$p_{\mathbf{z}}(\mathbf{z}) = p_{\mathbf{u}}\left(f_{\text{flow}}\left(\mathbf{z}\right)\right) \left|\det\left(\frac{\partial f_{\text{flow}}}{\partial \mathbf{z}}\right)\right|. \tag{7}$$

Normalizing flows require invertible transformations with efficiently computable Jacobian determinants. Various architectures have been proposed to satisfy these requirements [13, 22, 29], with RealNVP [14] being one of the most widely used approaches. RealNVP employs coupling layers with affine transformations, which are trivially invertible and have tractable Jacobian determinants due to their triangular structure, enabling scalable density modeling in high-dimensional spaces. Each coupling layer splits the input $\mathbf{z}_{\text{in}} \in \mathbb{R}^{dz}$ into two parts $\mathbf{z}_{\text{in},1} \in \mathbb{R}^{d_z/2}$ and $\mathbf{z}_{\text{in},2} \in \mathbb{R}^{d_z/2}$. The first part remains unchanged while the second part undergoes an affine transformation conditioned on the first part, leading to the equations

$$\mathbf{z}_{\text{out},1} = \mathbf{z}_{\text{in},1} \tag{8}$$

$$\mathbf{z}_{\text{out},2} = \mathbf{z}_{\text{in},2} \odot \exp(s(\mathbf{z}_{\text{in},1})) + t(\mathbf{z}_{\text{in},1}), \tag{9}$$

where $s$ and $t$ are deep neural networks computing the scale and translation parameters respectively. Multiple coupling layers are stacked with alternating permutations between layers to ensure all dimensions are transformed, yielding a flexible and tractable bijective function.

The anomaly score $a$ for each triplet $T$ is computed as the negative log-probability

$$a = -\log p(\mathbf{z}) = -\log p(\mathbf{u}) - \log \left|\det \frac{\partial f_{\text{flow}}}{\partial \mathbf{z}}\right|, \tag{10}$$

where higher scores indicate higher anomaly levels. For numerical stability, triplets with undefined scores (NaN or Inf) are assigned the maximum observed anomaly score. The normalizing flow is trained exclusively on normal data with maximizing the negative log-likelihood loss given by

$$\mathcal{L}_{\text{flow}} = -\frac{1}{2}\|\mathbf{u}\|_2^2 + \log \left|\det \frac{\partial f_{\text{flow}}}{\partial \mathbf{z}}\right|. \tag{11}$$

## 5. Experiments

Our approach is evaluated on the SARD dataset and compared to what is, to the best of our knowledge, the only

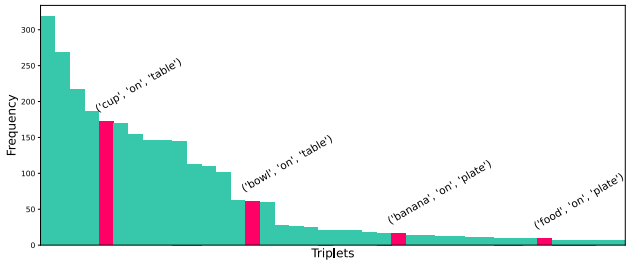


Figure 3. The 40 most frequent triplets of the **dining room** scene. The labels belong to the highlighted bars, showing example triplets.

available baseline. We also tested *BUSSARD* in a cross-dataset setting, where the model is trained on a subset of the MIT indoor dataset (MIT-67) [49] (for more details see appendix Sec. 8). Several ablation studies are conducted to show the robustness of *BUSSARD* and underline our design choices.

### 5.1. SARD Dataset

The SARD dataset contains images of two indoor scenes for anomaly detection, specifically created for the SARD task [32]. The first setting is a dining room scene with 60 normal and 54 anomalous images and the second scene consists of 60 normal and 63 anomalous images. All images are labeled according to whether they contain an anomaly or represent normal data. Additionally, the ground truths of anomalous images contain a description of the anomalous triplet. In each image only one anomaly can be found, but there can be varying valid descriptions for the same anomaly like 'cup-on-plate' and 'plate-under-cup'.

