BUSSARD

Normalizing Flows for Bijective Universal
Scene-Specific Anomalous Relationship Detection

Melissa Schween · Mathis Kruse · Bodo Rosenhahn

Institute of Information Processing (tnt), L3S – Leibniz University Hannover

CVPR JUNE 3-7, 2026 DENVER COLORADO

The Problem: What Counts as an Anomaly?

Anomalies are context-dependent.

Feet on a table in a dining room are unusual, while feet on a coffee table in the living room might be normal.

Pixel-level methods miss relationships.

Standard image-based anomaly detection treats scenes as textures and overlooks how objects interact.

Scene graphs offer structure.

Decomposing an image into (object – relation – object) triplets makes context explicit and learnable.

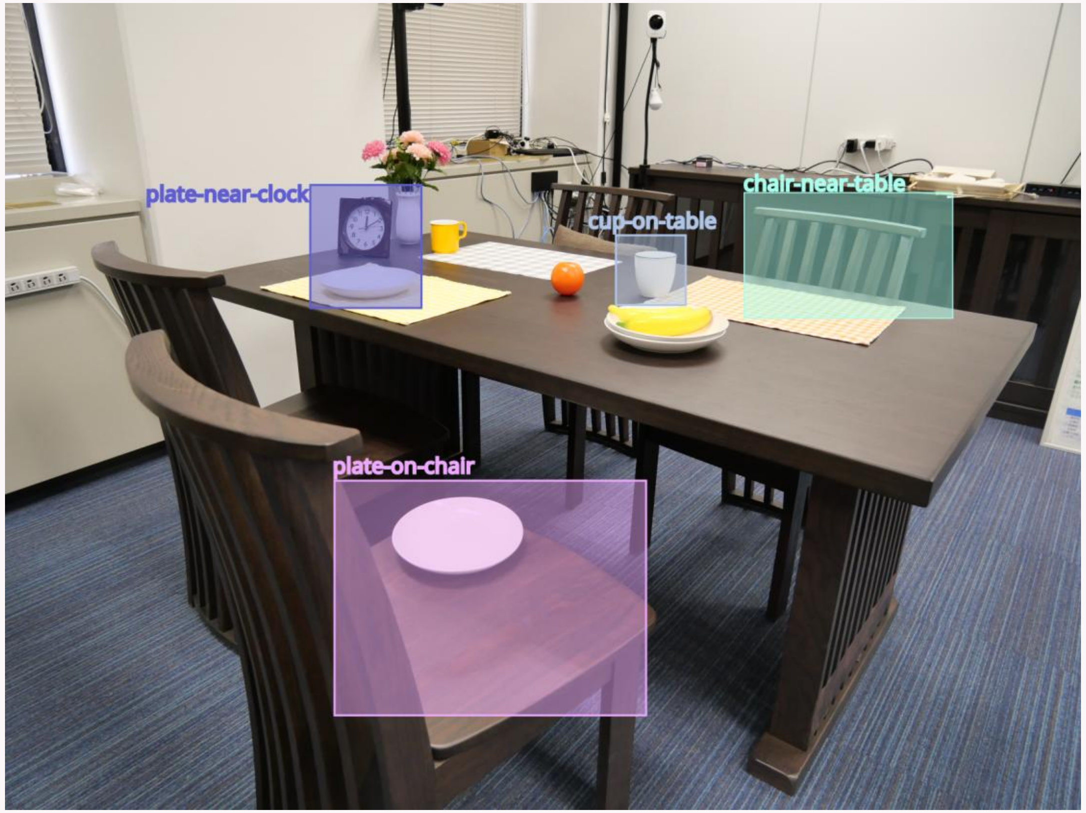


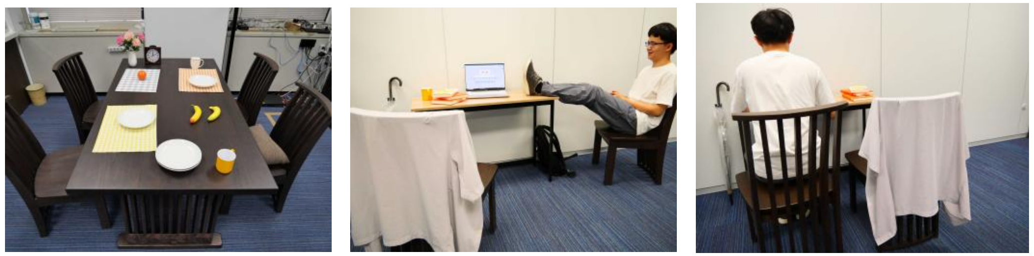
Figure 1: An example dining room scene with an anomalous relationship (plate-on-chair).

Datasets: SARD and MIT-67

Two datasets covering the same two scene types: dining room and office.

SARD

Lai et al., "Scene-Specific Anomalous Relationship Detection Using Scene Graph Summarization", CVPRW 2025.



60 normal + **~60 anomalous** images per scene

Controlled lab setup with annotated anomalous relationships

MIT-67

Quattoni & Torralba, "Recognizing Indoor Scenes", CVPR 2009.

274 dining-room + **109 office** images

No annotated anomalies, only used as normal images

Why both? SARD provides clean training and evaluation; MIT-67 tests generalization to in-the-wild scene variation.

