LEARNING SPATIALLY-AWARE REPRESENTATIONS FROM SCENE GRAPHS USING OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual similarity is typically considered from the semantic content of images, while the structure and the spatial organization of objects are often discarded. Spatial information curated from objects (such as bounding box coordinates and relative position relations between objects) is usually absent from the neural model similarity assessment. In the meantime, scene graphs computed from images provide an intermediate representation for explicitly modeling the structure of objects in images, yet their spatial configurations remain underexplored, limiting their usability for spatially-aware computer vision tasks like retrieval based on object arrangement. To address this gap and in this effort to better account for spatial information in images, we propose and formalize the Spatial Scene Graph (SSG), which encodes absolute and relative positions between objects as node and edge attributes, respectively. Building on this structured representation, we propose SPOT (SPatial scene graph Optimal Transport), a novel criterion for spatial similarity between SSGs, and indirectly between images. SPOT allows both the ability to evaluate neural models based on the spatial similarity found in the images and the training of neural backbones to learn more spatially-aware representations. We experimentally demonstrate these claims across two spatially-aware downstream tasks: content-based image retrieval and visual question answering.

1 INTRODUCTION

In computer vision, visual similarity is typically framed as a semantic problem: two images are similar if they share a large proportion of objects found in both scenes. Large-scale contrastive models such as CLIP (Radford et al., 2021) have shown robust performance in capturing rich semantic information from images, which enables strong semantic-aware retrieval results. However, these models fall short in capturing image compositionality (Thrush et al., 2022; Ma et al., 2023), relative positions (Kamath et al., 2023) and the underlying spatial structure between the objects in a scene (Yüksekgönül et al., 2023). These image components (compositionality and structure) are important aspects to be considered when quantifying visual similarity for many downstream tasks as Gestalt psychology emphasizes that humans organize visual scenes according to structural relationships (e.g. proximity, closure, layout), often prior to semantic interpretation (Koffka, 1935).

Attempts to formalize the layout of an image include scene graphs (Johnson et al., 2015). A scene graph is a structured representation of the image contents in which nodes represent objects and edges represent the relationships between them (Krishna et al., 2016), see e.g. Figure 1. While such structured representations are valuable for many scene understanding tasks such as visual question answering (Hudson & Manning, 2019) and image captioning (Anderson et al., 2016), scene graphs encode spatial relationships as categorical predicates (e.g. *to the left of*, *above*) rather than fine-grained spatial information and therefore fall short of capturing the spatial structure of the image. Furthermore, spatial relationships are prone to annotation biases (Tversky & Hard, 2009; Levinson, 1996) and polysemy in which a spatial relation may entail multiple spatial configurations (Liao et al., 2019; Chacra & Zelek, 2022). This predicament limits both the ability to learn fine-grained spatial representations (Servant et al., 2025; 2026) and the ability to accurately estimate the similarity between scene graphs, as it signals the presence of a different configuration without quantifying its extent (Anderson et al., 2016). Estimating the similarity between images represented as scene graphs then remains an open problem. Approaches such as graph-edit distance variants (Chaidos

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

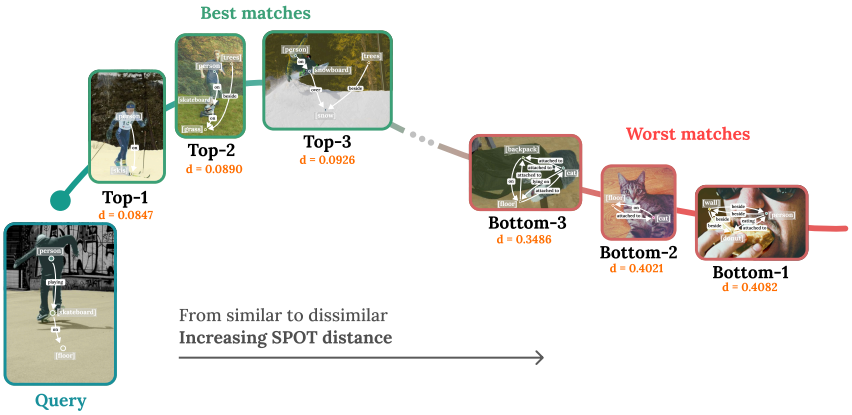

Figure 1: **Qualitative spatial retrieval with SPOT on the PSG dataset.** SPOT compares images according to the spatial arrangement of their objects and defines a quantitative notion of spatial similarity. Retrieved scenes are ranked by increasing SPOT scores, with lower values indicating more similar spatial configurations. Overlaid scene graphs show the objects and directed relations involved in the comparison.

et al., 2025) and graph-based embeddings (such as IRSGS (Yoon et al., 2021)) only approximate semantic similarity between scene graphs and fail to account for fine-grained spatial layout.

In this paper, we aim to move towards more spatialized image representations, enabling image comparisons that are more closely aligned with human perception of a scene. To achieve this, we use scene graphs derived from input images as intermediate proxy representations. We investigate the above mentioned challenges through the formalization of the Spatial Scene Graph (SSG), a scene graph that encodes absolute and relative spatial information between the objects from the associated image. Both types of spatial information are encoded as fine-grained quantitative vectors enabling the computation of pairwise distances between nodes and edges respectively, therefore framing the similarity between SSGs as an optimal transport problem. Building upon this, we introduce SPOT (SPatial scene graph Optimal Transport), a method for assessing spatial similarity between spatial scene graphs leveraging multi-dimensional fused Gromov–Wasserstein distance (Kawano et al., 2024). We experimentally show that SPOT can be used both as an evaluation criterion for spatial similarity between scene graphs (Figure 1) and involved in a spatially-aware representation learning framework with S3PE, a novel self-supervised training method which is leveraged to fine-tune vision encoders and train graph neural networks leading to spatially-informed model embeddings. To summarize, our contributions are as follows:

- We formalize the **Spatial Scene Graph** (SSG), a type of scene graph that encodes fine-grained spatial information of an image of a scene without semantic information;
- We introduce **SPOT**, a quantitative criterion for spatial similarity between SSGs leveraging multi-dimensional fused Gromov–Wasserstein distance, and experimentally demonstrate its usability as an evaluation criterion for spatial similarity in retrieval tasks;
- We ultimately show that SPOT can be efficiently leveraged as a supervision signal to train either image and graph models in a self-supervised fashion, leading to a novel method: **SPOT-Supervised SPatial Embeddings (S3PE)**.

2 Related Works

Similarity between scene graphs. The assessment of similarity between scene graphs is a complex task that involves comparing objects and their relationships, while accounting for the structural configurations. SPICE (Anderson et al., 2016) compares image captions by transforming sentences into scene graphs and comparing triplets of objects and relationships, which does not fully take into account the structural configuration of the scene graphs. The Graph Edit Distance (Chen et al., 2020a; Chaidos et al., 2025) induces both graph structure and semantic content in the similarity

computation, but does not take into account fuzzy or hierarchical similarities. Optimal Transport (OT) (Villani, 2009) enables this by providing a soft matching mechanism between nodes and edges. OT has been extensively used in various domains but remains underexplored with respect to scene graphs with most OT-based approaches used only for the generation of scene graphs (Kurita et al., 2023), but not their assessment.

In this context, embeddings-based methods for measuring the similarity between scene graphs have risen as an alternative, leveraging self-supervised image or graph learning to capture structural and semantic similarities. IRSGS (Yoon et al., 2021) trains a GNN to predict the relevance of the scene graph and the image caption. The authors of (Maheshwari et al., 2021) leverage a relative similarity metric based on semantic similarity with a contrastive ranking loss, while SPAN (Cong et al., 2026) aligns graph and image embeddings with a contrastive loss function. While powerful, these methods do not take into account the spatial configurations of objects in the images. Furthermore, as they leverage semantic information, the models may struggle with out-of-distribution semantic representations (Li et al., 2025). Our approach focuses on structural and spatial properties. SPOT is domain-agnostic and highlights fine-grained spatial similarities.

Relative Position Descriptors (RPDs). RPDs describe the spatial configuration between two objects in an image, first formalized by (Freeman, 1975) using elementary relations (e.g. directional, topological). The Force Histogram (Matsakis & Wendling, 1999) computes the force applied by one object to the other in each direction dependent on a force level. The Force Banner (Deléarde et al., 2022) extends the Force Histogram to stack Force Histograms with different force levels to provide a thorough spatial representation of the objects in the scene. Other RPDs such as the ϕ-descriptor (Matsakis et al., 2015) combine existing descriptors for thorough spatial descriptions.

