

Scene-Specific Anomalous Relationship Detection Using Scene Graph Summarization

Yu-Chen Lai^{1,2} Motoharu Sonogashira¹ Itthisak Phueaksri¹ Yasutomo Kawanishi¹

¹RIKEN GRP, Japan, ²National Cheng Kung University, Taiwan

marow17623@gmail.com

{motoharu.sonogashira, itthisak.phueaksri@riken.jp, yasutomo.kawanishi}@riken.jp

Abstract

This paper defines a novel task, Scene-Specific Anomalous Relationship Detection (SARD), designed to detect anomalous relationships between objects within a given scene by analyzing an image collection of that scene. The definition of “anomalous” can vary from scene to scene. For example, placing a cup near a book (“cup-near-book” relationship) is typical in general contexts. However, it may be considered anomalous in a specific scene, such as a library where drinking is usually prohibited. By observing a specific scene for a while, humans can discern the underlying rules or commonsense contexts within that scene. Building on this insight, this study proposes a method for SARD based on scene graph summarization that can capture scene-specific contexts. The method generates a scene graph from each image and then combines them into a summarized scene graph, retaining the occurrence count for each relationship. It calculates the anomaly score of each relationship within the summarized scene graph based on its rarity and outputs ranked relationships. To confirm the performance, this paper introduces a new dataset for the SARD task. By comparing the proposed method with baseline methods on the dataset, the proposed method achieves an increase of 19.77 in the AUC of Recall@k. We provide the code and dataset on <https://github.com/marow17623/SARD.git>.

1. Introduction

Recognizing unusual deviations is crucial for mitigating potential threats and ensuring stability in the everyday environment. Numerous studies have emphasized anomalous object detection [1, 16, 26], as this approach can identify a significant number of anomalies like a sofa on the street. However, some anomalies arise from relationships between objects that cannot be identified through anomalous object detection alone. Placing a book on a laptop keyboard may

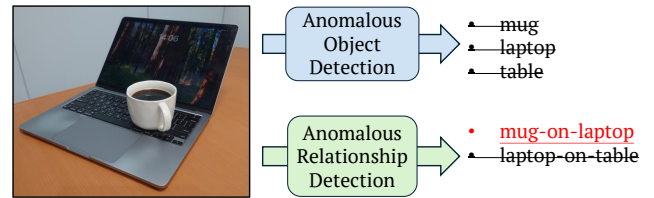


Figure 1. Illustration of the difference between anomalous object detection and anomalous relationship detection. This paper focuses on detecting anomalous relationships.

cause excessive pressure and should be regarded as anomalous. However, because a book and a laptop often appear together, this type of anomaly is challenging to detect for previous methods. Anomalous relationship detection can address this gap by analyzing object relationships. It offers a more comprehensive view of anomalies by considering both objects and their interactions, as shown in Figure 1. To detect the anomalous scene graphs, Garg et al. [4] used an auto-regressive scene graph generation model called Scene-GraphGen. In this method, anomalous relationships are identified based on the common sense learned by the generative model. However, typical normal relationships should be considered anomalous in specific scenes. For example, placing a cup near a book may be anomalous in a library where drinking is usually prohibited. The definitions of normal and anomalous may vary from scene to scene.

We aim to make anomalous relationship detection more flexible, adapting it to different anomalous definitions for each scene. By observing a specific scene for a while, humans can discern the underlying rules within that scene. This implies that multiple observations of the target scene provide contexts for the target scene. Therefore, we propose using an image collection from a specific scene as input to provide contextual information. With this in mind, we define a new task, Scene-specific Anomalous Relationship Detection (SARD), as shown in Figure 2. The methods of SARD need to recognize the unusual relationships from the

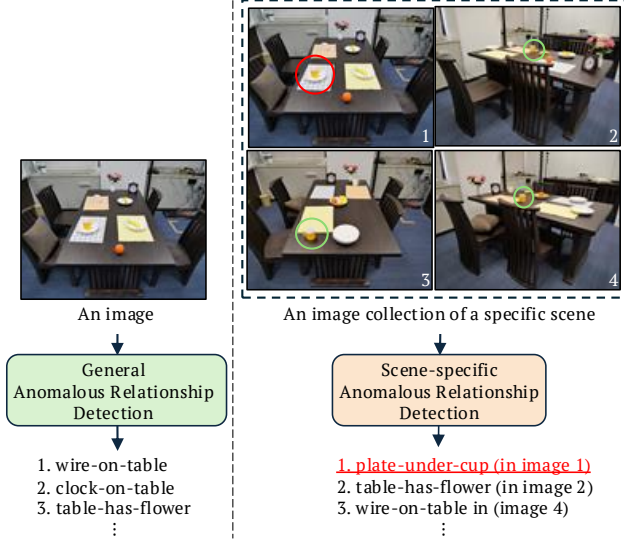


Figure 2. Illustration of the difference between general anomalous relationship detection (left) and scene-specific anomalous relationship detection (right). By analyzing an image collection, the SARD method can obtain scene patterns and identify “plate-under-cup” as the most anomalous.

given image collection. We provide a more detailed explanation of the problem definition in Sec. 3. Since no previous methods are directly applicable to this new task, we introduce a baseline method called counting-based SARD. This method uses the generic Scene Graph Generation (SGG) model to extract objects and their relationships for each image. We then summarize the relationships in each image and identify anomalies by analyzing their frequency, with unusual ones recognized as anomalous.

