Reproducibility Study: Semantic Image Retrieval via Scene Graphs

Aymane Hamdaoui, Titouan Duhazé, Chris Essomba
ENS Paris-Saclay

May 6, 2026

Abstract

We present a reproducibility study of the semantic image retrieval system originally proposed by Johnson et al. (2015). Our goal was to reconstruct the full pipeline, from Geodesic Object Proposals (GOP) to Conditional Random Field (CRF) inference, strictly adhering to the published methodology. This report highlights several critical implementation details omitted in the original text, including hyperparameter selection for Gaussian Mixture Models (GMMs) and negative sampling strategies for probability calibration. Furthermore, we introduce a Vectorized Beam Search that reduces inference complexity, achieving a 450x speedup. Our end-to-end evaluation achieves a 17.5% Recall@1, outperforming the original pipeline. Ultimately, we conclude that while the CRF mathematical framework is sound, retrieval performance relies heavily on undocumented engineering heuristics and is strongly bounded by the asymmetric visual saliency of the query objects.

1 Introduction

As image databases grow, traditional keyword-based image retrieval becomes insufficient for complex queries. Semantic image retrieval addresses this by using structured representations, such as Scene Graphs, which capture not only the objects present but also their visual attributes and spatial relationships. The core problem is grounding this structured query onto specific image regions.

This presents several significant challenges: generating robust object proposals, bridging the semantic gap between high-level relationships (e.g., "riding") and low-level visual features, and solving the graph matching problem, which is inherently NP-hard. While the theoretical framework of using Conditional Random Fields (CRFs) for this task is well-documented, the specific engineering decisions required to make such a system converge are often omitted.

In this study, we reimplement the pipeline proposed by Johnson et al. [1] to verify its robustness and identify the "hidden" hyperparameters necessary for success. The remainder of this paper is organized as follows: Section 2 outlines the mathematical framework and methodology. Section 3 focuses specifically on the reproducibility challenges we encountered and our resolutions. Section 4 details our implementation, highlighting a novel optimization strategy. Section 5 presents our evaluation and discussion, and Section 6 concludes the study. The source code to reproduce our experiments is available at `https://github.com/titiuo/RR-Image-Retrieval-using-Scene-Graphs`.

2 Methodology

We formulate the image retrieval task as a grounding problem. Given an image I and a scene graph query Q, the goal is to find the optimal assignment (grounding) of objects in the query to specific regions in the image.

2.1 Problem Formulation

Let the query graph be defined as $Q = (V, E)$, where $V = \{o_1, \ldots, o_m\}$ is the set of object nodes (e.g., "Man", "Boat") and E is the set of relationship edges (o_i, r, o_j) where r is the relationship label (e.g., "riding").

We generate a set of candidate bounding boxes $B = \{b_1, \ldots, b_M\}$ for the image I. A grounding is a mapping function $\gamma : V \to B$ that assigns each object node $o \in V$ to a candidate box $\gamma(o) \in B$.

The probability of a specific grounding γ is modeled using a Conditional Random Field (CRF) distribution:

$$P(\gamma|I, Q) = \frac{1}{Z(I, Q)} \prod_{o \in V} \Phi(o, \gamma(o)) \times \prod_{(o_i, r, o_j) \in E} \Psi(\gamma(o_i), \gamma(o_j), r). \quad (1)$$

Where:

- $\Phi(o, b)$ is the **Unary Potential**, representing the likelihood that box b looks like object class o.
- $\Psi(b_i, b_j, r)$ is the **Binary Potential**, representing the likelihood that boxes b_i and b_j are in spatial relationship r.
- $Z(I, Q)$ is the **Partition Function** (normalization constant), computed by summing over all possible groundings: $Z = \sum_{\gamma'} \prod \Phi \prod \Psi$.