**Scene Graph Dataset Analysis and Adaption.** The images in the dataset are converted to scene graphs using EGTR [24]. Following the protocol suggested by SARD, the top 30 triplets are used for further analysis, while removing triplets with minor objects as defined by SARD [32] using ConceptNet [56]. The top 30 triplets of all scene graphs contain approximately 145 unique triplets for the office scene, while the dining room scene contains about 80 unique triplets. An overview of the distribution of the 40 most common triplets in the dining room scene can be found in Fig. 3, while the distribution of the office scene can be found in the appendix (Fig. 6). The distribution exhibits the long-tail problem characteristic, as a few triplets are highly frequent while the majority appear very infrequently. This is why the counting-based approach of Lai *et al*. has weaknesses, as many normal triplets appear infrequently.

An analysis of all loaded triplets reveals that the anomalous triplet is not always within the top 30 triplets. This occurs as EGTR assigns lower confidence scores to actual anomalies, causing them to be ranked below the top 30 threshold and are thus excluded from the dataset. This is problematic since ground truth anomalies that are not present in the extracted triplets cannot be detected by any method. To ensure fair evaluation, these missing ground truth anomalous triplets are added in a preprocessing step. For comparability, the baseline SARD method is re-run with this corrected data (SARD-c) in addition to reporting their original results (SARD-o).

The authors of SARD used and evaluated the performance within subgroups of sizes 11, 21 and 31. Since *BUSSARD* needs to be trained (unlike the SARD model, which operates by counting directly), we use a training/test split of 80/20 for the normal data. This reduces the amount of normal data available for testing. Consequently, we adopt the subgroup size of 11, pairing each anomalous image with ten normal images from the test set.

### 5.2. Implementation Details

For converting the images to scene graphs, we use the model EGTR [24], which is pretrained on Visual Genome [30]. The word embedding model GloVe [7] is used to translate each word into a vector with dimension $d = 300$. The three vectors building one triplet are concatenated to one vector $\mathbf{t}$ with dimension $3d = 900$. The SGG and word embedding models are not trained or finetuned, but instead use frozen pretrained weights.

For dimensionality reduction, an autoencoder is pretrained separately for each scene using only normal training data from the SARD dataset. The $3d$-dimensional input is passed through four fully connected layers with ReLU activations, with the latent dimension being $d_z = 512$. Ablation studies (see Sec. 5.4.2) demonstrate that this latent dimension yields optimal performance. The decoder mirrors the architecture in reverse. The autoencoder is trained using the Adam optimizer [28] with learning rate $\eta = 0.001$ for 100 epochs, minimizing $\mathcal{L}_{\text{AE}}$. Only the encoder is used with frozen weights for subsequent anomaly detection, after the training of the AE.

We implement the normalizing flow using three RealNVP coupling layers [14]. The first two layers use alternating mask patterns, while the third uses a half mask pattern. The $s$- and $t$-networks of the coupling layers consist of multi-layer perceptrons (MLPs) that each have three fully connected layers with ReLU activations and a hidden dimension of 128. The normalizing flow is trained exclusively on normal data for 1000 epochs using the AdamW optimizer [38] and starting with an initial learning rate of $\eta = 10^{-4}$, while maximizing $\mathcal{L}_{\text{flow}}$. A plateau scheduler is employed to adaptively adjust the learning rate, by reducing the learning rate by a factor of 0.8 when the loss plateaus for 30 consecutive epochs, with a minimum learning rate of $10^{-7}$. *BUSSARD* is trained with ten different random seeds, while SARD is executed once as the counting-based method does not involve stochastic training.