However, simply applying the counting-based SARD often does not work. We found some issues with the method and proposed improvements in Sec. 5. First, generic SGG models have a bias toward common relationships and tend to assign higher confidence to them, affecting the ability of the baseline method to judge anomalies. Second, some noisy normal relationships detected by the SGG model may appear infrequently but are not truly anomalous. Lastly, relying only on the highest-confidence predicate limits the completeness of object relationship descriptions. Thus, we additionally propose some enhancements for the counting-based SARD method. For the evaluation of this task, we provide a novel dataset and establish a comprehensive evaluation in Sec. 6.

Our contributions are as follows:

1. **Scene-specific Anomalous Relationship Detection (SARD) task:** We introduce a new task to detect anomalous object relationships. This approach provides a more comprehensive description of anomalies for a scene.
2. **Counting-based SARD baseline:** We propose a base-

line that effectively identifies anomalies by analyzing the frequency of relationships across an image collection.

3. **Improvements for the counting-based SARD:** To address the several issues of baseline, we propose solutions as follows: balancing confidence bias from prior knowledge, reducing the influence of noisy normal relationships by using knowledge graph, and considering several predicates to represent each relationship completely.
4. **Dataset:** We built a dataset and suitable evaluation metrics to assess SARD performance.

2. Related Work

In this section, we introduce anomalous relationship detection methods related to but distinct from our task, as well as scene graph generation research that we reference or incorporate into our approach.

2.1. Anomaly Detection

While no existing methods directly address the task we propose, several anomaly detection approaches offer insights into related challenges. Bozkan et al. [1] proposed to detect anomalous objects for indoor scenes using image representations and location vectors (GPS coordinates) to learn typical scene patterns. Single-scene anomalous video detection [20] targets anomalies within videos of 120–200 frames, including unusual objects and interactions. However, they lack finer-grained relationship detection and require training for specific scenes. Garg et al. [4] introduced SceneGraphGen, an auto-regressive model that generates realistic scene graphs without an input image. It can also detect anomalous scene graphs by using negative log-likelihood to detect deviations from typical data patterns. However, it is limited to detecting commonsense anomalous relationships. Mullen Jr. et al. [7] presented a method for home robots to detect unsafe or unsanitary anomalies by using scene graphs and the Large Language Model (LLM) [17] to analyze spatial relationships between objects. Li et al. [13] introduced a method that uses an affordance-enhanced scene graph to detect misplaced items and guide household object rearrangement based on commonsense knowledge of LLM and user preferences. While effective at detecting anomalous relationships, it is focused on home organization scenarios.

2.2. Scene Graph Generation

Scene Graph Generation (SGG) is a well-known task that represents an image by detecting objects and their relationships. In a scene graph, the objects are represented as nodes, and the edges between nodes represent relationships. A scene graph is described using multiple triplets, each consisting of a subject, predicate, and object (e.g., “man-holding-cup” or “book-on-sofa”). This structured representation enables high-level image understanding, supporting

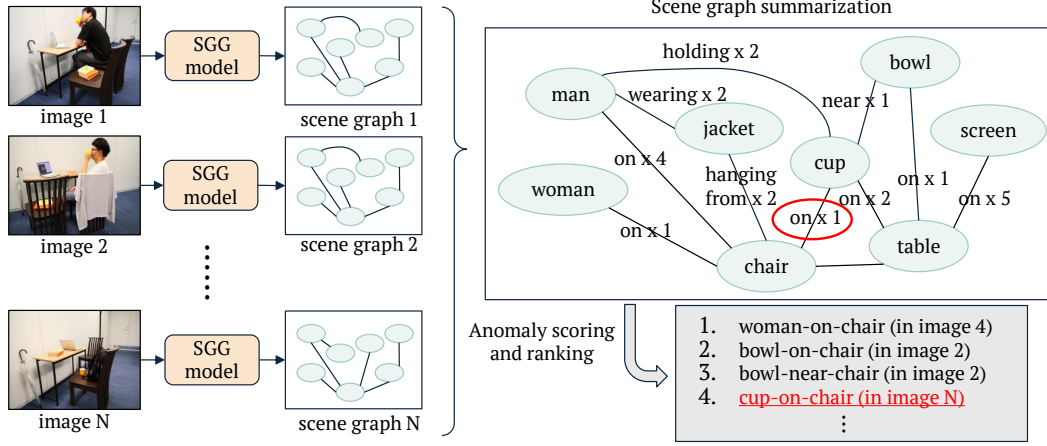


Figure 3. The framework of the counting-based scene-specific anomalous relationship detection (C-SARD). It uses a pre-trained SGG model to generate a scene graph of each image. The occurrence of each relationship type is then counted through scene graph summarization. Finally, the anomaly score of each object relationship is computed based on its rarity, with the most anomalous ones ranked at the top.

various downstream tasks such as image captioning [15] and visual question answering [12]. Common SGG datasets like Visual Genome [9] and Open Images [11] offer extensive image annotations of objects and relationships, ideal for training models to recognize complex interactions in diverse scenes.