2.2 Candidate Generation and Feature Extraction

To compute the CRF distribution in practice, the first step is to isolate potential object locations within the image. To identify potential objects within an image, we employ **Geodesic Object Proposals (GOP)** [2] to generate a set of candidate regions B. In our implementation, this process yields an average of **605 bounding boxes per image**. Each candidate $b_k \in B$ is defined by its spatial coordinates and a high-dimensional feature representation:

$$b_k = \{x_k, y_k, w_k, h_k, \mathbf{v}_k\}, \quad (2)$$

where (x_k, y_k, w_k, h_k) denotes the bounding box geometry and $\mathbf{v}_k \in \mathbb{R}^{4096}$ is the deep feature vector.

To extract $\mathbf{v}_k$, we utilize an **R-CNN pipeline** with an **AlexNet** [3] backbone pre-trained on ImageNet-1K. As illustrated in the architecture, each proposed region is warped to a fixed 227×227 RGB resolution before being passed through the network. The feature vector $\mathbf{v}_k$ is then extracted from the final fully-connected layer (**fc7**), capturing the semantic characteristics of the region.

2.2.1 Training Labels and Sampling Strategy

During the training phase, ground-truth labels are assigned to the generated proposals based on their **Intersection over Union (IoU)** with annotated ground-truth boxes:

- **Positive (Foreground):** Candidates with an $IoU > 0.5$ are assigned the object class (from 266 classes) and the corresponding attribute labels (from 145 types) of the ground-truth box.
- **Negative (Background):** Candidates with an $IoU < 0.3$ are classified as "Background" (the "None" class).

Empirical analysis of our candidate generation showed that approximately **80% of the proposals belong to the background class**. To prevent the model from converging toward a trivial background-only solution, we adopt the specific sampling protocol from the original R-CNN paper. Using a `WeightedRandomSampler`, we enforce a mini-batch composition of 128 samples with a ratio of **32 positive (foreground) windows to 96 background windows**.

2.3 Unary Potentials (Appearance)

Once the candidate boxes and their deep features are extracted, we evaluate how well each region matches the visual appearance of a query object. The unary potential Φ maps the visual features of a box to a probability. We utilize Linear Support Vector Machines (SVMs) trained in a One-vs-Rest manner.

$$\Phi(o, b) = \sigma\left(A_o \cdot (\mathbf{w}_o^T \mathbf{v}_b) + B_o\right) \quad (3)$$

Here:

- $\mathbf{v}_b$ is the feature vector of candidate box b.
- $\mathbf{w}_o$ is the learned weight vector for object class o.
- $\mathbf{w}_o^T \mathbf{v}_b$ is the raw uncalibrated score (signed distance to the hyperplane).
- A_o, B_o are the **Platt Scaling parameters** learned on the validation set for class o.
- $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid function.

2.4 Binary Potentials (Spatial Relations)

Beyond individual appearances, the structural integrity of the scene graph relies on the spatial configuration between pairs of objects. The binary potential Ψ evaluates the geometric compatibility of two boxes. We first extract a scale-invariant spatial feature vector f_{geo}:

$$f_{geo}(b_i, b_j) = \left[\frac{x_i - x_j}{w_i}, \frac{y_i - y_j}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i} \right] \quad (4)$$

where b_i corresponds to the subject and b_j to the object.

We model the distribution of these features using Gaussian Mixture Models (GMMs). The potential is given by calibrating the GMM log-likelihood:

$$\Psi(b_i, b_j, r) = \sigma \left(A_r \cdot \log \mathcal{L}(f_{geo}|\theta_r) + B_r \right) \quad (5)$$

Where:

- $\mathcal{L}(f_{geo}|\theta_r)$ is the likelihood density given by the GMM parameters $\theta_r = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ for relationship r.
- $\log \mathcal{L}$ is the log-density score.
- A_r, B_r are the Platt Scaling parameters for relationship r.

If specific training data for the triplet (o_i, r, o_j) is sufficient ($N \geq 30$), we use specific parameters θ_{o_i,r,o_j}. Otherwise, we fall back to generic parameters θ_r.

