Semantic and Structural Graph Enhancements for Scene Graph Generation

Abstract.

Scene Graph Generation (SGG) structures visual content into graphs of entities (objects) and relationships (predicates), enabling structured visual understanding. We present a systematic empirical study of semantic integration strategies in two state-of-the-art graph-based frameworks. Through controlled experiments, we find contextualized entity embeddings (BERT) drive most performance improvements: 19.90% R@50 (+0.60% over static Word2Vec) on Visual Relationship Dataset, and 29.09% R@100 (+0.49% absolute) on Visual Genome. Our key insight is that visual union features encode sufficient relational context, making explicit predicate semantics redundant. This counter-intuitive result is confirmed via comprehensive ablation. We further introduce a learned confidence-gating mechanism that adaptively filters noisy predicates, improving rare-relationship recall by 29.8% versus hard spatial pruning. Our findings provide a rigorous design framework for semantic integration in SGG, with implications for efficient model design.[1]

1 Introduction

The creation of scene graphs, which represent objects as entities (nodes) and their relationships as edges, enables structured visual understanding, an essential step for downstream tasks such as image captioning [1]. Graph neural networks (GNNs) excel at SGG by explicitly modeling structural dependencies between nodes, propagating contextual information to resolve relationship ambiguities through iterative message passing [2, 3]. However, constructing accurate scene graphs that balance computational efficiency with semantic richness remains challenging. Traditional methods process objects independently, missing crucial contextual dependencies. While GNNs address contextual learning through message passing, they still face challenges such as the long-tail predicate distribution (where 80% of relationships belong to only 20% of categories) and computational constraints arising from dense graph connectivity ($\mathcal{O}(N^2)$ complexity).

Early SGG methods [4, 5] used a two-stage pipeline, that separates object detection and relationship prediction. Later work added contextual modeling with MotifNet [5], graph refinement via Graph R-CNN [6], and hierarchical learning through VCTree [7]. More recent approaches use a bipartite graph [2], while addressing data imbalance through reweighting and bias mitigation. Semantic embeddings from Word2Vec [8], GloVe [9], and BERT [10] have shown promise in NLP but remain underexplored for visual relationship understanding.

Scene graph generation requires balancing semantic richness with computational efficiency. We extend two complementary architectures: a **homogeneous**

[1] Source code: `https://github.com/walidgeuttala/scene_graph_generation`.

graph with lightweight attention-based edges for efficiency, and a **bipartite graph** with confidence-gated passing for relational depth. Our architecture-dataset co-design achieves 19.90% R@50 on VRD and 29.09% R@100 on VG dataset. This is enabled by contextualized semantics: BERT+ attention improves +0.86% R@50 over static embeddings by dynamically modulating visual-semantic fusion. Critically, ablations reveal entity-level semantics contribute substantially more than predicate semantics, while learned attention suppresses false positives more effectively than hard spatial pruning. We evaluate on VRD [4] (5K images, 70 predicates) and VG dataset [11] (108K images, 41K relationships).

2 Method

We propose and evaluate two graph-based architectures for scene graph generation: a **homogeneous object graph** treating relationships as edge features, and a **bipartite entity-predicate graph** treating relationships as explicit nodes. Both share a common feature extraction pipeline but employ distinct graph construction, message passing, and classification strategies. Each architecture processes images through four stages (Fig. 1): (i) Feature Extraction via Faster R-CNN (Frozen), (ii) Graph Construction with filtering, (iii) Message Passing to refine representations, and (iv) Relationship Detection via specialized classifiers. Following standard two-stage practice, the detector remains frozen while graph components are trained end-to-end.

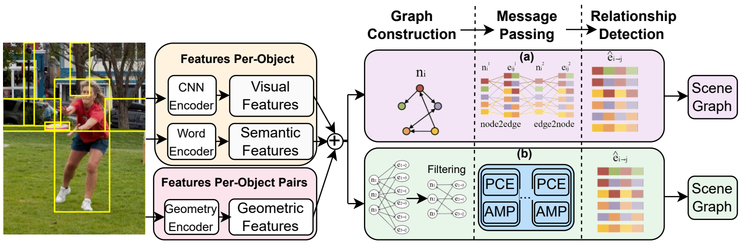

Fig. 1: Scene graph architectures. Three encoders extract features from Faster R-CNN detections: CNN (visual), Word (semantic), and Geometric (spatial/union). **(a) Homogeneous:** Objects as nodes, relationships as edges. Bidirectional node-edge message passing with attention weights. **(b) Bipartite:** Separate object and predicate nodes with confidence-gated updates.