A typical scene graph generation architecture has two-stage processes: object detection and relationship prediction. Recently, Faster R-CNN [21] and DETR [2] trained on Microsoft Common Objects in Context (MS-COCO) dataset [14] have been widely used as backbones for object detection. EGTR [6] is a one-stage method that extracts relational graphs from the multi-head self-attention layers of the DETR decoder using a lightweight relationship predictor to enhance the efficiency of scene graph generation. In our method, we use the state-of-the-art SGG method, EGTR, as our generic SGG model to extract scene graphs.

3. Problem Definition

In general anomaly detection, anomalies are defined as rare samples. Similarly, in Scene-specific Anomalous Relationship Detection (SARD), the goal is to identify and rank rare object relationships in an image collection. An image collection with N images from a specific scene is denoted as $\mathcal{I} = \{I_k\}_{k=1}^N$, where I_k represents the k -th image in the collection. Using a relationship detector θ , a set of object relationships, i.e., a scene graph $G_k = \theta(I_k)$, is estimated for each I_k . An object relationship requires a complete description, including subject s , predicate p , and object o . Following SGG methods [3, 6], structure $t = \langle s, p, o \rangle$ is used to represent an object relationship.

To rank the object relationships, the SARD method re-

quires a scoring function A_{rel} to calculate the anomaly score α_k^i of i -th object relationship $t_k^i = \langle s_k^i, p_k^i, o_k^i \rangle$ within k -th image I_k in the image collection $\mathcal{I}$ as follows:

$$\alpha_k^i = A_{\text{rel}}(t_k^i; \mathcal{I}). \quad (1)$$

The method then ranks the object relationships by their anomaly scores, with the higher anomaly scores ranked at the top.

4. Baseline: Counting-based Scene-specific Anomalous Relationship Detection

To our knowledge, we are the first to propose the SARD task. We introduce a baseline method called Counting-based SARD (C-SARD), which intuitively counts the occurrences of each relationship type in the image collection $\mathcal{I}$, considering the least frequent relationships as the most anomalous. Figure 3 illustrates the workflow of the baseline method. Scene Graph Generation (SGG) is performed for each image, followed by scene graph summarization and anomaly scoring to identify anomalous relationships. The following sections explain how to summarize the image collection and measure anomalies by counting.

4.1. Scene Graph Generation

This method follows the standard process of SGG methods [3, 6]. In this baseline, using the state-of-the-art SGG method EGTR [6], pre-trained on Visual Genome [9], as the SGG model, a scene graph G_k is generated from an image I_k . In the EGTR output, both the subject s and object o include class IDs $s_{\text{cls}}, o_{\text{cls}} \in \mathbb{N}$, confidence score vectors $s_{\text{score}}, o_{\text{score}} \in \mathbb{R}^{N_o}$, and bounding boxes $s_{\text{bbox}}, o_{\text{bbox}} \in \mathbb{R}^4$. The predicate p only includes the class ID $p_{\text{cls}} \in \mathbb{N}$ and confidence score vector $p_{\text{score}} \in \mathbb{R}^{N_p}$, as predicates lack spatial

positions. Here, N_o and N_p represent the total numbers of object and predicate categories, respectively. The class IDs $s_{cls}, p_{cls}, o_{cls}$ are determined by applying the $\arg \max$ function to their respective confidence score vectors.

To avoid including imprecise triplets, we retain only the top n_k triplets with the highest confidence from each I_k , where n_k is the lesser of the total relationships in k -th image and the maximum allowable number n_{max} . The triplet confidence is the product of the maximum confidence values among $s_{score}, p_{score}, o_{score}$.

4.2. Scene Graph Summarization and Counting

In this step, scene graphs are summarized to count the occurrence of each relationship type. To obtain all object relationships within the scene, the triplet set $\mathcal{T}$ is obtained as follows:

$$\mathcal{T} = \bigcup_{k=1}^N G_k = \{t_k^i \mid i = 1, \dots, n_k; k = 1, \dots, N\}, \quad (2)$$

where t_k^i denotes the i -th triplet in G_k , and n_k is the number of triplets in I_k .

The occurrences of each relationship type from $\mathcal{T}$ are then counted. For example, the relationship type “cup-near-book” appears once in each of the five images. This yields five triplets in $\mathcal{T}$, with their s_{cls}, p_{cls} , and o_{cls} corresponding to “cup,” “near,” and “book,” respectively. To efficiently manage this counting process, a counting tensor $C \in \mathbb{R}^{N_o \times N_p \times N_o}$ is introduced. The tensor records the occurrence count for each relationship type by incrementing the corresponding entry $C[s_{cls}, p_{cls}, o_{cls}]$, where s_{cls}, o_{cls} , and p_{cls} correspond to the class IDs of the subject, object, and predicate, respectively.