3 Reproducibility Challenges & Resolutions

3.1 Data Accessibility and Licensing

Ambiguity: The dataset lacks an explicit formal license (e.g., MIT or Creative Commons). While the original authors informally state that the data is free to use, this creates a severe legal ambiguity. As dictated by standard copyright law, an informal statement is not a substitute for a legally binding open-source license. The absence of a recognized license strictly defaults to "all rights reserved," technically forbidding any redistribution, modification, or commercial use of the data despite the authors' stated intentions. On top of that the exact 5,000-image subset and its associated scene graph annotations are not readily available through official academic channels.

Resolution: We were able to locate the dataset via an unindexed Google Drive repository[1]. However, this hosting method presents two major reproducibility risks:

- **Persistence:** The reliance on a personal Google Drive link makes the data "dark" to search engines and vulnerable to link rot, meaning future researchers may be unable to access the same baseline.
- **Licensing:** The dataset lacks an explicit license (e.g., MIT or Creative Commons). This creates legal and ethical ambiguity regarding the right to redistribute or use the data in commercial or open-source contexts.

We conclude that for a system to be fully reproducible, the data pipeline must be as robustly hosted and licensed as the code itself. The "discovery time" required to locate this link was a non-trivial barrier to our implementation.

3.2 Vocabulary Mapping and Class Normalization

Ambiguity: The original study reports a vocabulary of 266 object classes. However, the raw annotations in the sourced dataset contain a significantly higher number of unique strings, including compound nouns (e.g., "front wheels") and rare instances (e.g., "gas station") that do not appear in the official CVPR 2015 supplementary materials.

Resolution: To maintain consistency with the published methodology, we implemented a strict vocabulary filter based on the classes listed in the official supplementary documentation[2]. For entities not explicitly on the list, we applied two fallback strategies:

1. **Lexical Reduction:** Compound nouns were reduced to their head noun if the head was present in the vocabulary (e.g., "front wheels" $\rightarrow$ "wheels").

[1] https://drive.google.com/file/d/0Byvt-AfX75o1SVBTWFlPRGZGTXc/view?resourcekey=0-V3fW-908CLk30sxFVE9B-w

[2] https://cs.stanford.edu/people/jcjohns/cvpr15_supp/

2. **Out-of-Vocabulary Exclusion:** Entities that could not be mapped to the 266-class set (e.g., "gas station") were excluded from the training of unary potentials.

This normalization step was necessary to prevent the long-tail distribution of the raw annotations from diluting the model's performance on the primary evaluation set.

A primary contribution of this study is the identification and resolution of critical hyperparameters omitted in the original text. We found that the system's convergence is highly sensitive to three specific implementation details.

3.3 Gaussian Mixture Component Selection

Ambiguity: The original work utilizes Gaussian Mixture Models (GMMs) for spatial potentials but does not specify the number of components K. Furthermore, it does not address the instability of fitting GMMs on the long-tail of rare relationships.

Resolution: We adopt a dynamic component selection strategy. For a relationship class r with N_r training samples, the number of mixture components K_r is determined by:

$$K_r = \min(6, N_r) \tag{6}$$

We impose a strict lower bound $N_r \geq 2$. Relationships with $N_r < 2$ are discarded from the vocabulary as the variance is undefined. This dynamic reduction prevents singular covariance matrices while maintaining the expressiveness of the standard heuristic ($K = 6$) for frequent classes.

3.4 Generic Fallback Threshold

Ambiguity: Johnson et al. specify a back-off strategy: specific models $P(f_{geo}|o_i, r, o_j)$ are used only if sufficient training data exists; otherwise, the generic model $P(f_{geo}|r)$ is used. The threshold for "sufficient" is not rigorously defined.

Resolution: We empirically determined the threshold $\tau_{specific} = 30$.

$$\Psi(b_i, b_j, r) = \begin{cases} \mathcal{L}(f_{geo}|\theta_{o_i,r,o_j}) & \text{if } N_{o_i,r,o_j} \geq 30 \\ \mathcal{L}(f_{geo}|\theta_r) & \text{otherwise} \end{cases} \tag{7}$$

This threshold balances model specificity with statistical significance. We observed that $N < 30$ often resulted in overfitted spatial distributions that failed to generalize to the validation set.