Amongst existing RPDs, the Force Banner has been leveraged for predicting spatial relations (Deléarde et al., 2022), and for training models in a self-supervised manner (Servant et al., 2025; 2026). Bags-of-Spatial-Relations (Clément et al., 2016) leverage RPDs to model spatial relations between the pairwise structural components of complex objects. RPDs have also been leveraged with attributed relational graphs for scene recognition (Clément et al., 2018), highlighting the applicability of RPDs in graphs. As Force Banners model fine-grained spatial representations of directional and distance relations, we rely on Force Banners as edge representations of SSGs.

Self-supervised learning of spatial representation. Self-supervision loss functions (e.g. SimCLR (Chen et al., 2020b) and InfoNCE (van den Oord et al., 2018)) have been the staple method for training multimodal models, especially for aligning vision and text. GraphCL (You et al., 2020) leverages SimCLR to learn node representations with graph-focused augmentations such as edge dropping and node perturbations. More recently, self-supervised methods that consider spatial information as a dedicated modality have been developed for spatially-aware representations. SpIRL (Servant et al., 2026) leverages the triplet loss function and the reconstruction loss to build robust spatially-aware image representations, while C-SIP (Servant et al., 2025) trains image-spatial contrastive methods with InfoNCE. Riva et al. (2022) trains a U-Net for directional relations awareness. These three methods focus solely on pairs of objects, limiting the ability to capture spatial information across the whole image. Our approach, S3PE, captures the spatial similarity across the whole image, potentially composed of numerous objects, instead of focusing on solely object pairs.

3 Spatial Scene Graph

We introduce the Spatial Scene Graph (SSG), a quantitative representation of scene layout in which nodes encode absolute object geometry and directed edges encode relative spatial configurations.

3.1 Preliminaries and notation

Let $\mathcal{G}^{\text{sem}} = (\mathcal{V}, \mathcal{E})$ denote a semantic scene graph, where $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ is the set of nodes, in which each node $v_i \in \mathcal{V}$ is annotated with a semantic label c_i. $\mathcal{E}$ is the set of directed edges where each edge $e_{ij} \in \mathcal{E}$ corresponds to the semantic relationship between subject v_i and object v_j, forming a triplet (v_i, e_{ij}, v_j).

As argued earlier, the semantic scene graph does not enforce the spatial position of objects in the image space as it lacks fine-grained spatial information in the graph representation. We propose

instead to leverage a Spatial Scene Graph (SSG) $\mathcal{G}^{\text{sp}} = (\mathcal{V}, \mathcal{E})$ that retains the structure of $\mathcal{G}^{\text{sem}}$ in which nodes encode bounding box coordinates $b_i \in \mathbb{R}^4$ for each object $v_i \in \mathcal{V}$, and edges $e_{ij} \in \mathcal{E}$ encode quantitative spatial relations. We employ RPDs as edge representations (Wang et al., 2023), and specifically the Force Banner (Deléarde et al., 2022), which encodes the spatial configuration of two objects in a 2D representation, enabling the assessment of directional and distance relations.

3.2 Force Banner for edge attributes

The Force Banner (Deléarde et al., 2022) is a RPD that extends the Force Histogram (Matsakis & Wendling, 1999) which models the spatial relationship between two crisp objects in an image. Given two points a and b at a distance d, the force is computed as $\varphi_r(d) = \frac{1}{d^r}$ with r a force level acting as a weight parameter for points either closer ($r > 0$) or farther ($r < 0$) apart. The Force histogram computes the force value between two 1D segments I and J at a distance D^θ_{IJ}:

$$f_r(I, J) = \int_{D^\theta_{IJ}}^{|I|+D^\theta_{IJ}+|J|} \int_0^{|J|} \varphi_r(u - v)\, \mathrm{d}v\mathrm{d}u. \tag{1}$$

Given two crisp objects A and B, the global attraction $F^{AB}(r, \theta)$ along an angle θ is computed as:

$$F^{AB}(r, \theta) = \sum_{I \in C^\theta_A} \sum_{J \in C^\theta_B} f_r(I, J) \tag{2}$$

where C^θ_A and (respectively C^θ_B) is the set of segments that intersect between parallel lines L_θ and object A (respectively object B). The Force Histogram is then computed on the set of $\Theta = \{0, 0 + \delta_\theta, ..., 2\pi - \delta_\theta\}$ where δ_θ denotes a constant step between angles. The Force Banner extends the Force Histogram by computing the histogram along a set of force levels $R = \{r_{min}, r_{min} + \delta_r, ..., r_{max}\}$ into a 2D matrix such that:

$$\mathbf{FB}^{AB} = \left[F^{AB}(r_i, \theta_j)\right]_{i,j} \tag{3}$$

with $r_i \in R$, $r_{min} = -2.24$, $r_{max} = 2.23$ with $\delta_r = 0.08$ and $\theta_j \in \Theta$, $\delta_\theta = \pi/56$, leading to a final Force Banner dimension of $(56, 112)$. Additional information about the Force Banner is provided in Appendix A. Force Banners provide useful properties under geometric transformations, including translation invariance (Deléarde et al., 2022), making them suitable for comparing relative spatial configurations within an optimal transport framework.

4 SPOT: SPatial scene graph Optimal Transport

4.1 Problem formulation

We formulate spatial similarity between scene graphs as an optimal transport problem (Villani, 2009). Standard fused Gromov–Wasserstein formulations (Vayer et al., 2018) combine node-feature costs with pairwise structural differences, but do not explicitly compare rich edge attributes. Extensions such as multi-dimensional Gromov–Wasserstein (MFGW) (Kawano et al., 2024) generalize this framework to edge-attributed graphs. Building on this family of formulations, we introduce SPatial scene graph Optimal Transport (SPOT), a training-free method that quantifies spatial similarity between spatial scene graphs. SPOT combines absolute spatial information, encoded by object bounding boxes, with relative spatial information, encoded by Force Banners (Deléarde et al., 2022). These two components define node-to-node and edge-to-edge costs respectively, which are jointly integrated through optimal transport.

4.2 Spatial scene graph matching

Let $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ be two SSGs that contain $n = |\mathcal{V}_1|$ and $m = |\mathcal{V}_2|$ nodes, respectively. We associate the two node sets with probability vectors $\mathbf{p} \in \mathbb{R}^n$ and $\mathbf{q} \in \mathbb{R}^m$. Since all nodes are equally weighted, we set $\mathbf{p} = \frac{1}{n}\mathbf{1}_n$, $\mathbf{q} = \frac{1}{m}\mathbf{1}_m$, where $\mathbf{1}_d \in \mathbb{R}^d$ denotes the all-ones vector of dimension d.

A transport plan $\mathbf{T} \in \mathbb{R}_+^{n \times m}$ defines a soft correspondence between the nodes of the two graphs, where T_{ij} denotes the amount of mass transported from node $i \in \mathcal{V}_1$ to node $j \in \mathcal{V}_2$. We denote by $\Pi(\mathbf{p}, \mathbf{q})$ the set of admissible transport plans whose marginals match the node-mass distributions:

$$\Pi(\mathbf{p}, \mathbf{q}) = \left\{\mathbf{T} \in \mathbb{R}_+^{n \times m} \;\middle|\; \mathbf{T}\mathbf{1}_m = \mathbf{p},\ \mathbf{T}^\top \mathbf{1}_n = \mathbf{q}\right\}. \tag{4}$$

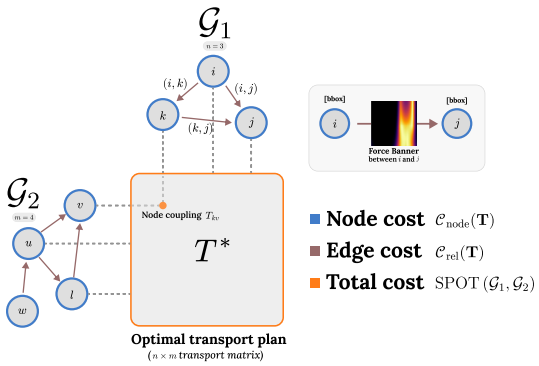

Figure 2: **Overview of SPOT for spatial similarity.** Nodes encode object geometry and directed edges encode Force Banners. SPOT jointly matches node and relation-level information through an optimal transport plan $\mathbf{T}^*$.