4.3. Anomaly Scoring

The relationships with fewer occurrences are regarded as more anomalous. From this, a counting tensor C is introduced to determine the anomaly score of each triplet t_k^i . The score is calculated as the inverse of the counting number $\text{count}_k^i = C[s_{cls}, p_{cls}, o_{cls}]$, where s_{cls}, p_{cls} , and o_{cls} are the object and predicate classes of s_i^k, p_i^k , and o_i^k , respectively. However, some triplets may have the same counting number, resulting in equal anomaly scores. Therefore, the confidence score conf_k^i of each triplet t_k^i is considered when calculating the anomaly score. The confidence score conf_k^i is defined as the product of the maximum values of three score vector s_{score}, p_{score} , and o_{score} corresponding to t_k^i . A higher confidence score indicates a more certain anomaly score, so it is directly proportional to the anomaly score. Finally, the anomaly score of the object relationship t_k^i is calculated as:

$$A_{\text{rel}}(t_k^i) = \frac{1}{\exp(\text{count}_k^i)} + \text{conf}_k^i. \quad (3)$$

Notably, we apply an exponential function to the count_k^i to ensure its influence is significantly greater than that of the conf_k^i . After that, we rank the triplets by their anomaly scores $A_{\text{rel}}(t_k^i)$.

5. Enhanced Counting-based SARD

Counting-based SARD achieves anomalous relationship detection by counting the occurrences of each relationship type. However, we found some issues with this approach and proposed several enhanced methods.

- **The bias toward common relationships:** The relationships commonly found in real-world scenarios tend to receive higher confidence scores from generic SGG models, potentially overshadowing the real anomalous ones. To address these issues, we propose rarity adjustment in Sec. 5.1 to balance this bias.
- **The influence of noisy normal relationships:** Certain objects play a minor role and add noise to scene analysis. Additionally, some relationships occur infrequently but are actually normal, which may be mistakenly identified as anomalies. To address this, we propose a noisy normal relationship reduction in Sec. 5.2.
- **Single predicate limitation:** Using only the highest-confidence predicate may not accurately describe each relationship since a single relationship can correspond to several predicates. We propose soft counting in Sec. 5.3 to account for several probable predicates.

5.1. Rarity Adjustment

The generic SGG model, pre-trained on a general dataset, tends to assign higher confidence scores to commonly found relationships, as it primarily learns from frequent examples. In contrast, rare relationships appear less in the training data, so the model has lower confidence in predicting them. Because our anomaly score function, shown in Equation (3), references the confidence score of relationships, we need to reduce the inherent bias of the pre-trained SGG model.

We are inspired by [27], which utilizes relationship frequency as prior knowledge in prediction tasks. We apply a similar idea to address this bias by adjusting confidence scores according to their rarity in general scenes, making them more balanced. Collecting statistics from the relationship ground truth in the Visual Genome dataset enables us to obtain prior knowledge counting tensor $PC \in \mathbb{R}^{N_o \times N_p \times N_o}$ that reflects the statistical distribution of each relationship type. From PC , we derive a rarity tensor $R \in \mathbb{R}^{N_o \times N_p \times N_o}$, assigning lower values to common relationships and higher values to rarer ones. The tensor R is defined as:

$$R[s_{cls}, p_{cls}, o_{cls}] = \frac{1}{PC[s_{cls}, p_{cls}, o_{cls}]}, \quad (4)$$

if $PC[s_{cls}, p_{cls}, o_{cls}] \neq 0$, else 0.

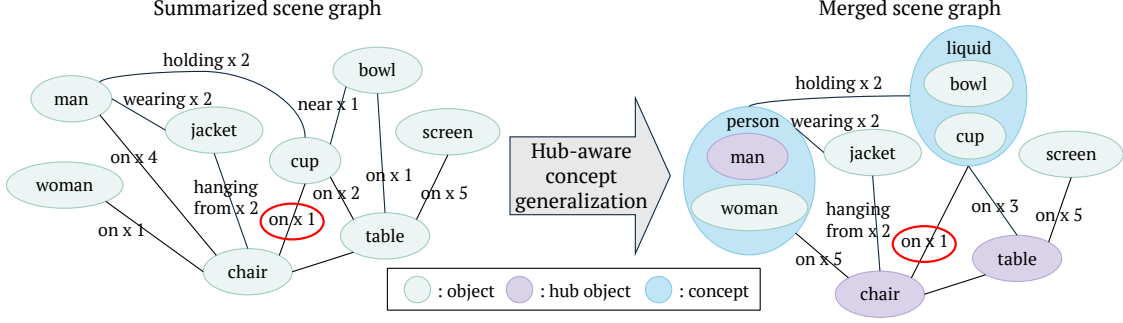


Figure 4. Illustration of hub-aware concept generalization. Left: scene graph summarization, with nodes as objects and edges indicating relationships with their counts. Some infrequent relationships are highlighted (red circle). Right: the result after applying hub-aware concept generalization, treating similar concept objects as the same type and sharing their counts, making the relationship between “woman” and “bowl” less unusual and reducing the impact of noisy normal relationships.