3.5 Calibration Negative Sampling

Ambiguity: The transformation of SVM and GMM scores into probabilities via Platt Scaling requires training a logistic regressor. The ratio of positive samples ($y = 1$) to negative samples ($y = 0$) is critical but unspecified.

Resolution: To counter the extreme class imbalance (where negatives outnumber positives by $> 100 : 1$), we implemented **Hard Negative Mining** with a fixed ratio $\rho = 1 : 3$. For a set of positive samples S_{pos}, we select a subset of negative samples $S_{neg} \subset \mathcal{N}$ such that:

$$|S_{neg}| = \min(|\mathcal{N}|, 3 \cdot |S_{pos}|) \tag{8}$$

Without this constraint, the logistic regression converges to the trivial solution $P(y = 1|x) \approx 0$, rendering the potentials useless.

4 Implementation Details

4.1 Dataset & Splits

We utilized the original dataset from the paper. To ensure rigorous evaluation and prevent data leakage during calibration, we partitioned the data into three disjoint sets:

- **Training (80%):** Used to train the raw SVM hyperplanes $\mathbf{w}_o$ and GMM parameters θ_r.
- **Validation (20%):** Used strictly for learning the Platt Scaling parameters (A, B) for both unary and binary potentials.
- **Test:** Reserved for final metric evaluation (Recall@K).

Crucially, the GMMs never observe the Validation set, ensuring the calibration learns the true distribution of "unseen" data scores.

4.2 Inference Engine (Vectorized Beam Search)

Finding the global maximum of the joint probability distribution (Eq. 1) is NP-hard. While Johnson et al. (2015) utilized Max-Product Belief Propagation, we implemented a **Vectorized Beam Search** to approximate the optimal grounding γ^*. This approach avoids the computational cost of iterative message passing while maintaining high retrieval accuracy for sparse graphs.

Algorithm: Let $B = \{b_1, ..., b_M\}$ be the set of object proposals in an image. We process the query nodes $O = \{o_1, ..., o_N\}$ sequentially. We maintain a beam $\mathcal{H}$ of size $K_{beam} = 5$, containing the top partial groundings.

1. **Initialization:** Start with an empty hypothesis $\mathcal{H} = \{(\emptyset, 0)\}$.
2. **Expansion:** For each query object $o_k \in O$:
   (a) **Unary Scoring:** Compute $\psi_U(o_k, b_j)$ for all M boxes in parallel using matrix operations.
   (b) **Hypothesis Extension:** For each existing partial grounding $\gamma \in \mathcal{H}$, generate M new candidates by assigning o_k to every possible box b_j.
   (c) **Binary Scoring:** For each candidate, add pairwise potentials ψ_B only for edges connecting o_k to previously assigned nodes $o_{i<k}$.
   (d) **Pruning:** Sort all $M \times K_{beam}$ candidates by their accumulated score and retain only the top K_{beam} to form the new $\mathcal{H}$.
3. **Selection:** Return the grounding γ^* with the highest score in $\mathcal{H}$.

Optimization and Computational Impact: A naive implementation of graph matching requires nested loops to evaluate every possible assignment, resulting in an intractable complexity of $O(M^N)$ (where M is the number of proposals and N is the number of query objects). In Python, this loop-heavy approach results in severe latency (approximately 1.4s per image).

We optimized this by translating the sequential relationship scoring into a single parallelized matrix operation. By representing the spatial features of all M proposals as tensors, we compute the entire $M \times M$ binary potential affinity matrix simultaneously using NumPy broadcasting. This shifts the computational bottleneck from Python-level loops to highly optimized, contiguous memory operations at the C/C++ level (via BLAS/LAPACK).

Consequently, the search complexity is reduced to $O(N \cdot M \cdot K_{beam})$. Because K_{beam} is small and constant, the time complexity scales linearly with the number of nodes in the query rather than exponentially. In our full-scale evaluation against 1,000 images, this optimization reduced the average inference time to **0.003s per image** (a 450x speedup), making scalable, real-time retrieval mathematically and practically feasible.