4.3 Node and Relational Costs

For the node cost, each bounding box is represented by its normalized center $\mathbf{c} = (c_x, c_y)$ and normalized size $\mathbf{s} = (w, h)$. The cost of matching node $i \in \mathcal{V}_1$ with node $j \in \mathcal{V}_2$ is defined as

$$M_{ij} = \frac{\|\mathbf{c}_i - \mathbf{c}_j\|_2 + \|\mathbf{s}_i - \mathbf{s}_j\|_2}{2\sqrt{2}}. \tag{5}$$

Since bounding box centers and sizes are normalized to $[0,1]^2$, M_{ij} is bounded in $[0,1]$.

For each directed edge, we sum the Force Banner along its force-level axis and normalize the resulting angular profile to unit mass over N_θ angular bins. Since the angular domain is periodic, we compare angular profiles using the closed-form circular 1-Wasserstein distance (Rabin et al., 2011). Given two normalized angular profiles $\mathbf{a}$ and $\mathbf{b}$, let $c_h = \sum_{t=1}^{h}(a_t - b_t)$ denote the cumulative mass difference up to angular bin $h = 1, \ldots, N_\theta$. We compute the 1-Wasserstein distance as

$$W_1(\mathbf{a}, \mathbf{b}) = \frac{2}{N_\theta} \sum_{h=1}^{N_\theta} |c_h - \text{median}(c_1, \ldots, c_{N_\theta})| \,. \tag{6}$$

The median term accounts for the circular nature of the angular domain, avoiding artifacts induced by the arbitrary cut between the first and last angular bins (Geist et al., 2024). As the angular profiles have unit mass and the circular domain is normalized, $W_1(\mathbf{a}, \mathbf{b}) \in [0,1]$.

We denote by $\mathbf{x}_{ik}$ and $\mathbf{x}_{jl}$ the normalized angular profiles derived from the Force Banners associated with the ordered node pairs (i,k) and (j,l), respectively. Their relational cost is defined as

$$L_{ikjl} = \begin{cases} 0, & (i,k) \notin \mathcal{E}_1,\ (j,l) \notin \mathcal{E}_2, \\ c_{\text{miss}}, & \text{if exactly one edge exists}, \\ W_1(\mathbf{x}_{ik}, \mathbf{x}_{jl}), & (i,k) \in \mathcal{E}_1,\ (j,l) \in \mathcal{E}_2. \end{cases} \tag{7}$$

The missing-edge cost c_{miss} penalizes the presence of a directed relation in only one of the two graphs. Since $W_1 \in [0,1]$, we set $c_{\text{miss}} = 1$, corresponding to the maximal relational cost on this normalized scale.

4.4 SPOT Objective and Optimization

Given a transport plan $\mathbf{T}$, the node-level and relational costs are respectively

$$\mathcal{C}_{\text{node}}(\mathbf{T}) = \sum_{i,j} M_{ij} T_{ij}, \qquad \mathcal{C}_{\text{rel}}(\mathbf{T}) = \sum_{i,k,j,l} L_{ikjl} T_{ij} T_{kl}. \tag{8}$$

We define SPOT as the minimum joint cost over all admissible transport plans:

$$\text{SPOT}\,(\mathcal{G}_1, \mathcal{G}_2) = \min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} \left[\alpha \mathcal{C}_{\text{node}}(\mathbf{T}) + (1-\alpha)\mathcal{C}_{\text{rel}}(\mathbf{T})\right], \tag{9}$$

where $\alpha \in [0, 1]$ controls the contribution of absolute and relative spatial information. Since M_{ij} and L_{ikjl} are bounded in $[0, 1]$ and $\mathbf{T}$ has unit total mass, both $\mathcal{C}_{\text{node}}$ and $\mathcal{C}_{\text{rel}}$ are also bounded in $[0, 1]$. We therefore set $\alpha = 0.5$ to assign equal weight to the two normalized cost terms, without assuming a prior preference for either source of spatial information. The influence of α is further illustrated in Appendix E. Lower SPOT values indicate greater spatial similarity.

We optimize equation 9 using the Frank–Wolfe algorithm. Starting from the independent coupling $\mathbf{pq}^\top$, each iteration solves a linear optimal transport subproblem based on the gradient of the current objective, followed by an exact line search. The procedure is repeated until convergence. An efficient reformulation of the objective and its computational complexity are detailed in Appendix B.

SPOT provides a quantitative signal of spatial similarity between images represented as scene graphs that can be used to supervise representation learning. Using this signal, we learn embedding spaces in which distances reflect differences in spatial scene organization, leading to our S3PE framework.

5 SPOT-SUPERVISED SPATIAL EMBEDDINGS (S3PE)

We introduce SPOT-Supervised SPatial Embeddings (S3PE), a model-agnostic framework for learning spatially-aware representations. We use SPOT-based spatial similarity to construct triplets of spatially similar and dissimilar scenes and train encoders with a triplet loss function, encouraging spatially similar scenes to be embedded closer than more dissimilar ones.

5.1 ARCHITECTURE

S3PE is a model-agnostic training method. To highlight its versatility, we consider two different model architectures with different inputs: a pre-trained DINOv2 (Oquab et al., 2023) image encoder with a projection head $\mathbf{z}^{\mathcal{I}} = g(E(\mathcal{I}))$ and two graph neural networks that leverage either semantic or spatial scene graphs ($\mathcal{G}^{sem}$ or $\mathcal{G}^{sp}$ respectively). We define hereinafter the GNN architectures.

Semantic scene graphs leverage word representations (Reimers & Gurevych, 2019) as both node and edge inputs. Spatial scene graphs are leveraged as in Section 3, such that each node $v_i \in \mathcal{V}$ is described by its object bounding box parametrization, while directed edges $e_{i,j} \in \mathcal{E}$ are associated with the Force Banners computed from the corresponding objects.

Given $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ either the semantic or spatial scene graph, our objective is to learn a graph-level representation that reflects the spatial similarity measured by SPOT. Node and edge features are first projected into a latent space with a feed-forward network:

$$\mathbf{h}_i^{(0)} = f_{\text{node}}(\mathbf{v}_i), \qquad \mathbf{r}_{ij}^{(0)} = f_{\text{edge}}(\mathbf{e}_{i,j}). \tag{10}$$

The resulting projections are processed by four GATv2 layers (Brody et al., 2022) to iteratively update node representations. Node and edge representations are then aggregated with a global mean pooling function, followed by a projection head $g(\cdot)$ to obtain a graph-level representation $\mathbf{z}_{\mathcal{G}} = g(\mathbf{h}_{\mathcal{G}})$. Additional information about the graph architecture are provided in Appendix D.

5.2 SPOT-GUIDED SUPERVISION

Given an encoder $E(x)$ (DINOv2 or graph encoders defined previously), with x either the image content, semantic or spatial scene graph, S3PE aims to encourage the learned embedding space to preserve the relative spatial similarity induced by SPOT thanks to a triplet loss function (see Figure 3). Triplets are constructed using SPOT as a spatial supervision signal. For each anchor graph $\mathcal{G}_a$, candidate graphs are ranked according to SPOT, from the most to the least spatially similar. Before training, a positive candidate pool $\mathcal{P}_a$ is formed from the closest graphs in this ranking, while a negative candidate pool $\mathcal{N}_a$ is formed from a more distant region of the ranking. During training, for each $\mathcal{G}_a$, a positive graph $\mathcal{G}_p$ and a negative graph $\mathcal{G}_n$ are randomly sampled from $\mathcal{P}_a$ and $\mathcal{N}_a$ respectively. Let $\mathbf{z}_a$, $\mathbf{z}_p$, and $\mathbf{z}_n$ denote their respective embeddings from the encoder $E(\cdot)$, with pairwise dissimilarity measured by cosine distance. The model is optimized using a margin-based triplet loss:

$$\mathcal{L}_{\text{triplet}} = \max\left(0,\, d_{\cos}(\mathbf{z}_a, \mathbf{z}_p) - d_{\cos}(\mathbf{z}_a, \mathbf{z}_n) + m\right), \tag{11}$$

where $m > 0$ denotes the margin. This objective encourages data samples with spatially similar graphs according to SPOT to be closer in the embedding space than spatially dissimilar graphs.