Table 1. AUROC and AUC-Recall@k on the SARD dataset. SARD-o refers to the original baseline results with incomplete data (some ground truth anomalies missing from top 30 triplets). SARD-c refers to the baseline re-run with corrected data where all ground truth anomalous triplets are present. BUSSARD is evaluated on the corrected data. Both baselines are included for transparency. *BUSSARD* was executed with ten different seeds and the average and standard deviation are provided. The best performing model is **highlighted**.

| Scene | Model | AUROC (↑) | AUC-Recall@k (↑) |
|---|---|---|---|
| Dining Room | SARD-o | 90.30 | 69.48 |
| | SARD-c | 90.50 | 91.50 |
| | *BUSSARD* **(ours)** | **97.85** ± 0.6 | **92.85** ± 1.52 |
| Office | SARD-o | 81.80 | 64.52 |
| | SARD-c | 82.50 | 82.70 |
| | *BUSSARD* **(ours)** | **95.29** ± 1.65 | **89.57** ± 3.46 |

## 5.3. Evaluation

For evaluation, we apply *BUSSARD* to the corrected SARD dataset and compare it primarily against SARD-c, ensuring both methods are evaluated on the same corrected scene graphs (see Sec. 5.1). The original SARD results (SARD-o) are also provided for completeness. The AUROC and AUC-Recall@k values for each scene are presented in Tab. 1.

*BUSSARD* outperforms SARD-o and SARD-c by a large margin in AUROC, achieving 97.85% and 95.29% on the dining room and office scenes, respectively, with an average improvement of 10.07% across both scenes. *BUSSARD* also improves ranking quality, with an average 4.11% improvement in the AUC-Recall@k metric compared to SARD-c. Notably, when comparing to the original implementation SARD-o, our method has improved in average 24.21% in the AUC-Recall@k metric. A qualitative analysis of a few selected examples is provided in the appendix (see Sec. 11).

**Computational Efficiency.** Beyond detection performance, *BUSSARD* demonstrates superior computational efficiency. Experiments conducted on an NVIDIA RTX 3090 show that, despite SARD-c's simpler counting-based implementation, our approach is on average five times faster.

## 5.4. Ablation Studies

We conduct several ablation studies with special focus on showing robustness, as well as justifying design decisions.

### 5.4.1. Robustness

**Synonyms.** To demonstrate universality in an open-vocabulary setting, we conduct a study where four chosen words are replaced with similar words (see Sec. 9.1 in appendix for more details). The replaced words were chosen to be common, while their synonyms are not yet present, making the effect of synonym substitution clearly measurable. For understanding the influence of the synonyms, words are replaced with different probabilities, here called *synonym rate*. All ground truths, with triplets containing at least one of the chosen words were modified by adding the triplets with the synonyms as well.

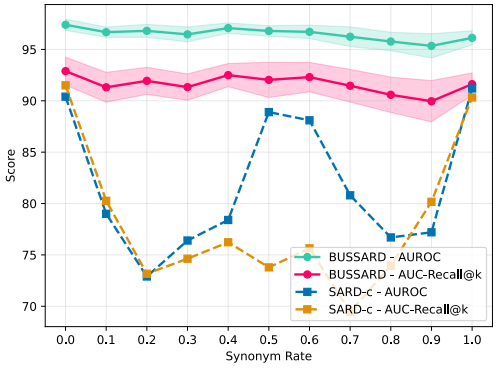


Figure 4. Ablation results with AUROC (↑) and AUC-Recall@k (↑) of *BUSSARD* and SARD-c for different synonym rates for the **dining room** scene. The synonym rate represents the probability of substituting words using synonym mappings. For *BUSSARD* the dots represent the average results after running with ten different seeds, and the shaded area visualizes the corresponding standard deviation. SARD-c was run only once for each rate as the calculation is deterministic.

The results of *BUSSARD* and SARD-c are shown in Fig. 4 for the different synonym rates of the dining room scene. Similar results of the office scene can be found in the appendix Fig. 7. Most notably, our model remains stable across all synonym rates with around 2% deviation. In contrast, the results of SARD-c show significantly stronger deviations with a maximum deviation of around 17.5%.