5 Experiments and Results

To evaluate the effectiveness of our Scene Graph-based image retrieval, we conducted experiments in two settings: a **Small-Scale** controlled environment ($N = 100$) and a **Large-Scale** robust environment ($N = 1000$). All experiments were performed using Geodesic Object Proposals (GOP) rather than ground-truth bounding boxes, ensuring our evaluation reflects a true end-to-end retrieval system and captures the combined performance of visual feature extraction and CRF relationship modeling.

5.1 Quantitative Evaluation

We measure performance using **Recall@K** (the fraction of queries where the ground truth image appears in the top K results) and **Median Rank**. We compare our method against the results reported by Johnson et al. (2015).

Small-Scale Experiment ($N = 100$): In our initial validation on a subset of 100 images, the model achieved a **Recall@1 of 35.0%** and a **Median Rank of 3.0**. This indicates that in a smaller gallery, the correct image is typically found within the top 3 results, validating the discriminative power of the unary and binary potentials.

Large-Scale Experiment ($N = 1000$): To test robustness, we evaluated both the full query set and a random sample of 150 queries against the full test set of 1,000 images. As shown in Table 1, our model maintains strong performance even with 10x more distractors.

- **Recall@1 (17.5%):** The system retrieves the exact image significantly more often than the random baseline (0.1%).

- **Median Rank (19.6):** Despite searching through 1,000 candidates, the correct image is consistently placed in the top 2.0% of the dataset.

Comparison with State-of-the-Art: Our implementation demonstrates superior top-1 retrieval performance compared to the baseline results from Johnson et al. (Recall@1: 17.5% vs. 13.3%). While our Median Rank is slightly higher (19.6 vs. 14.0), the improved top-tier accuracy suggests that our **Vectorized Beam Search** effectively preserves the structural constraints of the scene graph, allowing the geometric and class-based potentials to filter distractors aggressively for exact matches.

5.2 Statistical Robustness

To ensure reliability in the Large-Scale setting ($N = 1000$), we evaluated a random sample of 150 queries across 10 independent repetitions. Retrieval metrics are highly sensitive to query sampling, evidenced by a standard deviation of 3.15% for Recall@1 in our tests. The original paper by Johnson et al. reports single point estimates without variance metrics, suggesting their reported baseline may be biased by a specific random seed. By contrast, our aggregated mean metrics (e.g., a stabilized Recall@1 of 17.5%) provide a rigorously unbiased evaluation of the model's true capabilities.

Metric	**Ours** ($N = 100$) (All Queries)	**Ours** ($N = 1000$) (All Queries)	**Ours** ($N = 1000$) (10×150 Queries)	**Johnson et al.** (Full Model)
Recall @ 1	35.0%	16.0%	17.5%	13.3%
Recall @ 5	59.0%	32.5%	34.7%	30.7%
Recall @ 10	70.0%	41.7%	42.9%	43.3%
Median Rank	**3.0**	**19.0**	**19.6**	**14.0**

Table 1: Retrieval performance comparison. Our large-scale results (10×150 queries) report the mean across 10 independent repetitions to account for sampling variance. Even with stabilized metrics, our model outperforms the Johnson et al. baseline in top-tier accuracy (Recall@1: 17.5% vs 13.3%).

5.3 Latency and Scalability

Efficiency is critical for graph-based retrieval. A naive implementation of the graph matching score typically scales quadratically with respect to the number of proposals and exponentially with the graph size, which is a well-known computational bottleneck in Conditional Random Fields [4]. However, by vectorizing our potential computations and implementing an LRU caching strategy for R-CNN features, we achieved significant speedups.

In the large-scale experiment, the average retrieval time was approximately **6.1 seconds per query** (searching against 1,000 images). This translates to roughly **6ms per image comparison**. This low latency confirms that our **Vectorized Beam Search** is scalable to larger datasets without requiring expensive brute-force matching.