Non-existent relationships are given a rarity of 0 rather than an extremely large value because we consider such relationships unreliable. An example from our results is “chair-holding-banana.” Such object relationships might be reasonable from a certain perspective (e.g., imagining the chair as a person), but humans would not describe the relationship in this way, as it falls outside the scope of prior knowledge. Thus, we remove any t_k^i with a rarity of 0 from the triplet set $\mathcal{T}$. For the remaining t_k^i , the confidence score conf_k^i in Equation (3) is adjusted as follows:

$$\text{adjusted\_conf}_k^i = \text{conf}_k^i \cdot R[s_{\text{cls}}, p_{\text{cls}}, o_{\text{cls}}]. \quad (5)$$

We adjust the confidence score conf_k^i by incorporating rarity information from R , ensuring balanced confidence scores for relationships. This process adjusts the confidence scores of the triplets and removes those outside the scope of prior knowledge before anomaly scoring in Sec. 4.3.

5.2. Noisy Normal Relationship Reduction

In this section, we address the issue of noisy normal relationships. Primarily, there are two sources of noisy normal relationships. One source of noisy normal relationships arises from minor objects, such as “a leg of a table” or “an ear of a head”. These objects and their relationships are challenging for SGG models to detect due to their subtle features or occlusion. Since these minor objects are part of larger main objects, their relationships are typically encompassed within those of the main objects. Therefore, we can exclude all relationships involving minor objects from the triplet set $\mathcal{T}$. To identify minor objects, we leverage ConceptNet [24], a large knowledge graph that connects words and phrases through various relationships, helping computers understand human language. By using the “PartOf” relationship in ConceptNet, we determine an object as a minor object if it is part of another object in the scene. This process operates on $\mathcal{T}$ in Sec. 4.2 to obtain a filtered triplet set $\mathcal{T}_{\text{filtered}}$ without relationships involving minor objects.

The detailed algorithm is shown in Supplementary Section B.

The second source of noisy normal relationships arises from objects that occur infrequently but are actually normal. For example, a single “banana-on-table” does not seem unusual if multiple oranges are also present, as both are fruits and commonly appear together. Additionally, the SGG model may misclassify objects, especially when viewed from specific angles or partially obscured, such as mistaking a cup for a bowl or a man for a woman. These infrequent relationships should not be treated as anomalous relationships due to their semantic or visual similarity.

Inspired by previous works [18, 19] that use ConceptNet to extract common object concepts for image summarization, we also utilize ConceptNet to find relevant concepts for each object. While preventing over-grouping and maintaining detail, we select only the concepts most closely associated with each object. Additionally, we avoid merging frequently occurring objects, termed “hub objects.” An object is regarded as a hub object if its occurrence frequency exceeds a threshold τ , as shown in the formula below. Note that the frequency is specifically calculated from $\mathcal{T}$.

$$\text{HubObjs} = \{\text{obj} \mid \text{freq}(\text{obj}) > \tau, \forall \text{obj} \in \text{ObjClass}\}. \quad (6)$$

Here, ObjClass contains all the object class IDs from $\mathcal{T}$. The occurrence frequency of an object, $\text{freq}(\text{obj})$, is defined as:

$$\text{freq}(\text{obj}) = \frac{1}{\sum_{k=1}^N n_k} \sum_{k=1}^N \sum_{i=1}^{n_k} \delta(\text{obj} \in t_k^i), \quad (7)$$

where $\delta(\cdot)$ is the indicator function that equals 1 if the condition is true and 0 otherwise.

To address these aforementioned issues, we propose a hub-aware concept generalization algorithm, as shown in Figure 4. This algorithm selects highly relevant concepts, accounts for hub objects, and groups objects with the same

concept. First, the algorithm searches ConceptNet for concepts related to each object through the ‘RelatedTo’ relationship. It then constructs an undirected graph with objects and concepts as nodes, where edges reflect relationship strength. Concepts are selected iteratively. In each step, the concept with the highest average weight is selected. The concept and its associated objects are stored in the concept bank and removed from the graph. If the selected concept connects more than two hub objects, it is removed and not stored in the concept bank, as hub objects are not infrequent and are not the objects we want to handle. Before the next iteration, unimportant concepts (those with zero or one degree) are also removed from the graph. These steps are repeated until no concepts remain in the graph. This approach can select concepts that cover less frequent but typical objects. By referencing the concept bank, we group objects according to their shared concepts.

In the implementation, the algorithm achieves grouping by allowing objects within the same concept to share their counting numbers. The algorithm performs after the scene-graph summarization stage. By the algorithm, a summarized scene graph shown on the left side of Figure 4 is converted to a merged scene graph shown on the right side of Figure 4. Finally, the counting tensor C is updated by referring to the merged scene graph. The detailed algorithm is shown in Supplementary Section B.

5.3. Soft Counting

For each object relationship, the baseline C-SARD explained in Sec. 4 only considers the predicate with the highest confidence score, i.e., *hard counting*. However, the highest-confidence predicate is not always apparently definitive; certain predicates, like “on” and “above,” can be easily confused from different perspectives. Thus, it may be beneficial to consider several possible predicate candidates simultaneously, i.e., *soft counting*, for a more robust representation of an object relationship. We consider the top K_{soft} predicates and increment C by one for each.

Table 1. The number of images containing an anomalous relationship in the office and the dining room scenes.