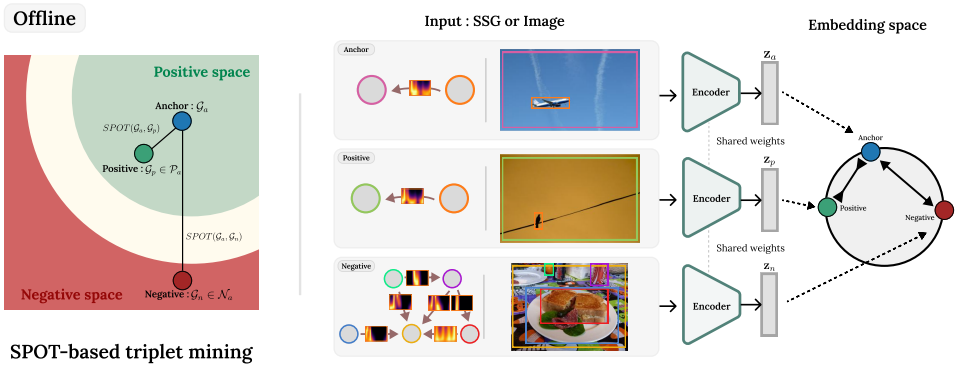

Figure 3: **Overview of the S3PE training framework.** For each anchor graph $\mathcal{G}_a$, SPOT is used to define a positive candidate pool $\mathcal{P}_a$ of spatially similar graphs and a negative candidate pool $\mathcal{N}_a$ of more spatially dissimilar graphs. A positive graph $\mathcal{G}_p \in \mathcal{P}_a$ and a negative graph $\mathcal{G}_n \in \mathcal{N}_a$ are sampled and represented either by their spatial scene graphs or associated images. The three inputs are processed by an encoder with shared weights to obtain the final embeddings $\mathbf{z}_a$, $\mathbf{z}_p$, and $\mathbf{z}_n$. The triplet objective encourages the anchor embedding to be closer to the positive than to the negative in the learned embedding space.

6 Experimental Study

We indirectly evaluate the interest of SPOT through the S3PE encoders with two spatially-aware downstream tasks: image retrieval, which aims to highlight the ability of the models to capture spatial similarity instead of semantic similarity and Visual Question Answering (VQA), to showcase that the image representations of S3PE can be leveraged for scene understanding. An additional downstream task, i.e. scene recognition in remote sensing imagery, is provided in Appendix H to demonstrate the cross-domain generalizability of S3PE.

Training dataset and preprocessing. We conduct our training on the PSG (Yang et al., 2022) dataset, which contains 48,749 scene graphs from natural images drawn from the overlap between COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2016). The dataset covers 133 object categories (e.g. *bench*, *car*) and 56 predicates categories (e.g. *beside*, *walking on*). We construct a fixed random 80/10/10 train/validation/test split.

For each image–scene graph pair, we remove nodes and associated edges corresponding to small objects whose bounding boxes account for less than 3% of the image size, to avoid modeling spatial configuration from barely visible objects in the image. COCO segmentation masks (Lin et al., 2014) are required to compute the Force Banners required for computing SPOT. Due to the computational complexity of SPOT (see Appendix B), the computation of the whole SPOT matrix is expensive. We provide in Appendix C an approximation algorithm retaining the similarity ordering with a Lipschitz embedding.

Baselines. We compare S3PE against representations obtained from both images and semantic scene graphs. For image-based representations, we evaluate CLIP-ViT-B/16 and CLIP-ViT-L/14 (Radford et al., 2021), DINOv2 (Oquab et al., 2023), and RelTR (Cong et al., 2022). For semantic scene graph-based retrieval, we compare against IRSGS (Yoon et al., 2021). Implementation details are provided in Appendix D.

6.1 Content-based image retrieval

We evaluate the learned representations through content-based image retrieval on the PSG (Yang et al., 2022) test set. To characterize how the different representation spaces organize visual similarity, we evaluate retrieval along two complementary dimensions: semantic similarity and spatial similarity. This allows us to assess whether SPOT-guided supervision increases sensitivity to the spatial organization of the scene, while also measuring how this spatial specialization affects the semantic information captured by the embeddings.

Table 1: **Semantic and spatial retrieval performance on the PSG (Yang et al., 2022) test set.** Retrieval rankings are evaluated according to semantic relevance derived from SPICE and spatial relevance derived from SPOT. These two criteria characterize complementary properties of the learned representation spaces. Values in gray parentheses indicate the absolute gain over the corresponding random baseline. Bold values indicate the best performance within each method family.

		Semantic relevance SPICE-based		**Spatial relevance** SPOT-based	
Model	Input	NDCG@25 ↑	MP@100 ↑	NDCG@25 ↑	MP@100 ↑
Reference rankings					
Random	–	18.04	2.09	62.60	2.09
SPOT ranking	–	31.66	7.44	100.00	100.00
SPICE ranking	–	100.00	100.00	72.64	7.44
Image-based methods					
CLIP-ViT-B/16	Image	62.33 (+44.29)	29.21 (+27.12)	69.09 (+6.49)	5.79 (+3.70)
CLIP-ViT-L/14	Image	62.35 (+44.31)	28.95 (+26.86)	68.96 (+6.36)	5.69 (+3.60)
DINOv2	Image	28.68 (+10.64)	5.45 (+3.36)	66.83 (+4.23)	3.46 (+1.37)
RelTR	Image	**64.08** (+46.04)	**32.35** (+30.26)	70.82 (+8.22)	6.77 (+4.68)
S3PE (Ours)	Image	35.18 (+17.14)	8.86 (+6.77)	**73.78** (+11.18)	**10.68** (+8.59)
Scene-graph methods					
IRSGS-GIN	SG	**74.35** (+56.31)	**40.34** (+38.25)	73.19 (+10.59)	6.26 (+4.17)
IRSGS-GCN	SG	73.82 (+55.78)	40.32 (+38.23)	72.93 (+10.33)	6.23 (+4.14)
S3PE (Ours)	SG	44.02 (+25.98)	12.22 (+10.13)	**78.30** (+15.70)	**12.39** (+10.30)
Spatial scene-graph method					
S3PE (Ours)	SSG	**25.86** (+7.82)	**5.11** (+3.02)	**87.27** (+24.67)	**31.29** (+29.20)

Semantic similarity. We adapt the SPICE metric (Anderson et al., 2016) as a semantic similarity ($\mathrm{rel}_{\mathrm{sem}}$), in which each graph is transformed into sets of 1-tuples (object type) and 3-tuples (subject-predicate-object). For a given query graph $\mathcal{G}_q$ and a candidate graph $\mathcal{G}_i$, we resolve one-to-one set bipartite matching. The relevance score is computed from the F-Score between the two sets of tuples such that $\mathrm{rel}_{\mathrm{sem}}(q,i) = \mathrm{F1}_{\mathrm{SPICE}}(\mathcal{G}_{\mathrm{q}}, \mathcal{G}_{\mathrm{i}})$.

Spatial similarity. We leverage SPOT to define spatial relevance between a query graph $\mathcal{G}_q$ and a candidate graph $\mathcal{G}_i$. Since lower SPOT scores indicate greater spatial similarity, we convert them into a relevance score using an exponential decay:

$$\mathrm{rel}_{\mathrm{sp}}(q,i) = \exp\left(-\frac{\mathrm{SPOT}(\mathcal{G}_q, \mathcal{G}_i)}{\tau}\right), \qquad (12)$$

where $\tau > 0$ controls the scale of the relevance decay and is fixed using the median heuristic, i.e., the median pairwise SPOT score computed on a subset of the training split. The resulting value is then kept fixed for all methods evaluated on the test set.

Retrieval metrics. Given either $\mathrm{rel}_{\mathrm{sem}}$ or $\mathrm{rel}_{\mathrm{sp}}$, we leverage the Mean Precision (MP@k) metric and the NDCG metric to measure the ability of the models to retrieve relevant images or graphs.

Let π_q denote the ranking induced by decreasing cosine similarity between model embeddings, such that $\pi_q(r)$ is the candidate retrieved at rank r. We define the normalized discounted cumulative gain as

$$\mathrm{DCG}\,@k(q) = \sum_{r=1}^{k} \frac{\mathrm{rel}(q, \pi_q(r))}{\log_2(r+1)}, \qquad \mathrm{NDCG}\,@k = \frac{\mathrm{DCG}\,@k}{\mathrm{IDCG}\,@k}. \qquad (13)$$

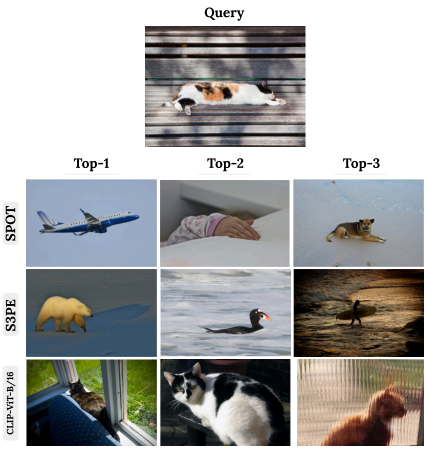

Figure 4: Top-3 retrievals for a query image using the SPOT ground truth ranking, S3PE, and CLIP-ViT-B/16, illustrating the contrast between spatial and semantic notions of similarity.

