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

Supplementary Material

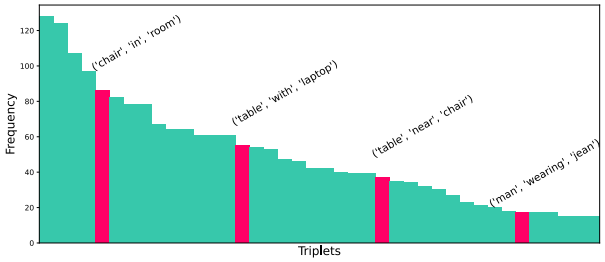


Figure 6. The 40 most frequent triplets of the **office scene**. The labels belong to the highlighted bars, showing example triplets.

7. Additional Information to the Scene Graph Dataset

Figure 6 shows the frequency distribution of the 40 most frequent triplets of the office scene.

8. MIT-67 Dataset

The MIT-67 dataset consists of various indoor scenes, originally designed for indoor scene recognition [49]. Among the overall 67 scenes, there are office and dining room scenes, with each having 109 and 274 samples per scene. The dataset contains no anomalies and has no specific labels.

9. Additional Robustness Results

In this section we provide the additional results for the robustness experiments.

9.1. Synonym Experiment Details

The list of the original words and their chosen corresponding synonyms, used for the experiment, are: *table* → *surface*, *chair* → *stool*, *laptop* → *notebook* and *plate* → *dish*.

Figure 7 shows the results of the synonym experiments for the office scene. The plateau at 50% synonym rate occurs since, the original word and its synonym both appear with similar frequency; consequently, when counting, neither word is sufficiently rare to be falsely classified as anomalous.

9.2. Gaussian Noise Experiments

We test robustness to error-prone SGGs by adding Gaussian noise directly to the concatenated word embedding features. Specifically, we perturb the clean embedding $\mathbf{t}$ as

$$\mathbf{t}_{\text{noisy}} = \mathbf{t} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I) \tag{12}$$

and use $\mathbf{t}_{\text{noisy}}$ as input to the autoencoder. The intuition is that a small deviation of the embedding vectors corresponds to a shift towards semantically similar words, thereby simulating variability

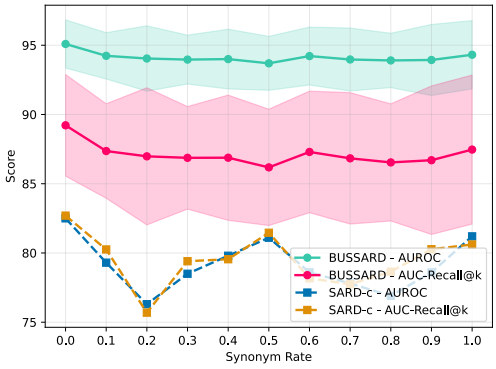


Figure 7. Ablation results with AUROC (↑) and AUC-Recall@k (↑) of *BUSSARD* and SARD-c for different synonym rates for the **office scene**. The synonym rate represents the probability of substituting words using synonym mappings. For *BUSSARD* the dots represent the average results after running with 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.

Table 4. Results for experiments for the scene graph generator robustness, with a different scene graph generator and gaussian noise with different σ. AUROC and AUC-Recall@k on the SARD dataset.

Scene	Experiment	AUROC (↑)	AUC-Recall@k (↑)
Dining Room	RelTR	95.57 ± 0.88	87.29 ± 3.2
	$\sigma = 0.01$	97.43 ± 0.57	93.08 ± 1.41
	$\sigma = 0.05$	97.46 ± 0.47	93.44 ± 1.18
	$\sigma = 0.10$	96.11 ± 1.22	90.37 ± 2.54
Office	RelTR	96.09 ± 1.16	89.47 ± 3.11
	$\sigma = 0.01$	95.39 ± 1.54	89.77 ± 3.22
	$\sigma = 0.05$	95.60 ± 1.46	90.24 ± 3.05
	$\sigma = 0.10$	92.91 ± 1.69	84.97 ± 3.45

in object and relationship descriptions. Tab. 4 shows the results for different values of σ. As the noise does not yield a consistent improvement, we exclude it from the final model.

9.3. Encoding Independence

Scene Graph Generator Independence. To show the independence of *BUSSARD* from the chosen scene graph generator, we replaced the modular SGG component with RelTR [11] and provide the results in Tab. 4. The similar performance underlines the stability of *BUSSARD* regarding different scene graph generators.

Word Embedding Independence. To demonstrate independence from the chosen word embedding model, we conducted ex-

Table 5. Results for experiments with different word embedding models. All variations were executed with ten different seeds and the average and standard deviation is provided. The best performing model is **highlighted**.