Office scene		Dining room scene	
Relationship	Num	Relationship	Num
plate-under-cup	18	bag-on-table	20
pillow-on-table	15	cup-on-chair	15
fruit-on-chair	10	man-holding-umbrella	12
plate-on-chair	6	shoe-on-table	11
plate-has-flower	4	shoe-on-chair	3
paper-on-chair	1	book-on-laptop	2

6. Experimental Results

This section presents the proposed dataset and evaluation metrics for the SARD task. We compare Enhanced C-SARD not only with C-SARD but also with traditional approaches on image-level detection to provide a comprehensive analysis.

6.1. Dataset and Evaluation Metrics

Dataset. For the SARD task, we propose an Indoor Scene Dataset. This dataset contains two scenes: an office scene and a dining room scene. In the office scene, the objects include a table, chair, laptop, book, man, cup, jacket, and umbrella. In the dining room scene, the objects include a table, chair, paper, flower, vase, banana, orange, pillow, plate, cup, and clock. We defined a set of object relationships as normal relationships and other relationships as anomalous relationships. The images are split into normal and anomalous images for each scene. The normal images only contain normal relationships, while the anomalous images contain exactly one anomalous relationship in each image. The office scene includes 60 normal and 63 anomalous images, while the dining room scene includes 60 normal and 54 anomalous images. The size of each image is $5,184 \times 3,888$ pixels.

Table 1 presents the number of images containing each anomalous relationship. Using the normal and anomalous image sets, we built image collections. In a collection of N images, one is anomalous, while the rest $N - 1$ are normal images randomly sampled from the normal image set, meaning the collection includes an anomalous relationship. We hope SARD can detect the anomalous relationship within this image, prioritizing it as high as possible in the anomalous ranking. This results in 63 image collections for the office scene and 54 for the dining room scene. We evaluate our proposed methods using three values of N : 11, 21, and 31.

Evaluation metrics. To assess the performance of SARD, we employ Recall@1 and the Area Under the Curve (AUC) of Recall@k. Previous SGG studies [3, 6, 27] use Recall@k to measure the percentage of correct relationships among the top k predictions. In our task, we use Recall@1 to measure how accurately the model detects the anomalous relationship by checking whether it ranks at the top. To comprehensively evaluate under multiple k values, we use the AUC of Recall@k to calculate the area under the recall curve across varying k values. A higher AUC indicates that the anomalous relationship tends to be ranked closer to the top. In our experiments, k is varied from 1 to 100. We do not use the same evaluation metric as the general anomalous scene graph detection method [25], which classifies an image based on whether it contains an anomalous relationship rather than ranking multiple relationships.

Table 2. Performance comparison of the original Counting-based SARD (pure hard counting) and enhanced versions with rarity adjustment, noisy normal relationship reduction, and soft counting. The top row shows the original C-SARD (baseline, explained in Sec. 4), and the bottom row shows the Enhanced C-SARD (full model, explained in Sec. 5). Recall values are presented as ratios (0 to 1), while the AUC of Recall@k values are from 1 to 100. Bold indicates the best, and underscore indicates the second best.

Counting type	Rarity adjustment	Noisy normal relationship reduction	AUC-Recall@k			Recall@1		
			$N = 11$	$N = 21$	$N = 31$	$N = 11$	$N = 21$	$N = 31$
Hard counting	✓	✓	61.25	50.55	39.76	0.01	0.02	0.02
			63.73	56.25	47.25	0.18	0.16	<u>0.16</u>
	✓	✓	66.22	55.40	43.42	0.04	0.02	0.05
			68.18	63.29	56.59	<u>0.19</u>	0.16	0.17
Soft counting	✓	✓	61.43	47.72	34.20	0.00	0.03	0.01
			64.68	56.65	44.96	0.18	<u>0.17</u>	0.17
	✓	✓	66.98	49.20	37.06	0.07	0.11	0.07
			69.79	<u>61.94</u>	59.53	0.20	0.18	0.17

6.2. Hyperparameter Setting

The maximum number $n_{\max}$ used in the C-SARD (Sec. 4) is set to 30, filtering out low-confidence ones. For noisy normal relationship reduction (Sec. 5.2), we set the maximum number of searched entries in ConceptNet M to 30, and the threshold for defining a hub object τ is set to 0.1. For soft counting (Sec. 5.3), K_{soft} is set to 10 to account for several predicates representing an object relationship. Additional details on hyperparameter selection are provided in Supplementary Section A. Following the default setting of the EGTR object detector (DETR [2]), we resize images to $1,066 \times 800$ pixels.

6.3. Results

The performance of the baseline C-SARD and the proposed improvement methods are shown in Table 2. We propose three improvements in Sec. 5: the rarity adjustment, noisy normal relationship reduction, and soft counting. The baseline method C-SARD uses only hard counting without these two components, while the full method Enhanced C-SARD uses all the options. The table presents average scores across two scenes, showing significant performance increases when three improvements are used together. An output example is shown in Figure 6. Other examples are shown in Supplementary Section D. To demonstrate the robustness of our enhancements, we also try to use a different SGG method, RelTR [3], in Supplementary Section C.