IDCG $@k(q)$ denotes the ideal ranking π_q^* obtained by sorting the candidate graphs by decreasing SPOT or SPICE-based relevance score.

Results. Experimental results are provided in Table 1. CLIP-ViT-L/14 reaches 62.35% semantic similarity and 68.96% spatial similarity for the NDCG metric. However, compared to the random rankings, CLIP-ViT-L/14 gets +44.31% and +6.36% absolute improvements respectively, highlighting that although spatial similarity is higher, models are not necessarily better at spatial similarity than semantic. We emphasize that spatial relevance values are more fuzzy than semantic relevance values, leading to higher spatial similarity baselines compared to semantic similarity.

Models leveraging scene graphs reach higher similarities scores as scene graphs are underlying structures of images that induce better assessment of either similarity dimension.

We consider the results of S3PE-SSG as upper-bound scores of the S3PE method as it leverages the same spatial information used by SPOT. Whether images or scene graphs are used as inputs to the model, S3PE outperforms all other methods on spatial similarity. On the other hand, S3PE-based models get lower semantic similarity scores than their corresponding baseline models, which is an inherent consequence of the spatial similarity training. Specifically, Figure 4 illustrates that given a query image containing a bench and a cat, baseline models like CLIP-ViT-B/16 retrieve images based on the semantic content, while SPOT reference ranking and S3PE-Image retrieve images that do not comprise similar semantic contents but retain the spatial configurations of the query image. Appendices F and G provide additional quantitative and qualitative retrieval results of SPOT and S3PE-Image on synthetic and natural image datasets.

6.2 Visual Question Answering

We leverage the GQA (Hudson & Manning, 2019) dataset to perform a spatially-aware VQA downstream task. We follow the same experimental protocol of Servant et al. (2025) in which we retain the image-question pairs whose questions are *Verify Relation* types (leading to a 37,743/4,194/5,975 train/val/test split). S3PE is compared to C-SIP (Servant et al., 2025), a method focused on training spatially-aware image representations with a contrastive loss function. C-SIP corresponds to the spatial-only encoder, while C-SIP++ is the concatenation of semantic and spatial models provided in Servant et al. (2025). The pre-trained baseline models and S3PE-Image are frozen and their image embeddings are concatenated with the BERT (Devlin et al., 2019) embeddings of the question. Experimental results are provided in Table 2, and highlight that S3PE captures a spatially- and structure-aware representation of the image contents for related downstream tasks.

Table 2: **Accuracy on the VQA downstream task.** The Text-only model corresponds to BERT embeddings of the question without the image input.

Model	Accuracy
Random	50.00
Text-only	54.64
C-SIP	72.06
C-SIP++	77.10
S3PE-Image (Ours)	**80.71**

7 Conclusion

In this paper, we addressed the limitations of existing similarity metrics that do not capture fine-grained spatial content of the objects and the overall structure of images for visual similarity retrieval. To this end, we introduced the Spatial Scene Graph (SSG), a scene graph representation that encodes fine-grained spatial information from an image, disentangled from semantic content. Building on this representation, we proposed SPOT, a quantitative criterion for similarity between spatial scene graphs framed as an optimal transport problem, that indirectly captures the spatial similarity of objects in images. We experimentally demonstrated that SPOT can serve as an evaluation criterion for spatial similarity in retrieval tasks, and as a training signal for SPOT-Supervised SPatial Embeddings (S3PE), a self-supervised representation learning method that mines positive and negative pairs according to SPOT-based spatial similarity.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *ECCV*, volume 9909 of *LNCS*, pp. 382–398, 2016.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *ICLR*, 2022.

David Abou Chacra and John S. Zelek. The topology and language of relationships in the visual genome dataset. In *CVPR Workshops*, pp. 4859–4867. IEEE, 2022.

Nikolaos Chaidos, Angeliki Dimitriou, Maria Lymperaiou, and Giorgos Stamou. SCENIR: visual semantic clarity through unsupervised scene graph retrieval. *CoRR*, abs/2505.15867, 2025.

Lichang Chen, Guosheng Lin, Shijie Wang, and Qingyao Wu. Graph edit distance reward: Learning to edit scene graph. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *ECCV*, volume 12364 of *LNCS*, pp. 539–554. Springer, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020b.

Michaël Clément, Camille Kurtz, and Laurent Wendling. Bags of spatial relations and shapes features for structural object description. In *ICPR*, pp. 1994–1999. IEEE, 2016.

Michaël Clément, Camille Kurtz, and Laurent Wendling. Learning spatial relations and shapes for structural object description and scene recognition. *Pattern Recognition*, 84:197–210, 2018.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE TPAMI*, 45:11169–11183, 2022.

Yuren Cong, Wentong Liao, Bodo Rosenhahn, and Michael Ying Yang. SPAN: learning similarity between scene graphs and images with transformers. *IEEE TPAMI*, 48(3):2886–2901, 2026.

Robin Deléarde, Camille Kurtz, and Laurent Wendling. Description and recognition of complex spatial configurations of object pairs with force banner 2d features. *Pattern Recognition*, 123:108410, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186. ACL, 2019.

John Freeman. The modelling of spatial relations. *Computer graphics and image processing*, 4(2):156–171, 1975.

Andreas René Geist, Jonas Frey, Mikel Zhobro, Anna Levina, and Georg Martius. Learning with 3d rotations, a hitchhiker's guide to SO(3). In *ICML*, 2024.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pp. 3668–3678, 2015.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, pp. 9161–9175. ACL, 2023.

Keisuke Kawano, Satoshi Koide, Hiroaki Shiokawa, and Toshiyuki Amagasa. Multi-dimensional fused gromov wasserstein discrepancy for edge-attributed graphs. *IEICE Trans. Inf. Syst.*, 107(5):683–693, 2024.

Kurt Koffka. Principles of gestalt psychology. 1935.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2016.

Sorachi Kurita, Satoshi Oyama, and Itsuki Noda. Predicate classification using optimal transport loss in scene graph generation. *CoRR*, abs/2309.10430, 2023.

Stephen C Levinson. Frames of reference and molyneux's question: Crosslinguistic evidence. 1996.

Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, Bo Dang, Yongjun Zhang, Yi Yu, and Junchi Yan. STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE TPAMI*, 47(3): 1832–1849, 2025.

Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. Natural language guided visual relationship detection. In *CVPR*, pp. 444–453, 2019.

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 *ECCV*, pp. 740–755, 2014.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: can vision-language foundation models reason compositionally? In *CVPR*, pp. 10910–10921. IEEE, 2023.

Paridhi Maheshwari, Ritwick Chaudhry, and Vishwa Vinay. Scene graph embeddings using relative similarity supervision. In *AAAI*, pp. 2328–2336, 2021.

Pascal Matsakis and Laurent Wendling. A new way to represent the relative position between areal objects. *IEEE TPAMI*, 21(7):634–643, 1999.

Pascal Matsakis, Mohammad Naeem, and Farhad Rahbarnia. Introducing the ϕ-descriptor-a most versatile relative position descriptor. In *ICPRAM*, volume 2, pp. 87–98, 2015.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Rabin, Delon, and Gousseau. Transportation distances on the circle. *Journal of Mathematical Imaging and Vision*, 41(1):147–167, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pp. 8748–8763, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pp. 3980–3990. ACL, 2019.

Mateus Riva, Pietro Gori, Florian Yger, and Isabelle Bloch. Is the u-net directional-relationship aware? In *ICIP*, pp. 3391–3395, 2022.

Logan Servant, Michaël Clément, Laurent Wendling, and Camille Kurtz. Contrastive learning of image representations guided by spatial relations. In *WACV*, pp. 2124–2133, 2025.