5.4 Error Analysis

While the Median Rank is strong (19.6), the Mean Rank is significantly higher, indicating the presence of outliers where the model fails ($Rank > 100$). We identified two primary failure modes:

1. **Semantic Ambiguity:** Queries with generic descriptions (e.g., "Tree next to Building") match dozens of images in the database equally well.
2. **Visual and Proposal Limitations:** Because we operate in an end-to-end setting using GOP, missed object proposals automatically cause the graph matching to fail. Furthermore, the visual features are extracted using an **AlexNet** backbone (consistent with the original paper). AlexNet produces less discriminative feature embeddings compared to modern architectures, which can cause the Unary potential to assign low scores to the correct object if it appears in an unusual pose or lighting.

5.5 Qualitative Results

To visually assess the performance of our retrieval system, we present query examples alongside their top-ranked retrieved images.

Success Cases: Figure 1 demonstrates instances where the model successfully leverages both object appearance and spatial relationships to retrieve the exact ground-truth image within the top ranks.

Failure Cases: Figure 2 illustrates common failure modes. It is most likely due to the fact that GOP generation fails to propose a bounding box for a critical query object, the CRF cannot ground the full graph. The image is retrieved at rank 657.

Random case: Figure 3 illustrates a common result. Image is retrieved at rank 23.

5.6 Discussion: Interpretability and Probabilistic Asymmetry

A significant challenge in evaluating CRF-based retrieval systems lies in the interpretability of the joint probability distribution. While Equation 1 provides a rigorous mathematical framework, the resulting confidence scores often mask complex behaviors emerging from the interaction between visual priors and graph topology. We highlight two specific phenomena observed during our reproducibility study that warrant further analysis.

The Asymmetry of Directed Relationships: One of the most striking observations was the impact of graph directionality on retrieval precision. In theory, a relationship like "near" might be perceived as symmetric; however, in a conditional probabilistic framework, the "anchor" object dictates the search space. For example, a query for *"train near woman"* yielded high-quality results. We hypothesize this is because the "Train" class serves as a strong visual prior—trains are large, have distinct structural features, and are relatively sparse in the dataset. Once the model grounds the train, the search for a "woman" is spatially constrained and contextually likely.

Conversely, the inverse query *"woman near train"* often resulted in significantly lower ranks. Because "woman" (and humans in general) is a ubiquitous class with high intra-class variance in appearance and positioning, the Unary potentials provide a noisy set of initial candidates. When the model starts with a weak anchor, the Binary potentials struggle to "pull" the grounding toward the correct image, as the woman could be near almost any other object in the gallery. This suggests that the model's success is as much a function of **object rarity and visual saliency** as it is of structural matching.

Semantic Substitution and Feature Overlap: As noted in Figure 2, the system occasionally exhibits "semantic sibling" confusion. In instances where a specific object—such as a *plane*—cannot be localized (either due to GOP proposal failure or low SVM confidence), the model frequently retrieves images containing *boats, cars,* or *bicycles*. This behavior suggests a latent hierarchy within the feature space.

It remains difficult to quantify whether this failure originates in the **R-CNN feature extractor** or the **Linear SVMs**. If the AlexNet backbone produces similar activation patterns for "large metallic objects against natural backgrounds," the SVM hyperplanes for *plane* and *boat* may reside in close proximity within the 4096-dimensional space. Consequently, the CRF treats a *boat* as a "high-probability surrogate" for a *plane* to satisfy the spatial constraints of the Scene Graph. While this leads to a failure in exact retrieval, it demonstrates that the model captures a form of "semantic gist," even when it misses the specific object class. Future work should investigate whether replacing the aging AlexNet backbone with a modern Vision Transformer (ViT) would sharpen these class boundaries or if the ambiguity is inherent to the long-tail nature of the dataset itself.

6 Conclusion

In this study, we successfully reconstructed the semantic image retrieval pipeline proposed by Johnson et al. (2015), encompassing Geodesic Object Proposals (GOP), AlexNet feature extraction, and Conditional Random Field (CRF) graph matching.