2.1 Feature Representation

Features per Objects. Faster R-CNN with backbone ϕ (VGG16/ResNet-101) detects objects with bounding boxes $\{b_i\}_{i=1}^{N}$. For each object i, the CNN Encoder extracts L_2-normalized visual embeddings $\tilde{\mathbf{v}}_i \in \mathbb{R}^{d_v}$ via RoI Align from $\mathcal{F} = \phi(\mathcal{I})$. The Word Encoder maps labels c_i to semantic vectors $\mathbf{s}_i \in \mathbb{R}^{d_s}$ using pretrained embeddings (Word2Vec, GloVe, or BERT). The Position Encoder projects box coordinates to spatial embeddings $\mathbf{l}_i \in \mathbb{R}^{d_l}$. Concatenated features $[\tilde{\mathbf{v}}_i \oplus \mathbf{s}_i \oplus \mathbf{l}_i] \in \mathbb{R}^{d_v+d_s+d_l}$ form initial node embedding $\mathbf{n}_i^{(0)}$.

Features per Object Pairs. For each pair (i, j), the *Geometric Encoder* computes spatial features

$$\mathbf{1}_{ij} = [\Delta(b_i, b_j) \oplus \Delta(b_i, b_{ij}) \oplus \Delta(b_j, b_{ij}) \oplus \mathrm{IoU}(b_i, b_j) \oplus \mathrm{dis}(b_i, b_j)] \tag{1}$$

where $\Delta(b_i, b_j) = [\frac{x_i - x_j}{w_j}, \frac{y_i - y_j}{h_j}, \log(\frac{w_i}{w_j}), \log(\frac{h_i}{h_j})] \in \mathbb{R}^4$ encodes relative position/scale, b_{ij} is the union box, IoU measures overlap, and $\mathrm{dis}(b_i, b_j) = \|c_i - c_j\|_2$ is the center distance. These are embedded into object nodes during initialization. The Union Visual Encoder extracts features $\mathbf{u}_{ij}$ from region b_{ij} to form initial edge representation $\mathbf{e}_{ij}^{(0)} = f_{\mathrm{edge}}(\mathbf{u}_{ij})$.

2.2 Homogeneous Object Graph

Construction (Fig. 1a, top). Directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has object nodes $\mathcal{V}$ and relationship edges $\mathcal{E}$. Edges are created via spatial heuristics: $\mathbf{1}_{[e_{ij} \in \mathcal{E}]} = \mathbf{1}_{[\mathrm{dis}(b_i, b_j) < 0.5 \vee \mathrm{IoU}(b_i, b_j) > 0.1]}$, where $\mathbf{1}_A(\cdot)$ is the indicator function of A.

Message Passing. Through L iterations, edges aggregate node embeddings via attention-weighted summation

$$\mathbf{e}_{ij}^{(l+1)} = \mathrm{GRU}_e \left(\mathbf{e}_{ij}^{(l)}, \alpha_{ij}^s \mathbf{W}_e \mathbf{n}_i^{(l)} + \alpha_{ij}^o \mathbf{W}_e \mathbf{n}_j^{(l)} \right) \tag{2}$$

where α are attention weights and GRU_e is the edge GRU cell. Nodes aggregate from incident edges

$$\mathbf{n}_i^{(l+1)} = \mathrm{GRU}_n\big(\mathbf{n}_i^{(l)}, \ \tfrac{1}{|\mathcal{N}_{\mathrm{out}}(i)|} \sum_{j \in \mathcal{N}_{\mathrm{out}}(i)} \beta_{ij} \mathbf{W}_n \mathbf{e}_{ij}^{(l+1)} + \tfrac{1}{|\mathcal{N}_{\mathrm{in}}(i)|} \sum_{k \in \mathcal{N}_{\mathrm{in}}(i)} \beta_{ki} \mathbf{W}_n \mathbf{e}_{ki}^{(l+1)}\big) \tag{3}$$

where β are learned attention scores, $\mathcal{N}_{\mathrm{out}}(i)$ and $\mathcal{N}_{\mathrm{in}}(i)$ denote neighbors, and GRU_n is the node GRU cell.

Detection (Fig. 1a, top). Final edge embeddings are classified via $\mathbf{p}_{ij} = \mathrm{softmax}(\mathbf{W}_{\mathrm{rel}} \mathbf{e}_{ij}^{(L)})$, preserving directionality for asymmetric relationships.

2.3 Bipartite Entity-Predicate Graph

Construction (Fig. 1b, bottom). Bipartite graph $\mathcal{G} = (\mathcal{V}_e, \mathcal{V}_p, \mathcal{E}_{e2p}, \mathcal{E}_{p2e})$ has entity nodes $\mathbf{n}_i \in \mathcal{V}_e$ and predicate nodes $\mathbf{r}_{i \to j} \in \mathcal{V}_p$ initialized as $\mathbf{r}_{i \to j}^{(0)} = f_r(\mathbf{u}_{ij})$. Edges $\mathcal{E}_{e2p}$ and $\mathcal{E}_{p2e}$ connect entities to their subject/object predicates.