Scene	Embedding Model	AUROC (↑)	AUC-Recall@k (↑)
Dining Room	all-MiniLM-L6-v2 [63]	96.34 ± 0.69	90.98 ± 1.35
	all-mpnet-base-v2 [55]	97.25 ± 0.55	93.16 ± 1.22
	Qwen3-Embedding-8B [68]	**97.99** ± 0.58	**94.77** ± 1.48
	GloVe [7]	97.85 ± 0.6	92.85 ± 1.52
Office	all-MiniLM-L6-v2	95.43 ± 2.30	89.68 ± 4.81
	all-mpnet-base-v2	95.40 ± 2.74	89.69 ± 5.79
	Qwen3-Embedding-8B	**96.09** ± 1.71	**91.30** ± 3.52
	GloVe	95.29 ± 1.65	89.57 ± 3.46

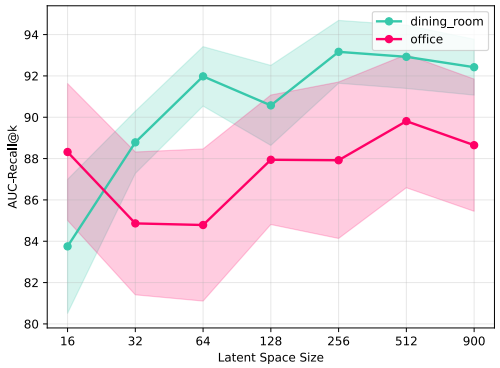


Figure 8. Ablation results with the AUC-Recall@k (↑) for different latent space dimensions of the autoencoder.

periments substituting GloVe [7] with stronger sentence embedding models, which encode complete sentences rather than individual words. For these models, each triplet is encoded as a full sentence of the form: `'In a {scene_type}, {obj1} {relation} {obj2}'`, incorporating both the scene context and the triplet components. The results are provided in Tab. 5. Despite being a simpler word-level model, GloVe achieves competitive performance compared to the more computationally expensive sentence embedding models, justifying its use in *BUSSARD*.

10. Additional Design Study Results

Table 6 and Fig. 8 show the office scene and AUC-Recall@k counterparts to the design study results presented in Sec. 5.4.2.

11. Qualitative Analysis

To better understand the strengths and weaknesses of *BUSSARD*, we conduct a qualitative analysis across both scenes. True positives include unusual object placements such as 'banana-on-chair' or 'shoe-on-desk'. The false negative 'pillow-on-table' scored near the decision threshold, indicating a genuinely ambiguous relationship, likely due to rare training occurrences. A notable false positive is the triplet 'flower-under-clock' which is detected by the scene graph generator despite not being in the image. However the detection of this triplet as anomalous is correct.

Table 6. Ablation study results for **office** 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	94.18 ± 2.45	87.35 ± 5.23
Feature Mult	128	94.65 ± 3.18	88.24 ± 6.74
Node Only	512	90.89 ± 4.33	80.16 ± 9.26
No AE	900	85.74 ± 2.98	71.82 ± 5.18
BUSSARD (ours)	512	**95.29** ± 1.65	**89.57** ± 3.46

Figure 12 shows subsets of the generated scene graphs with normalized triplet anomaly scores, with the two highest-scoring triplets highlighted. In Fig. 12a, 'flower-on-table' correctly receives the highest anomaly score. In Fig. 12b, however, 'bag-on-chair' scores highest while the true anomaly 'shoe-on-chair' ranks second, suggesting that bags on chairs were rarely seen during training, causing an elevated anomaly score.

12. Example Scene Graphs of Images

We present example images from the dataset with their corresponding top 30 triplets. Three examples each from the dining room and office scenes can be found in Fig. 9. The corresponding scene graphs for the dining room (see Fig. 10) and office (see Fig. 11) examples are also presented.