Method: The BUSSARD Pipeline

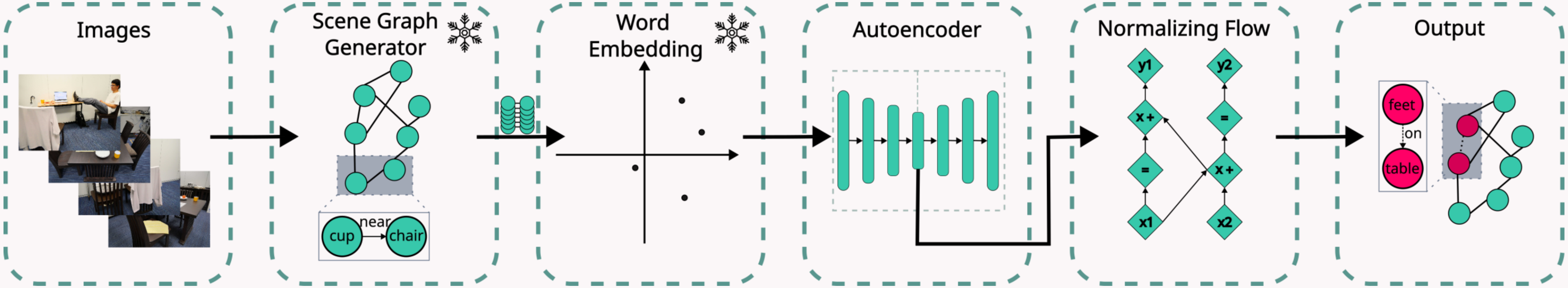


Figure 2: Four-stage pipeline of BUSSARD, from images to predicted anomaly likelihoods.

Scene Graph

EGTR [1] (pretrained) → top-30 triplets per image

Embedding

GloVe [2] word embeddings on object + relation tokens

Autoencoder

Compresses triplet embeddings to $d = 512$

Normalizing Flow

Three RealNVP [3] coupling layers
Negative log-likelihood as anomaly score

1. Im et al., "EGTR: Extracting Graph from Transformer for Scene Graph Generation", CVPR 2024.
2. Carlson et al., "A New Pair of GloVes", arXiv 2025.
3. Dinh et al., "Density Estimation using Real NVP", ICLR 2017.

Results: Outperforming the Counting Baseline

BUSSARD beats SARD-c by ~10% AUROC on both scenes and is more robust.

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**

Table 1: Results of BUSSARD and baseline. SARD-o are original results and SARD-c with corrected preprocessing.

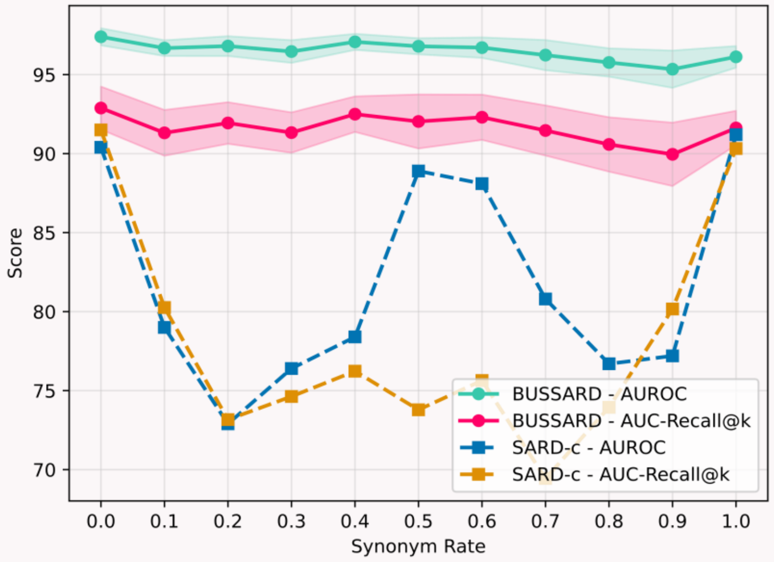


Figure 3: Ablation study results for different synonym rates. BUSSARD stays flat across substitution rates, SARD-c collapses.

AUC-Recall@k:
+4% over SARD-c, +24% over SARD-o

5× faster inference on a single RTX 3090

Synonym deviation:
~2% (BUSSARD) vs. ~17.5% (SARD-c)

Ablation Studies

Autoencoder is critical
Removing it causes the largest performance drop.

Relationships > objects alone
The object-only variant performs noticeably worse.

GloVe is enough
Larger sentence-embedding models give no clear advantage.

Cross-dataset generalization
Training on MIT-67 works well for dining room scenes, but poorly for office, likely due to distribution shift.

Cross-dataset generalization

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 2: Multi-scene training on MIT-67 and SARD. Evaluated on SARD.

Why offices fail to transfer

MIT-67 dining rooms look very different from SARD's controlled lab setup, yet BUSSARD still generalizes there. Offices vary even more across MIT-67, breaking the learned distribution of normal relationships.

Conclusions

- **First learning-based approach** for scene-specific anomalous relationship detection.
- **Semantic embeddings + normalizing flows** effectively capture the distribution of normal relationships.
- **Robust to vocabulary variation** well-suited for open-world settings with synonym noise.
- **Next steps** extend to video via dynamic scene graphs or integrate LLMs for textual anomaly feedback.

Link to GitHub.

Code: github.com/mschween/BUSSARD
Contact: schween@tnt.uni-hannover.de · tnt.uni-hannover.de/staff/schweenm