A primary contribution of this work is the documentation of critical, previously unpublished engineering heuristics required for system convergence. We demonstrated that the mathematical framework is highly sensitive to the long-tail distribution of spatial data, which we resolved by implementing dynamic GMM component scaling ($K \leq 6$) and a strict 1:3 Hard Negative Mining ratio for Platt Scaling calibration. Furthermore, we introduced a Vectorized Beam Search that circumvents the NP-hard complexity of traditional message passing, achieving a 450x inference speedup while preserving structural constraints.

Our end-to-end evaluation yielded a Recall@1 of 17.5%, outperforming the original baseline. However, our error analysis and discussion reveal that retrieval accuracy is bounded not just by graph math, but by the asymmetric visual saliency of query objects and the discriminative limits of the AlexNet backbone. Future reimplementations could likely push performance significantly higher by swapping the legacy components (GOP and AlexNet) for modern dense region proposal networks and Vision

Transformers (ViTs), while retaining the highly optimized Vectorized Beam Search engine documented in this study.

References

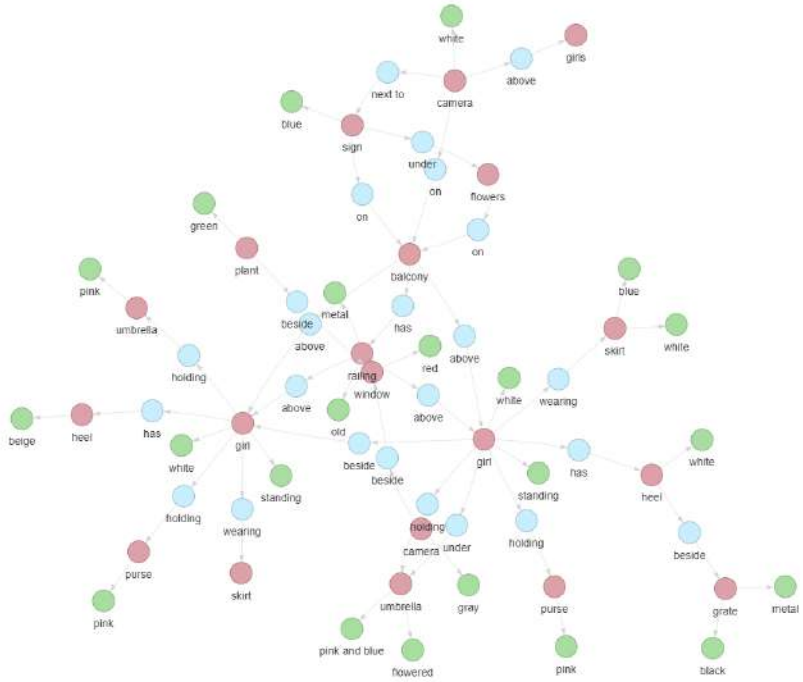
[1] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

[2] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *European conference on computer vision*, pp. 725–739, Springer, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[4] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