Logan Servant, Michaël Clément, Laurent Wendling, and Camille Kurtz. Spirl: Spatially-aware image representation learning under the supervision of relative position descriptors. *Pattern Recognition*, 170:112013, 2026.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pp. 5228–5238, 2022.

Barbara Tversky and Bridgette Martin Hard. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, 110(1):124–129, 2009.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *CoRR*, abs/1811.02834, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. 2009.

Yang Wang, Huilin Peng, Yiwei Xiong, and Haitao Song. Spatial relationship recognition via heterogeneous representation: A review. *Neurocomputing*, 533:116–140, 2023.

Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, volume 13687, pp. 178–196, 2022.

Sangwoong Yoon, Woo-Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *AAAI*, pp. 10718–10726, 2021.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023.

APPENDIX

A FORCE BANNERS MATHEMATICAL PROPERTIES

The Force Banner is an extension of the Force Histogram (Matsakis & Wendling, 1999), such that the whole 2D representation of the Force Banner between objects A and B is computed as:

$$\mathbf{FB}^{AB} = \begin{pmatrix} F^{AB}(r_{min},\theta_0) & \cdots & F^{AB}(r_{min},\theta_{j\delta_\theta}) & \cdots & F^{AB}(r_{min},\theta_{2\pi-\delta_\theta}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ F^{AB}(r_{min}+i\delta_r,\theta_0) & \cdots & F^{AB}(r_{min}+i\delta_r,\theta_{j\delta_\theta}) & \cdots & F^{AB}(r_{min}+i\delta_r,\theta_{2\pi-\delta_\theta}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ F^{AB}(r_{max},\theta_0) & \cdots & F^{AB}(r_{max},\theta_{j\delta_\theta}) & \cdots & F^{AB}(r_{max},\theta_{2\pi-\delta_\theta}) \end{pmatrix} \tag{14}$$

The Force Banner exhibits well-defined behaviour under common geometric transformations of images (e.g. translation, rotation) defined as follows:

Translation. The Force Banner remains invariant to any translation applied to the whole image. Given a translation t_u by a vector u:

$$FB[t_u(AB)] = FB(AB) \tag{15}$$

Rotation. Under a rotation by an angle α, the Force Banner is deduced as:

$$FB[rot_\alpha(AB)](\cdot,\theta) = FB[AB](\cdot,\theta-\alpha) \tag{16}$$

Scaling. Applying a scaling factor k with a homothety h on an image results in the related Force Banner to be computed as:

$$FB[h_k(AB)](r,\cdot) = k^{3-r}FB[AB](r,\cdot) \tag{17}$$

Reflection. Given a reflection transformation ref at an angle α of an image, the Force Banner can be computed as:

$$FB[\mathrm{ref}_\alpha(AB)](\cdot,\theta) = FB[AB](\cdot,2\alpha-\theta) \tag{18}$$

These transformation properties enable fine-grained assessments of spatial similarity between object pairs, and are therefore well suited to the computation of SPOT.

B EFFICIENT COMPUTATION AND OPTIMIZATION OF SPOT

A direct implementation of the relational term would require storing a fourth-order tensor with $\mathcal{O}(n^2m^2)$ entries. In practice, we use an algebraically equivalent formulation that only requires explicit comparisons between existing directed edges, reducing the relational memory requirement to $\mathcal{O}(|\mathcal{E}_1||\mathcal{E}_2|)$. We optimize the resulting objective using the Frank–Wolfe algorithm.

Let A_{ik} and B_{jl} denote the directed edge-presence indicators:

$$A_{ik} = \begin{cases} 1, & (i,k) \in \mathcal{E}_1, \\ 0, & \text{otherwise}, \end{cases} \qquad B_{jl} = \begin{cases} 1, & (j,l) \in \mathcal{E}_2, \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

The relational cost in equation 7 can then be equivalently written as

$$L_{ikjl} = c_{\mathrm{miss}}A_{ik} + c_{\mathrm{miss}}B_{jl} + A_{ik}B_{jl}\left[W_1(\mathbf{x}_{ik},\mathbf{x}_{jl}) - 2c_{\mathrm{miss}}\right]. \tag{20}$$

Substituting equation 20 into the definition of $\mathcal{C}_{\mathrm{rel}}(\mathbf{T})$ in equation 8 gives

$$\begin{aligned}\mathcal{C}_{\text{rel}}(\mathbf{T}) = c_{\text{miss}} &\sum_{(i,k)\in\mathcal{E}_1}\left[\left(\sum_j T_{ij}\right)\left(\sum_l T_{kl}\right)\right] \\ &+ c_{\text{miss}} \sum_{(j,l)\in\mathcal{E}_2}\left[\left(\sum_i T_{ij}\right)\left(\sum_k T_{kl}\right)\right] \\ &+ \sum_{\substack{(i,k)\in\mathcal{E}_1\\(j,l)\in\mathcal{E}_2}} \left[W_1(\mathbf{x}_{ik},\mathbf{x}_{jl}) - 2c_{\text{miss}}\right] T_{ij}T_{kl}. \end{aligned} \tag{21}$$

This reformulation is algebraically identical to the relational term in equation 8. The first two terms aggregate missing-edge contributions through the row and column marginals of the transport plan, while the last term requires explicit comparisons only between existing directed edges in $\mathcal{E}_1 \times \mathcal{E}_2$. Consequently, the relational representation requires $\mathcal{O}(|\mathcal{E}_1||\mathcal{E}_2|)$ memory instead of $\mathcal{O}(n^2m^2)$.

C Approximation of SPOT

The training set of PSG contains $n = 38,258$ samples. The computation of the SPOT distance matrix would thus require $\binom{n}{2} \approx 8 \times 10^8$ optimal transport computations that require computing node- and edge-costs as well as optimizing with the Frank–Wolfe algorithm. To reduce the computational burden, we employ a landmark-based approximation method, which constructs a low-rank approximation of the full matrix using a subset of $m \ll n$ landmark samples.

Let $K \in \mathbb{R}^{n\times n}$ denote the full SPOT distance matrix, and $S = \{s_1, \ldots, s_m\} \subset \{1, \ldots, n\}$ be a set of m landmark indices chosen by random sampling. We define

$$C = K_S \in \mathbb{R}^{n\times m}, \tag{22}$$

the submatrix of SPOT distances computed with the method from Section 4. Each of the n samples can then be represented by a landmark embedding (a Lipschitz embedding with singleton reference sets), i.e. its vector of distances to the m landmarks such that $v_i = C_{i,\cdot}$. Given two non-landmark indices $i \notin S$ and $j \notin S$, we then approximate the SPOT distance as:

$$\widehat{\text{SPOT}}(i,j) = \|v_i - v_j\|_2 = \sqrt{\sum_{l=1}^{m} \left(C_{i,l} - C_{j,l}\right)^2}. \tag{23}$$

The intuition behind this approach is that samples that are close in the SPOT space have similar distances to every landmarks.

We leveraged this approach to approximate the distance matrix of the training set of PSG, but computed the full distance matrices of both validation and test sets. With $m = 1000$, this approach enabled a $n \times m \approx 4 \times 10^7$ OT computations, leading to a 20-fold reduction of the total number of optimal transport computations.

From the set of m landmarks, we consider $C_m = K_{SS} \in \mathbb{R}^{m\times m}$ the subset of the matrix of SPOT solves, and $\hat{C_m} \in \mathbb{R}^{m\times m}$ the matrix of approximations obtained with the abovementioned approach, and compare them. The Pearson correlation coefficient between the two matrices is computed as $\rho = 0.94$.

We further computed SPOT solves and their approximations of 10 random samples $s \notin S$. The average correlation across samples is $\rho = 0.92(\pm 0.02)$. We additionally report the top-k recall, as exact values of SPOT are not required for training S3PE, as the ranking matters more for positive and negative triplet mining. We set $k = 100$ in S3PE for positive triplet mining (See Appendix D). Following this setup, we obtain a mean top-k recall of $r = 0.93, (\pm 0.02)$ across the 10 randomly selected samples. This result indicates that using approximated SPOT values instead of performing numerous exact optimal transport computations substantially reduces the computational cost while largely preserving the neighborhood rankings used for S3PE triplet mining.

D IMPLEMENTATION DETAILS

D.1 GNN ARCHITECTURE

Node and edge representations are first projected into a latent space with a feed-forward network. To account for the periodicity of the Force Banner, edge representations in spatial scene graphs used as inputs to the GNN are instead projected with 2D convolutions with circular padding.