We additionally conduct experiments adding *Gaussian noise* to the word embeddings to simulate slight word variances, confirming robustness to small variations. Further details are provided in the appendix (see Sec. 9.2).

Overall, the synonym experiment and the noise robustness highlight the more universal nature of *BUSSARD*, which maintains stable performance across vocabulary variations, in contrast to the less robust enhanced counting approach of SARD-c.

**Multi-Scene and Cross-Dataset Evaluation.** To show the robustness of our approach, we provide the results of two additional experiments, summarized in Tab. 2. In both experiments, training is performed on the combined dining room and office scenes, and evaluation is conducted per scene on the SARD test set. In the first experiment, we use MIT-67 as the training source, and use the combined SARD scene training sets for the second experiment. Tab. 2 shows that training on both SARD scenes improves office performance with only a small dining room decline. Using MIT-67 reveals strong cross-dataset performance on the dining room scene (92% AUROC), whereas the office scene

Table 2. AUROC and AUC-Recall@kk k comparing multi-scene training on MIT-67 and SARD, evaluated per scene on the SARD test set.

| Train Set | Test Scene | AUROC (↑) | AUC-Recall@k (↑) |
|---|---|---|---|
| SARD | Dining Room | 97.46 ± 0.59 | 91.60 ± 2.11 |
| | Office | 96.73 ± 2.06 | 88.53 ± 7.31 |
| MIT-67 | Dining Room | 92.45 ± 2.50 | 74.79 ± 7.93 |
| | Office | 60.03 ± 4.45 | 20.70± 5.63 |

Table 3. Ablation study results for the **dining room** scene. All experiments were executed with ten different seeds and the average results are listed with their standard deviation. $d_z$ is the input dimension to the normalizing flow.

| Experiment | $d_z$ | AUROC (↑) | AUC-Recall@k (↑) |
|---|---|---|---|
| Feature Sum | 128 | 91.22 ± 2.49 | 78.88 ± 4.59 |
| Feature Mult | 128 | 96.48 ± 0.83 | 90.02 ± 2.11 |
| Node Only | 512 | 95.20 ± 1.66 | 89.33 ± 3.32 |
| No AE | 900 | 85.44 ± 2.33 | 63.40 ± 4.78 |
| *BUSSARD* (ours) | 512 | **97.85** ± 0.6 | **92.85** ± 1.52 |

suffers from a serious decline (60% AUROC). We attribute this to the distribution shift between the MIT-67 dataset and SARD's office scenes.

**Encoding Independence.** Using RelTR [11] as the scene graph generator yields similar performance (see appendix for more details Sec. 9.3), confirming independence from a specific generator. Similarly, substituting GloVe [7] with stronger sentence embedding models [55, 63, 68] confirms embedding model independence. More details are provided in the appendix (see Sec. 9.3).

### 5.4.2. Design Study

We perform a design study to assess the influence of individual components and justify our design choices.

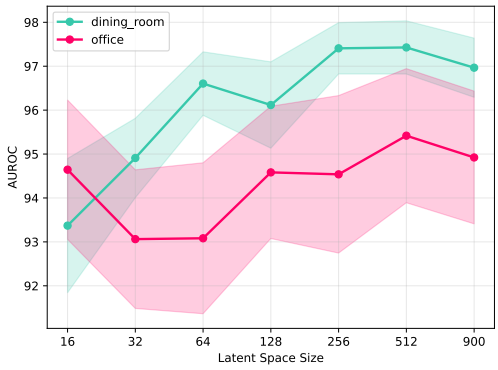


Figure 5. Ablation results with the AUROC (↑) for different latent space dimensions of the autoencoder.