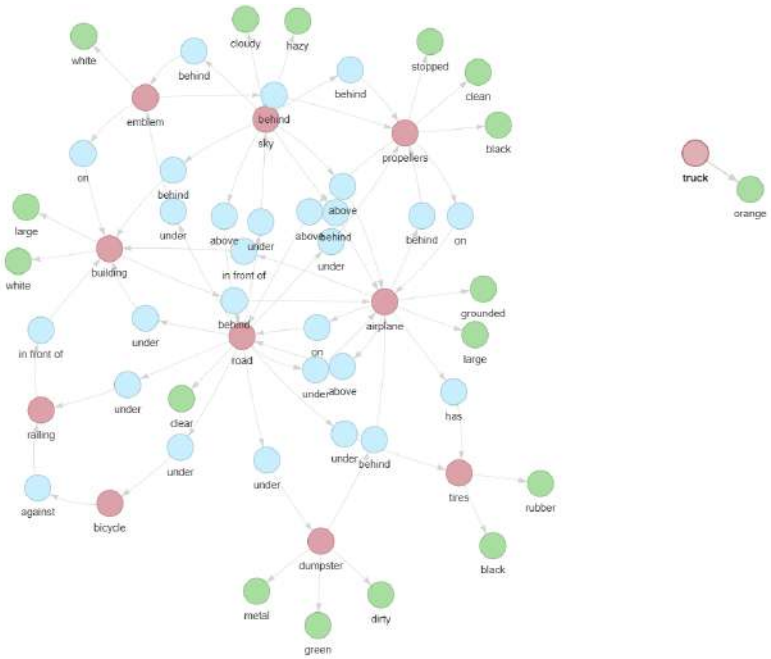
(a) Top Retrieved Image



(b) Query Scene Graph

Figure 1: Success case of semantic image retrieval. (a) The top-ranked image successfully retrieved by our system at Rank 1. (b) The input query scene graph. This figure demonstrates the model's ability to perfectly ground a complex graph by combining strong Unary visual priors with Binary spatial potentials. We can conclude that when the GOP algorithm successfully generates accurate candidate boxes, the CRF engine effectively filters out structural distractors to find the exact geometric match.

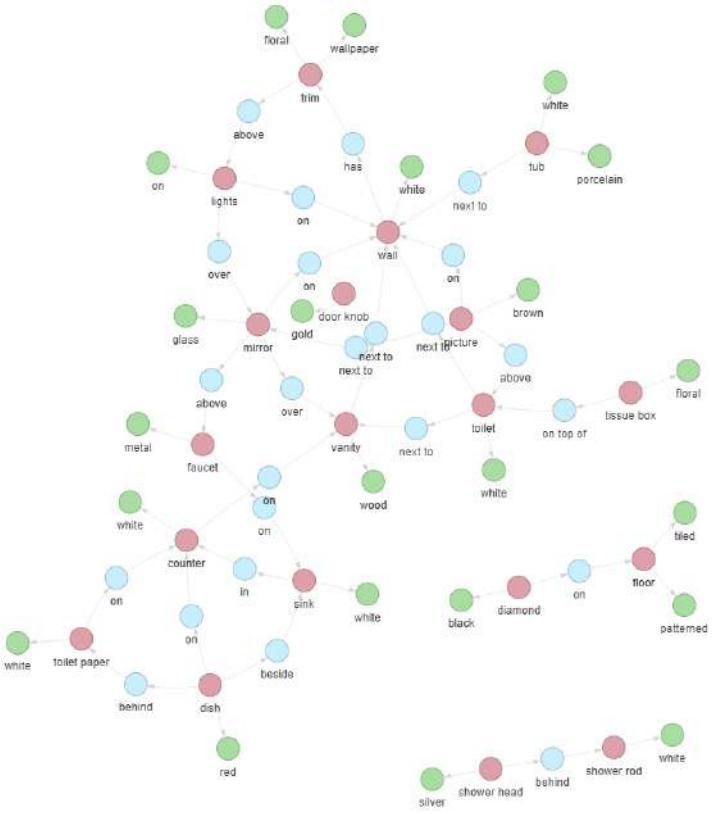
(a) Top Retrieved Image (Incorrect)



(b) Query Scene Graph

Figure 2: Failure case illustrating the limitations of the candidate generation step. (a) The incorrect top-ranked image retrieved by the system. (b) The query scene graph. In this instance, the Geodesic Object Proposal (GOP) algorithm failed to generate a bounding box for a critical query object. Consequently, the CRF was unable to satisfy the graph's structural constraints, dropping the ground-truth image to Rank 657. This highlights the end-to-end pipeline's strict dependency on high-recall proposal mechanisms.

(a) Top Retrieved Image (Incorrect)



(b) Query Scene Graph

Figure 3: Average performance case. (a) The highest-ranked retrieved image, which is incorrect but structurally related. (b) The query scene graph. The ground-truth image in this case was retrieved at Rank 23. This result illustrates the issue of semantic ambiguity and feature overlap; the model retrieves an image with a matching spatial arrangement but substitutes visually analogous objects. This suggests that while the geometric constraints are robust, the AlexNet visual features sometimes lack the fine-grained discriminative power needed to separate similar classes.