We present results for different collections configuration with 11, 21, and 31 images. Although larger sets contain several times more images, the performance declines slightly, demonstrating the stability of our proposed method. Pure soft counting may sometimes perform worse than hard counting because the triplet set $\mathcal{T}$ may include inaccurate relationships, leading to incorrect counts. Rarity adjustment filters out relationships beyond the scope of

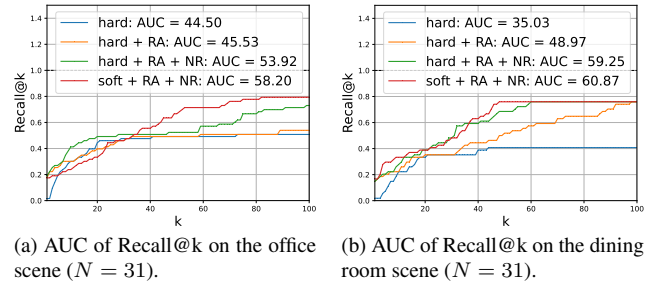


Figure 5. Visualization of results on the office scene and the dining room scene. The y-axis represents Recall@k values (from 0 to 1), while the x-axis represents k values (from 1 to 100). “RA” and “NR” denote rarity adjustment and noisy normal relationship reduction, respectively.

prior knowledge. Noisy normal relationship reduction removes the relationships with misclassified objects. We can observe from the AUC of Recall@k visualization in Figure 5 that their combination achieves optimal performance.

6.4. Anomalous Image Detection

In this section, we explore whether other methods can effectively detect unusual situations by analyzing an image collection. Since other methods do not focus on relationship-level detection, we conduct comparisons at image-level detection. We calculated an anomaly score for each image in the collection. If the anomalous image receives the highest anomaly score in the collection, it indicates that the model has correctly detected it, representing image-level Recall@1. For our method, the anomaly score of image I_k is defined as the average anomaly score of all relationships within the image, as shown below:

$$A_{\text{img}}(I_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} A_{\text{rel}}(t_k^i). \quad (8)$$

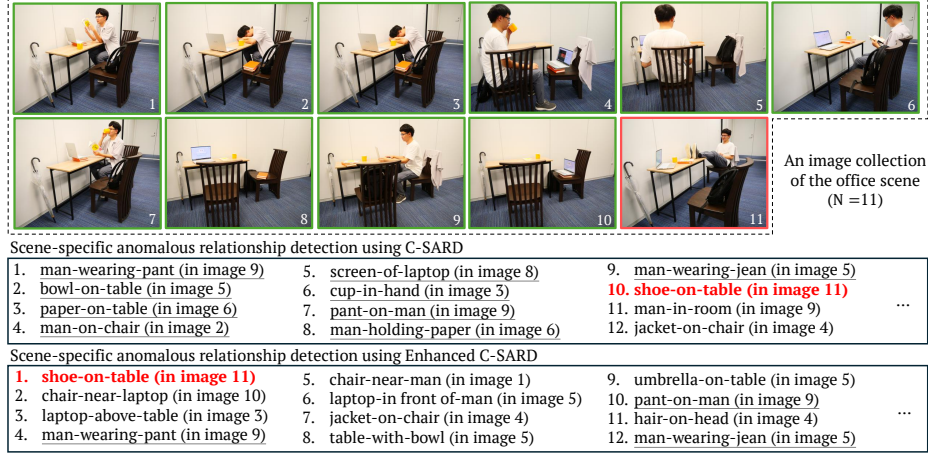


Figure 6. An example of the scene-specific anomalous relationship detection results on the office scene ($N = 11$), including both 10 normal images (green border) and 1 anomalous image (red border). Enhanced C-SARD successfully ranked the anomalous relationship (red text) at the top. The relationships incorrectly ranked above the anomalous one by C-SARD (underline) are moved down.

To benchmark our approach, we wanted to compare it with other image-level detection methods. Since many methods are designed for different tasks [1, 4, 20] or application scenarios [7, 13], it is hard to compare them. Thus, we selected two generally used methods for comparison: One-Class Support Vector Machine (OC-SVM) [22] and Variational Autoencoder (VAE) [8]. Under the same conditions, these methods are trained on the image collection. Since the collection predominantly consists of normal images, with only one noise (an image containing an anomalous relationship), traditional methods were expected to be more familiar with the common, normal features. After training, they were applied to the same image collection.

In the implementation, we extract image features using pre-trained ResNet50 [5] and train an OC-SVM with an RBF kernel using scikit-learn. The ν parameter is set to 0.1, and γ is set to its default value, which depends on the number of features. During testing, we use the negative decision function as the anomaly score. For the VAE model, the training loss combines binary cross entropy [23] as the reconstruction loss with Kullback-Leibler divergence loss [10]. We set a latent dimension of 256 and apply full-batch training for 1,000 epochs. In the testing phase, we serve reconstruction loss as the anomaly score.

We compare the performance of our method with OC-SVM and VAE on image-level Recall@1. Table 3 shows that our method outperformed traditional approaches across different image collection configurations since our method performs scene analysis on the entire image collection.

7. Conclusion

In this paper, we present Scene-specific Anomalous Relationship Detection (SARD) task that identifies anomalous

Table 3. Comparison of image-level Recall@1 performance between our method, OC-SVM, and VAE. The bold is the best.