**Autoencoder Latent Dimensions** Since normalizing flows are sensitive to input dimensions [10], we conduct experiments with varying latent dimension sizes to validate our hyperparameter choice. The AUROC results across different latent dimensions are presented in Fig. 5. It shows that smaller latent dimensions yield inferior performance compared to larger ones, justifying our choice of $d_z = 512$. Notably, performance drops again when the latent dimension approaches the input dimensionality (here $d_z = 900$). The AUC-Recall@k results, which exhibit similar behavior, can be found in the appendix (see Fig. 8).

**Feature Aggregation Methods.** In *BUSSARD*, the feature vectors of each word in a triplet are concatenated. To evaluate whether a more compact representation could be suitable, we performed two experiments that reduce the representation size. The first variant sums all feature vectors instead of concatenating them, resulting in a three times smaller autoencoder input dimension of $d = 300$ (called *Feature Sum*). Since the chosen latent dimension exceeds this input dimension, the autoencoder is adjusted to compress vectors to $d_z = 128$. The second variant applies element-wise multiplication of the vectors instead of summation, called *Feature Mult*. The results of the dining room can be seen in Tab. 3 and the office results are in the appendix (see Tab. 6). Both experiments yield inferior results, with Feature multiplication performing closest to *BUSSARD*.

**Normalizing Flow with only Object Encoding.** We evaluate the contribution of relationship features by testing a variant that concatenates only object embeddings, called *Node Only* (see Tab. 3 and appendix Tab. 6). The resulting performance drop confirms that incorporating all triplet components, both objects and their relationship, is crucial for effective anomaly detection.

**Remove Autoencoder.** To validate the autoencoder's necessity, we tested a variant without it (*No AE*). The results in Tab. 3 and appendix Tab. 6 show that this variant performs significantly worse than all other tested configurations. This underscores the critical role of the autoencoder in enabling effective normalizing flow performance.

## 6. Conclusion

We propose the first model using a normalizing flow for solving the SARD task. By using a word embedding model in combination with an autoencoder we are able to create a rich embedding of the triplets that help the normalizing flow to achieve state-of-the-art results on the SARD dataset. Simultaneously we are able to show the universality of *BUSSARD*, improving the applicability to open-word settings. For future work, the approach could be applied to videos by leveraging DSGG. Additionally large language models could be included to generate textual feedback.

**Acknowledgements.** This work was supported by the MWK of Lower Saxony within Hybrint (VWZN4219) and LCIS (VWZN4704), the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC2122) and Quantum Frontiers (EXC2123), the European Union under grant agreement no. 101136006 – XTREME.

## References

[1] Oliver Atkinson, Akanksha Bhardwaj, Christoph Englert, Vishal S Ngairangbam, and Michael Spannowsky. Anomaly detection with convolutional graph neural networks. *Journal of High Energy Physics*, 2021(8):1–19, 2021. 3

[2] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. BMAD: benchmarks for medical anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 4042–4053. IEEE, 2024. 3

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 3

[4] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *Int. J. Comput. Vis.*, 130(4):947–969, 2022. 1, 3

[5] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2022, Volume 5: VISAPP, Online Streaming, February 6-8, 2022*, pages 202–213. SCITEPRESS, 2022. 3

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 2

[7] Riley Carlson, John Bauer, and Christopher D. Manning. A new pair of gloves. *CoRR*, abs/2507.18103, 2025. 2, 5, 6, 8

[8] Mu Chen, Liulei Li, Wenguan Wang, and Yi Yang. Diffvsgg: Diffusion-driven online video scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 29161–29172. Computer Vision Foundation / IEEE, 2025. 2

[9] NF Chen, Zhiyuan Du, and Khin Hua Ng. Scene graphs for interpretable video anomaly classification. In *Conference on Neural Information Processing Systems Workshop on Visually Grounded Interaction and Language*, 2018. 1, 3

[10] Andrea Coccaro, Marco Letizia, Humberto Reyes-González, and Riccardo Torre. Comparison of affine and rational quadratic spline coupling and autoregressive flows through robust statistical tests. *Symmetry*, 16(8):942, 2024. 8

[11] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11169–11183, 2023. 2, 8, 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[13] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. 5