Node representations are then updated with GATv2 layers by aggregating information from their neighboring nodes $\mathcal{N}_i$ and their associated edge representations, such that:

$$\mathbf{h}_i^{(k+1)} = \mathrm{GATv2}\left(\mathbf{h}_i^{(k)}, \{\mathbf{h}_j^{(k)}, \mathbf{r}_{ij} \mid j \in \mathcal{N}_i\}\right). \tag{24}$$

To obtain a graph-level representation after $K = 4$ layers, we aggregate node and edge representations independently using mean pooling and apply a readout network:

$$\mathbf{h}_{\mathcal{G}} = \mathrm{READOUT}\left(\mathrm{MEAN}\{\mathbf{h}_i^{(K)}\}_{i\in\mathcal{V}}, \mathrm{MEAN}\{\mathbf{r}_{ij}\}_{e_{ij}\in\mathcal{E}}\right). \tag{25}$$

D.2 S3PE TRAINING DETAILS

Triplet mining hyperparameters. On PSG, the positive pool $\mathcal{P}_a$ contains the 100 nearest neighbors according to SPOT, while the negative pool $\mathcal{N}_a$ contains 100 graphs sampled between ranks 1000 and 5000. Preliminary experiments using the most distant graphs as negatives led the triplet objective to saturate rapidly, as these negatives were too easily separated from the positives. We therefore sample negatives from an intermediate ranking range to provide more informative training constraints while maintaining a clear spatial separation from the positive pool. Positive and negative samples are resampled from their respective pools at each training epoch. We use a cosine triplet loss with margin $m = 0.2$.

Training hyperparameters. All S3PE models are trained for 20 epochs using the SPOT-guided cosine triplet objective described in the main paper, with a margin $m = 0.2$. The loss is applied to the 128 dimensional projected representations.

For the scene graph models, S3PE-SG and S3PE-SSG, we use AdamW with a learning rate of 10^{-4} and a weight decay of 10^{-4}. Models are trained with batches of 128 triplets. Both GNNs use a hidden dimension of 128, four GATv2 attention heads, a dropout rate of 0.1, and a 128 dimensional projection space.

For S3PE-Image, we fine-tune a DINOv2 ViT-B/14 backbone initialized from pretrained weights. Images are resized to 224×224 while preserving their aspect ratio. We use a batch size of 32 triplets and AdamW with separate learning rates of 10^{-5} for the image backbone and 10^{-4} for the projection head, with a weight decay of 0.05. The projection head contains a 512 dimensional hidden layer followed by the 128 dimensional output representation, with a dropout rate of 0.1.

E EFFECT OF BALANCING NODE AND RELATIONAL COSTS

We qualitatively illustrate the effect of the weighting parameter α, which controls the contribution of node-level geometry and relational information in SPOT. Figure 5 shows retrieval examples for $\alpha \in \{0, 0.5, 1\}$.

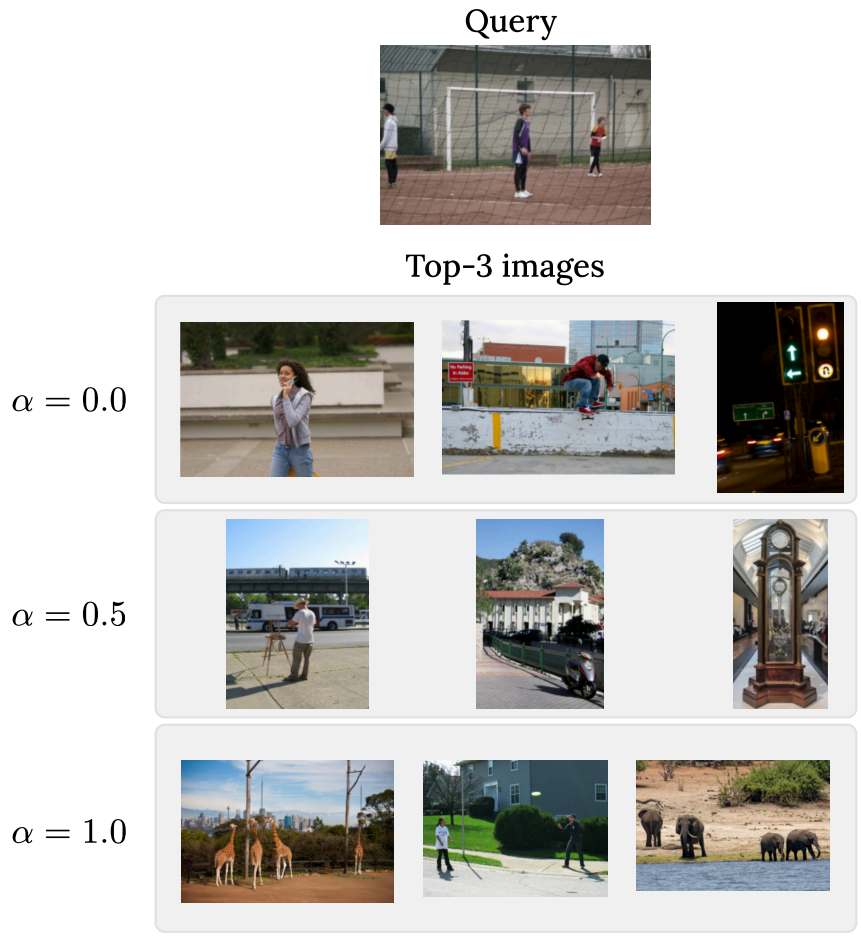

Figure 5: **Qualitative effect of the SPOT weighting parameter α.** Retrieval results are shown for $\alpha \in \{0, 0.5, 1\}$. With $\alpha = 1$, SPOT considers only node-level geometry, whereas $\alpha = 0$ relies exclusively on relational information encoded by Force Banners. The balanced setting $\alpha = 0.5$ jointly accounts for absolute object geometry and relative spatial configurations.

F CONTROLLED SPOT EXPERIMENT

We highlight in this section the ability of SPOT to capture similar scenes. We first perform this experimental validation on a synthetic dataset, and then demonstrate its application on natural image datasets.

F.1 SYNTHETIC DATASET

We leverage SharvitSR (Servant et al., 2026), a synthetic dataset of 44,100 pairs of objects in images annotated with 8 directional spatial relations (e.g. *left*, *Above-right*). Figure 6 presents query images with different spatial configurations of the pairs of objects and the five closest images in the SPOT space. The retrieved images demonstrate that (1) the absolute spatial information (i.e. bounding box coordinates) is preserved in the SPOT reference ranking, and (2) relative positions of objects are also preserved in the SPOT space, which enables a fine-grained spatial representation that cannot be accurately represented with semantic spatial relations. To illustrate this claim, we highlight that the query image of the last row is annotated with the *Under-Left* spatial relation, while the fifth retrieved image is annotated with the *Left* spatial relation, despite their spatial configuration being almost identical. This predicament would lead these two visually similar images farther apart if semantic similarity metrics are leveraged.