Method	Image-level Recall@1		
	$N = 11$	$N = 21$	$N = 31$
ResNet + OCSVM	0.11	0.09	0.02
VAE	0.06	0.03	0.04
Our	0.18	0.14	0.13

relationships within specific scenes. We developed a baseline method, Counting-based SARD (C-SARD), which ranks anomalies by analyzing relationship frequencies across an image collection. In addition, we enhanced C-SARD by using rarity adjustment, noisy normal relationship reduction, and soft counting to address issues such as biases toward common relationships, noisy normal relationships, and single-predicate limitation. Furthermore, we created a new dataset to validate that our approach maintains stability across various dataset configurations. We also showed its effectiveness in image-level anomaly detection by comparing it with traditional methods like OC-SVM and VAE. In future work, we will explore more effective summarization techniques, refine our anomaly scoring approach, and adopt open-vocabulary SGG to further improve the generalization of Enhanced C-SARD.

Acknowledgements

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research (JP24H00733 and JP22K17920).

References

- [1] Ilker Bozcan, Jonas Le Fevre, Huy X. Pham, and Erdal Kayacan. GridNet: Image-agnostic conditional anomaly detection for indoor surveillance. *IEEE Robot. Autom. Lett.*, 6(2): 1638–1645, 2021. 1, 2, 8
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. of the 18th Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020. 3, 7
- [3] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. ReTR: Relation transformer for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11169–11183, 2023. 3, 6, 7
- [4] Sarthak Garg, Helisa Dhama, Azade Farshad, Sabrina Musatian, Nassir Navab, and Federico Tombari. Unconditional scene graph generation. In *Proc. of the 18th Int. Conf. Comput. Vis.*, pages 16362–16371, 2021. 1, 2, 8
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the 2016 IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 8
- [6] Jinbae Im, JeongYeon Nam, Nokyoung Park, Hyungmin Lee, and Seunghyun Park. EGTR: Extracting graph from transformer for scene graph generation. In *Proc. of the 2024 IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24229–24238, 2024. 3, 6
- [7] James F. Mullen Jr, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Dinesh Manocha, and Reza Ghanadan. “Don’t forget to put the milk back!” Dataset for enabling embodied agents to detect anomalous situations. *IEEE Robot. Autom. Lett.*, 9:9087–9094, 2024. 2, 8
- [8] Diederik P Kingma. Auto-encoding variational bayes. *Proc. of the 2nd Int. Conf. Learn. Represent.*, pages 1–14, 2014. 8
- [9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123:32–73, 2017. 3
- [10] Solomon Kullback and Richard A Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951. 8
- [11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The Open Images Dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.*, 128(7):1956–1981, 2020. 3
- [12] Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. Symbolic Replay: Scene graph as prompt for continual learning on vqa task. In *Proc. of the 37th AAAI Conf. Artif. Intell.*, pages 1250–1259, 2023. 3
- [13] Wenhao Li, Zhiyuan Yu, Qijin She, Zhinan Yu, Yuqing Lan, Chenyang Zhu, Ruizhen Hu, and Kai Xu. LLM-enhanced scene graph learning for household rearrangement. *Proc. of the 2024 ACM SIGGRAPH Conf. and Exhibit. Comput. Graph. Interact. Tech. Asia*, pages 1–19, 2024. 2, 8
- [14] 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 *Proc. of the 13th Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 3
- [15] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? In *Proc. Assoc. Comput. Linguist. 10th Int. Joint Conf. Nat. Lang. Process.*, pages 504–515, 2020. 3
- [16] Jongsuk Oh, Hong-In Kim, and Rae-Hong Park. Context-based abnormal object detection using the fully-connected conditional random fields. *Pattern Recognit. Lett.*, 98:16–25, 2017. 1
- [17] OpenAI. GPT-4 technical report, 2024. 2
- [18] Itthisak Phueaksri, Marc A Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. An approach to generate a caption for an image collection using scene graph generation. *IEEE Access*, pages 128245–128260, 2023. 5
- [19] Itthisak Phueaksri, Marc A. Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. Image-collection summarization using scene-graph generation with external knowledge. *IEEE Access*, 12:17499–17512, 2024. 5
- [20] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2293–2312, 2020. 2, 8
- [21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Inf. Process. Syst.*, pages 91–99, 2015. 3
- [22] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Adv. Neural Inf. Process. Syst.*, pages 582–588, 1999. 8
- [23] C. E. Shannon. A mathematical theory of communication. *Bell Syst. tech. j.*, 27(3):379–423, 1948. 8
- [24] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proc. of the 31st AAAI Conf. Artif. Intell.*, pages 4444–4451, 2017. 5
- [25] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proc. of the 2020 IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3716–3725, 2020. 6
- [26] Haitao Xiong, Jiaqing Wu, Qing Liu, and Yuanyuan Cai. Research on abnormal object detection in specific region based on Mask R-CNN. *Int. J. Adv. Robot. Syst.*, 17(3): 1729881420925287, 2020. 1
- [27] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene graph parsing with global context. In *Proc. of the 2018 IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5831–5840, 2018. 4, 6