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 5, 6

[15] Dario Fontanel, Fabio Cermelli, Massimiliano Mancini, and Barbara Caputo. Detecting anomalies in semantic segmentation with prototypes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 113–121. Computer Vision Foundation / IEEE, 2021. 3

[16] Sarthak Garg, Helisa Dhamo, Azade Farshad, Sabrina Musatian, Nassir Navab, and Federico Tombari. Unconditional scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16342–16351. IEEE, 2021. 1, 3

[17] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 1932–1940. AAAI Press, 2024. 3

[18] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 20405–20415. Computer Vision Foundation / IEEE, 2025. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[20] Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec AD 2 dataset: Advanced scenarios for unsupervised anomaly detection. *CoRR*, abs/2503.21622, 2025. 3

[21] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *IEEE/CVF International*

*Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 13499–13508. IEEE, 2023. 1, 3

[22] Christian Horvat and Jean-Pascal Pfister. Denoising normalizing flow. In *Advances in Neural Information Processing Systems*, pages 9099–9111. Curran Associates, Inc., 2021. 3, 5

[23] Yu Hsuan Hsieh and Shang-Hong Lai. CSAD: unsupervised component segmentation for logical anomaly detection. In *35th British Machine Vision Conference, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA Press, 2024. 3

[24] Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. EGTR: extracting graph from transformer for scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24229–24238. IEEE, 2024. 1, 2, 3, 6

[25] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19606–19616. IEEE, 2023. 3

[26] Soohwan Jeong, Jongmin Park, Mingyu Choi, Yongjin Kwon, and Sungsu Lim. Llm-powered scene graph representation learning for image retrieval via visual triplet-based graph transformation. *Expert Syst. Appl.*, 286:127926, 2025. 2

[27] 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-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 4129–4137. AAAI Press, 2025. 3

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[29] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10236–10245, 2018. 5

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 6

[31] Mathis Kruse and Bodo Rosenhahn. Multi-flow: Multi-view-enriched normalizing flows for industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pages 3933–3944. Computer Vision Foundation / IEEE, 2025. 3

[32] Yu-Chen Lai, Motoharu Sonogashira, Itthisak Phueaksri, and Yasutomo Kawanishi. Scene-specific anomalous relationship detection using scene graph summarization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pages 3995–4003. Computer Vision Foundation / IEEE, 2025. 1, 3, 4, 6

[33] Chenghao Li, Lei Qi, and Xin Geng. A sam-guided two-stream lightweight model for anomaly detection. *ACM Trans. Multim. Comput. Commun. Appl.*, 21(2):67:1–67:23, 2025. 3

[34] Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21628–21638. IEEE, 2023. 2

[35] Nanjun Li, Faliang Chang, and Chunsheng Liu. A self-trained spatial graph convolutional network for unsupervised human-related anomalous event detection in complex scenes. *IEEE Trans. Cogn. Dev. Syst.*, 15(2):737–750, 2023. 3

[36] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 28076–28086. IEEE, 2024. 2

[37] Yixin Liu, Shirui Pan, Yu Guang Wang, Fei Xiong, Liang Wang, Qingfeng Chen, and Vincent Cheng-Siong Lee. Anomaly detection in dynamic graphs via transformer. *IEEE Trans. Knowl. Data Eng.*, 35(12):12081–12094, 2023. 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6

[39] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. 2

[40] Yukuan Min, Aming Wu, and Cheng Deng. Environment-invariant curriculum relation learning for fine-grained scene graph generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 13250–13261. IEEE, 2023. 2

[41] Furkan Mumcu, Michael J. Jones, Yasin Yilmaz, and Anoop Cherian. Complexvad: Detecting interaction anomalies in video. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025 - Workshops, Tucson, AZ, USA, February 28 - March 4, 2025*, pages 1003–1012. IEEE, 2025. 1, 3