Message Passing (PCE and AMP). The bipartite graph employs two key modules: Predicate Confidence Estimation (PCE) and Adaptive Message Passing (AMP). A confidence estimator $s_{ij}^b = \sigma(\mathbf{w}_b^\top g_x([\mathbf{r}_{i \to j} \oplus \mathbf{n}_i \oplus \mathbf{n}_j]))$ assesses predicate reliability, which modulates information flow via learnable gating

$$\gamma_{ij} = T(s_{ij}^b) = \begin{cases} 0 & s_{ij}^b \leq \beta \\ \alpha s_{ij}^b - \alpha\beta & \beta < s_{ij}^b < \frac{1}{\alpha} + \beta \\ 1 & s_{ij}^b \geq \frac{1}{\alpha} + \beta \end{cases} \tag{4}$$

where $T(\cdot)$ is the piecewise gating function with learnable parameters α and β. Bidirectional updates with affinity weights d_s and d_o for subject/object roles are

$$\mathbf{r}_{i\to j}^{(l+1)} = \mathbf{r}_{i\to j}^{(l)} + \varphi(d_s \mathbf{W}_r^\top \mathbf{n}_i^{(l)} + d_o \mathbf{W}_r^\top \mathbf{n}_j^{(l)}) \tag{5}$$

$$\mathbf{n}_i^{(l+1)} = \mathbf{n}_i^{(l)} + \varphi\big(\tfrac{1}{|B_s(i)|}\sum_{k\in B_s(i)}\gamma_k d_s \mathbf{W}_n^\top \mathbf{r}_k^{(l)} + \tfrac{1}{|B_o(i)|}\sum_{k\in B_o(i)}\gamma_k d_o \mathbf{W}_n^\top \mathbf{r}_k^{(l)}\big) \tag{6}$$

where $B_s(i)$ and $B_o(i)$ denote predicates with entity i as subject and object respectively, and φ is ReLU activation.

Detection (Fig. 1b, bottom). Predicate classification uses $\mathbf{p}_{r_{i\to j}} = \mathrm{softmax}(\mathbf{W}_{\mathrm{rel}}^\top \mathbf{r}_{i\to j}^{(L)} + \log(\hat{\mathbf{p}}_{\mathrm{freq}}))$ with frequency priors $\hat{\mathbf{p}}_{\mathrm{freq}}$. Entity classification balances contextual refinement and visual evidence via $\mathbf{p}_{n_i} = \mathrm{softmax}(\mathbf{W}_{\mathrm{ent}}^\top(\rho \cdot \mathbf{n}_i^{(L)} + (1-\rho)\tilde{\mathbf{v}}_i))$ where $\rho \in [0,1]$ is a learnable interpolation weight and L is the total number of iterations.

3 Experimental Results

Table 1 shows results from our architecture-dataset co-design approach. **For VRD** (5K images, 70 predicates), we use the lightweight homogeneous graph with attention mechanisms. **For Visual Genome** (108K images, 41K relationships), we employ the bipartite graph with confidence-gating to handle the dataset's scale and complexity. This pairing leverages each architecture's strengths: homogeneous graphs for efficient processing of moderate scenes, bipartite graphs for managing dense relational structures. The homogeneous architecture extends the message passing framework of (Word2Vec) with semantic embeddings and attention mechanisms [3]. The bipartite architecture extends the entity-predicate graph formulation of (No semantics) with BERT semantics and adaptive gating [2].

On VRD, BERT+Attention achieves best performance (19.90% R@50, 23.58% R@100), representing +0.60% (R@50) and +0.66% (R@100) absolute improvements over the Word2Vec baseline [3]. The attention mechanism provides greater benefits for contextual embeddings: BERT shows +0.38%/+0.44% gains compared to Word2Vec's +0.25%/+0.31%, indicating stronger synergy between transformer based representations and learned edge weighting. On Visual Genome, BERT entity embeddings outperform the no-semantics baseline by +0.86/+0.49 points (R@50/R@100) while achieving competitive mean recall, demonstrating the value of contextualized representations in complex relational learning.

The homogeneous architecture's attention weights effectively suppress spatially proximate but semantically weak relationships, substantially reducing false positives in cluttered scenes through learned prioritization (e.g., up-weighting "person-riding-elephant" while down-weighting "person-near-tree"). The bipartite graph's explicit predicate modeling combined with confidence-gating achieves superior recall on complex scenes (29.09% R@100) while maintaining rare relationship coverage, distinguishing fine-grained interactions like "holding" versus "carrying" through explicit node representations.