(a) Dining room example 1

(d) Office example 1

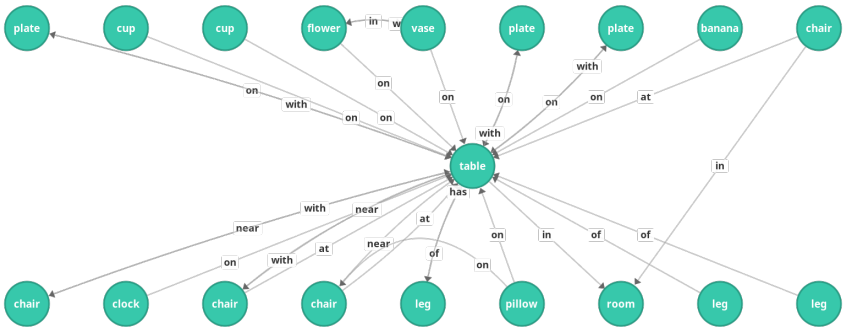
(b) Dining room example 2

(e) Office example 2

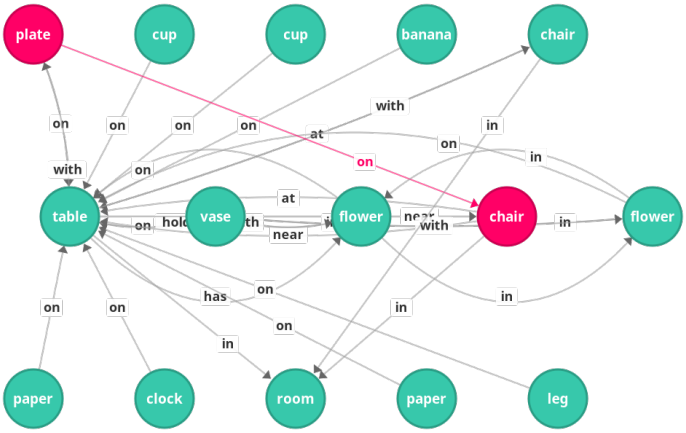
(c) Dining room example 3

(f) Office example 3

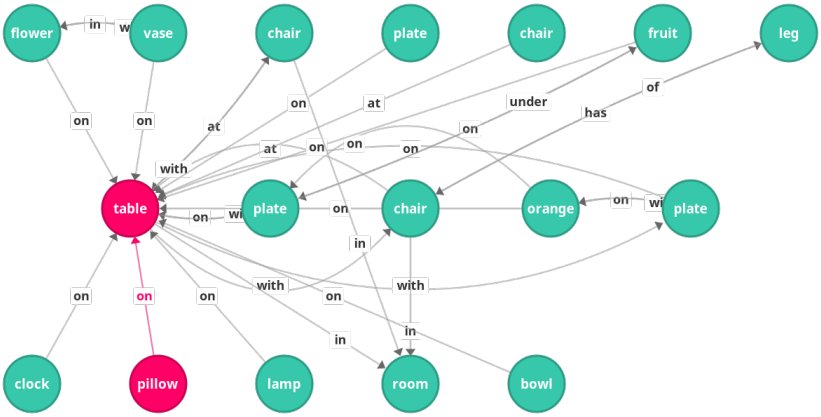
Figure 9. Example images from **dining room** (left) and **office** (right). The top images (a) and (d) are normal, while the others contain anomalies.