[42] Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2024 - Workshops, Waikoloa, HI, USA, January 1-6, 2024*, pages 995–1004. IEEE, 2024. 1, 3

[43] Sotheany Nou, Joong-Sun Lee, Nagaaki Ohyama, and Takashi Obi. Human pose feature enhancement for human

anomaly detection and tracking. *International Journal of Information Technology*, 17(3):1311–1320, 2025. 1, 3

[44] OpenAI. Chatgpt, 2025. 3

[45] George Papamakarios, Iain Murray, and Theo Pavlakou. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2338–2347, 2017. 3

[46] Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Raghavi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, Yejin Choi, and Ranjay Krishna. Synthetic visual genome. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9073–9086. Computer Vision Foundation / IEEE, 2025. 1, 2

[47] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014. 2

[48] Masoud Pourreza, Mohammadreza Salehi, and Mohammad Sabokrou. Ano-graph: Learning normal scene contextual graphs to detect video anomalies. *arXiv preprint arXiv:2103.10502*, 2021. 3

[49] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 413–420. IEEE Computer Society, 2009. 6, 1

[50] Pallavi Raj. Anoedgepred: A novel method for detecting anomalous edges in social networks. *IETE Technical Review*, pages 1–17, 2025. 3

[51] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1530–1538. JMLR.org, 2015. 3

[52] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14298–14308. IEEE, 2022. 3

[53] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1906–1915. IEEE, 2021. 3

[54] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 2591–2601. IEEE, 2023. 3

[55] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 8, 2

[56] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press, 2017. 6

[57] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6479–6488. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 3

[58] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 184–192. ACM, 2020. 3

[59] Shuzhou Sun, Shuaifeng Zhi, Qing Liao, Janne Heikkilä, and Li Liu. Unbiased scene graph generation via two-stage causal modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12562–12580, 2023. 2

[60] Kees M Van Hespen, Jaco JM Zwanenburg, Jan W Dankbaar, Mirjam I Geerlings, Jeroen Hendrikse, and Hugo J Kuijf. An anomaly detection approach to identify chronic brain infarcts on mri. *Scientific Reports*, 11(1):7714, 2021. 3

[61] Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. OED: towards one-stage end-to-end dynamic scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27938–27947. IEEE, 2024. 2

[62] Kai Wang, Shuaiyi Lyu, Yang Liu, and Bailing Wang. Graph optimization via decoupled edge tuning for efficient industrial anomaly detection. *IEEE Trans. Netw. Sci. Eng.*, 12(5): 4028–4043, 2025. 3

[63] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 8, 2

[64] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8032–8041. IEEE, 2023. 3

[65] Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 14158–14168. Computer Vision Foundation / IEEE, 2025. 1, 2

[66] Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-Seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual

scene. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24539–24549. Computer Vision Foundation / IEEE, 2025. 2

[67] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 649–655. IEEE, 2022. 3

[68] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. 8, 2

[69] Dongil Yang, Minjin Kim, Sunghwan Kim, Beong-woo Kwak, Minjun Park, Jinseok Hong, Woontack Woo, and Jinyoung Yeo. LLM meets scene graph: Can large language models understand and generate scene graphs? A benchmark and empirical study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21335–21360. Association for Computational Linguistics, 2025. 1, 2

[70] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18675–18685. IEEE, 2023. 2

[71] Yi Yang, Yuren Cong, Hao Cheng, Bodo Rosenhahn, and Michael Ying Yang. FDSG: forecasting dynamic scene graphs. *CoRR*, abs/2506.01487, 2025. 2

[72] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 15204–15213. Computer Vision Foundation / IEEE, 2025. 3

[73] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 589–597. IEEE Computer Society, 2016. 1, 3

[74] Yiheng Zhang, Yunkang Cao, Xiaohao Xu, and Weiming Shen. Logicode: An llm-driven framework for logical anomaly detection. *IEEE Trans Autom. Sci. Eng.*, 22:7712–7723, 2025. 3

[75] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[76] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 3