Table 1: Performance comparison across graph architectures and semantic encoders. Starred (*) entries denote our contributions. **Metrics:** R@K = Recall at top-K predictions while mR@K = Mean Recall at K (per-category average, crucial for assessing long-tail performance on VG's imbalanced distribution). N/A: VRD's relatively balanced distribution makes mR@K less critical, baseline [3] reports only R@K.

Word Encoder	**R@50**	**R@100**	**mR@100**
Homogeneous Graph (VRD)			
Word2Vec [3]	19.30	22.92	N/A
Word2Vec + Attention*	19.55	23.23	N/A
BERT*	19.52	23.14	N/A
BERT + Attention*	**19.90**	**23.58**	N/A
Bipartite Graph (VG)			
No semantics [2]	23.77	28.60	4.64
Word2Vec*	24.32	28.75	5.70
GloVe*	23.77	28.60	**6.16**
BERT*	**24.63**	**29.09**	6.02

3.1 Ablation Study on Semantic Integration

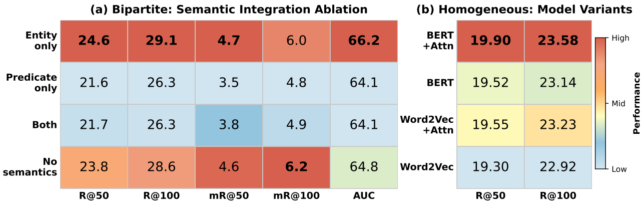

Fig. 2: Semantic ablation across architectures. **(a) Bipartite (VG):** Entity-only semantics dominate recall/AUC while no-semantics preserves tail diversity. **(b) Homogeneous (VRD):** BERT+Attention achieves consistent gains across metrics. Color intensity encodes relative performance. Bold values indicate best per metric.

Fig. 2 systematically ablates semantic integration across both architectures. On VG's bipartite graph (a), entity-only features achieve optimal recall (29.09% R@100) and AUC (66.2), while no-semantics maximizes tail performance (6.16% mR@100), revealing a precision-diversity trade-off. Predicate-only semantics underperform across all metrics, and combined features show redundancy suggesting visual union features encode sufficient relational context. On VRD's homogeneous graph (b), BERT+Attention consistently outperforms alternatives (+0.60% R@50, +0.66% R@100 over Word2Vec), with contextualized embeddings showing stronger synergy with learned attention than static representations. This cross-architecture validation confirms: (1) entity semantics provide maximal discriminative power, (2) predicate semantics add limited marginal value, (3) learned attention mechanisms amplify contextualized embeddings more effectively than static ones.

4 Conclusion

We demonstrate scalable graph-based SGG through architecture-dataset co-design: homogeneous graphs with BERT+Attention achieve competitive VRD performance (19.90% R@50) efficiently, while bipartite graphs with entity-focused semantics excel on complex VG scenes (29.09% R@100). Key principles emerge from systematic ablation: (1) Entity semantics dominate predicate semantics across both architectures, (2) Contextualized BERT embeddings outperform static alternatives with stronger attention synergies, (3) Learned prioritization surpasses hard pruning for preserving rare relationships. Future work includes causal inference for long-tail predicate, hybrid architectures with dynamic topology adaptation, and vision-language models for open-vocabulary detection.

References

[1] T. Ghandi, H. Pourreza and H. Mahyar. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1-39, 2023.

[2] R. Li, S. Zhang, B. Wan and X. He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. *Proc. IEEE/CVF CVPR*, 11109-11119, 2021.

[3] Y. Hu, S. Chen, X. Chen, Y. Zhang and X. Gu. Neural message passing for visual relationship detection. *ArXiv:2208.04165*, 2022.

[4] C. Lu, R. Krishna, M. Bernstein and L. Fei-Fei. Visual relationship detection with language priors. *Proc. ECCV*, 852-869, 2016.

[5] R. Zellers, M. Yatskar, S. Thomson and Y. Choi. Neural motifs: Scene graph parsing with global context. *Proc. IEEE CVPR*, 5831-5840, 2018.

[6] J. Yang, J. Lu, S. Lee, D. Batra and D. Parikh. Graph R-CNN for scene graph generation. *Proc. ECCV*, 670-685, 2018.

[7] K. Tang, H. Zhang, B. Wu, W. Luo and W. Liu. Learning to compose dynamic tree structures for visual contexts. *Proc. IEEE/CVF CVPR*, 6619-6628, 2019.

[8] T. Mikolov, K. Chen, G. Corrado and J. Dean. Efficient estimation of word representations in vector space. *ArXiv:1301.3781*, 2013.

[9] J. Pennington, R. Socher and C. Manning. GloVe: Global vectors for word representation. *Proc. EMNLP*, 1532-1543, 2014.

[10] J. Devlin, M. Chang, K. Lee and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL*, 4171-4186, 2019.

[11] R. Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Computer Vision*, 123(1):32-73, 2017.