(a) Scene graph dining room example 1

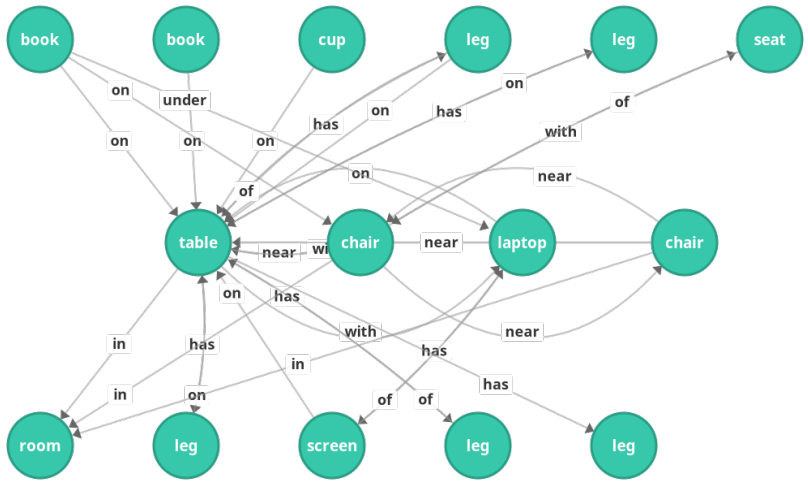


(b) Scene graph dining room example 2

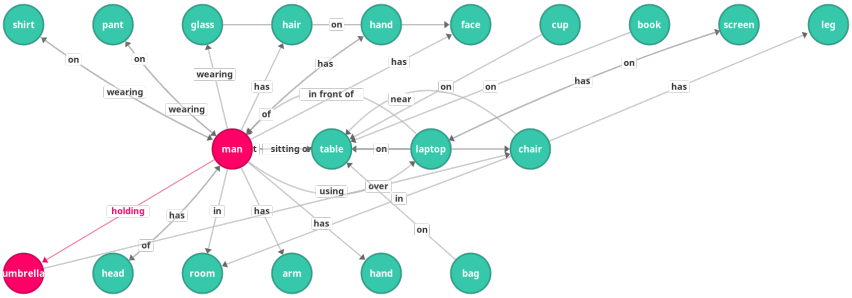


(c) Scene graph dining room example 3

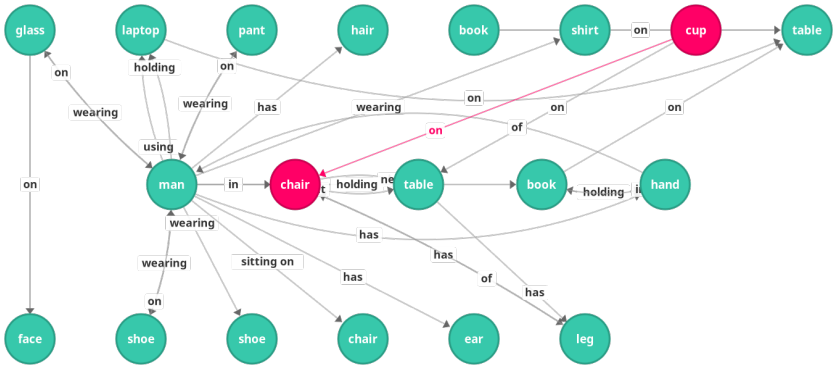
Figure 10. Example scene graphs from **dining room** images, constructed from the top 30 triplets. The last two scene graphs contain anomalies, which are highlighted red.



(d) Scene graph office example 1

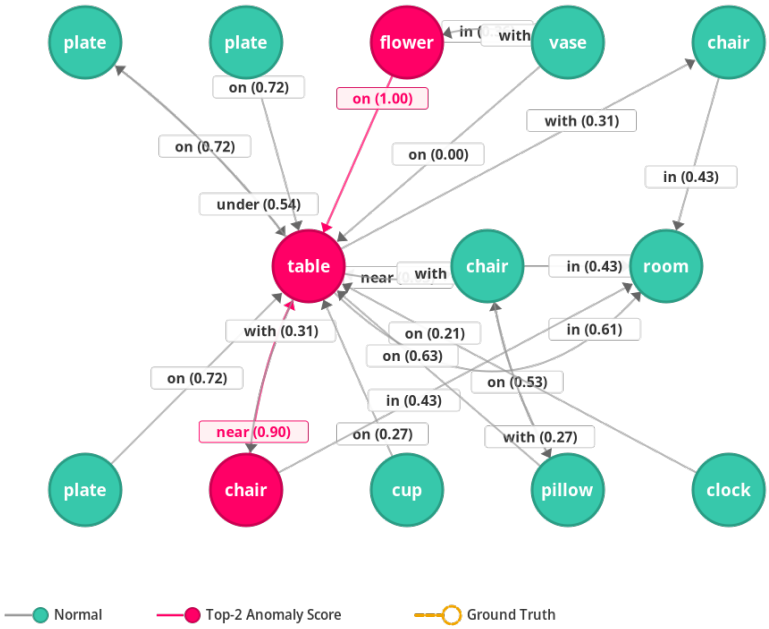


(e) Scene graph office example 2

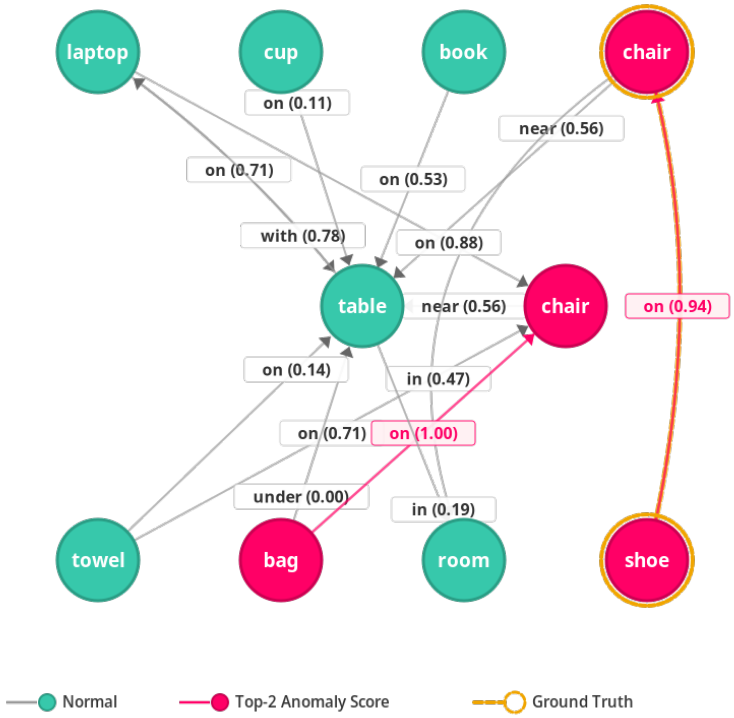


(f) Scene graph office example 3

Figure 11. Example scene graphs from **office** images, constructed from the top 30 triplets. The last two scene graphs contain anomalies, which are highlighted red.



(a) Scene graph dining room example with normalized anomaly scores



(b) Scene graph office example with normalized anomaly scores

Figure 12. Example subsets of scene graphs with normalized anomaly scores. The red highlighted elements belong to the two triplets with the highest anomaly score and the ground truth is marked with an additional, yellow circle.