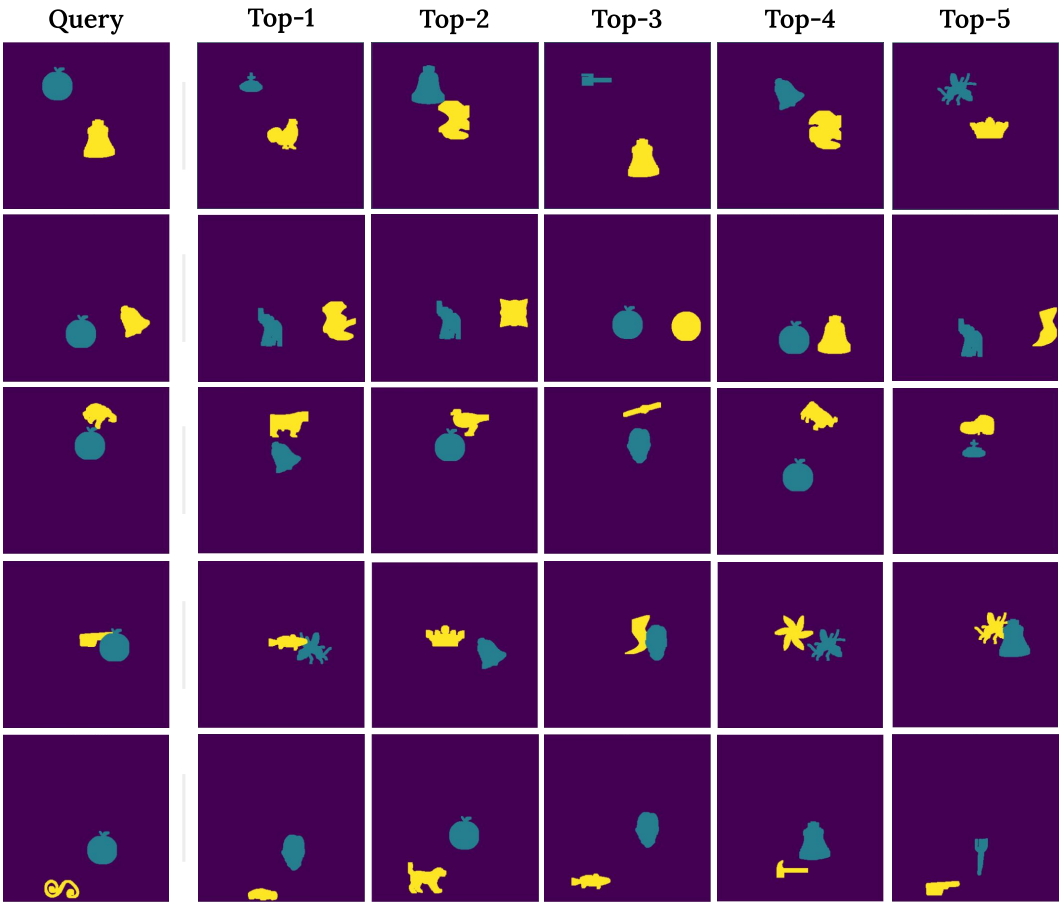

Figure 6: **Qualitative spatial retrieval with SPOT on the SharvitSR synthetic dataset.** Each row shows a query scene followed by its top-5 nearest scenes ranked by increasing SPOT distance. Each scene contains two binary objects with varying shapes and spatial configurations. The retrieved examples illustrate that SPOT primarily captures similarities in the relative spatial arrangement of the two objects despite variations in their shapes.

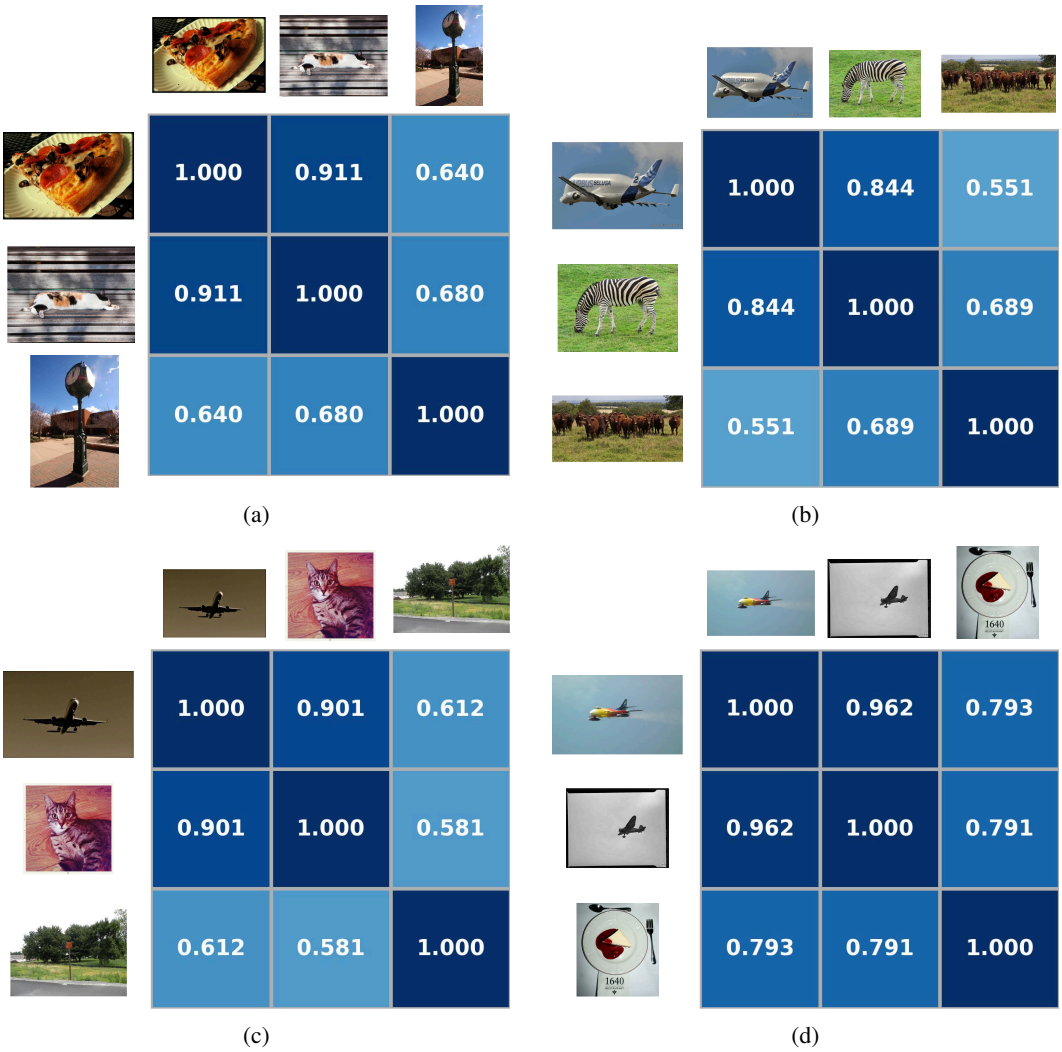

Figure 7: Examples of pairwise spatial similarity on natural images from the PSG dataset. Each panel shows three images and the corresponding pairwise similarity matrix. SPOT distances are converted into similarity scores as $1 - \mathrm{SPOT}(\mathcal{G}_i, \mathcal{G}_j)$, such that higher values indicate more similar spatial configurations. Diagonal entries correspond to self-similarity and are equal to 1.

F.2 Natural images dataset

We present in Figure 7 additional examples of the SPOT similarities drawn from the spatial scene graphs associated with each image. Each similarity matrix provides two visually similar images (according to SPOT values) with a third image that is spatially dissimilar from the other two.

G Additional retrieval results

G.1 Quantitative results

We provide in this section additional quantitative results of spatial similarity. Specifically, Figure 8 and Figure 9 display the performances of the baseline models and S3PE over the NDCG@k and MP@k metrics respectively over growing values of k.

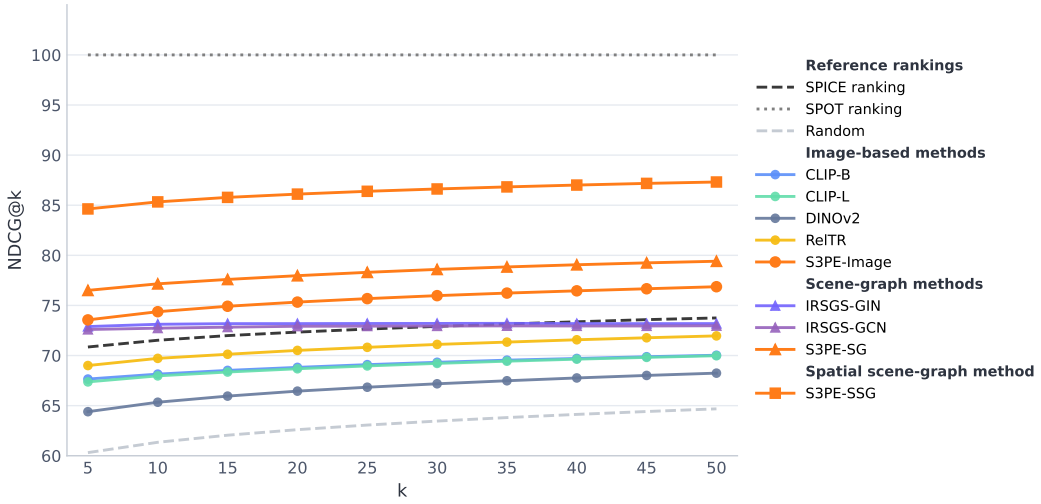

Figure 8: **Spatial retrieval performance on the PSG test set for different values of k.** NDCG@k is computed using SPOT-based spatial relevance, with higher values indicating stronger agreement between the embedding-based ranking and the reference ranking induced by SPOT. Results are reported for $k \in \{5, 10, \ldots, 50\}$.

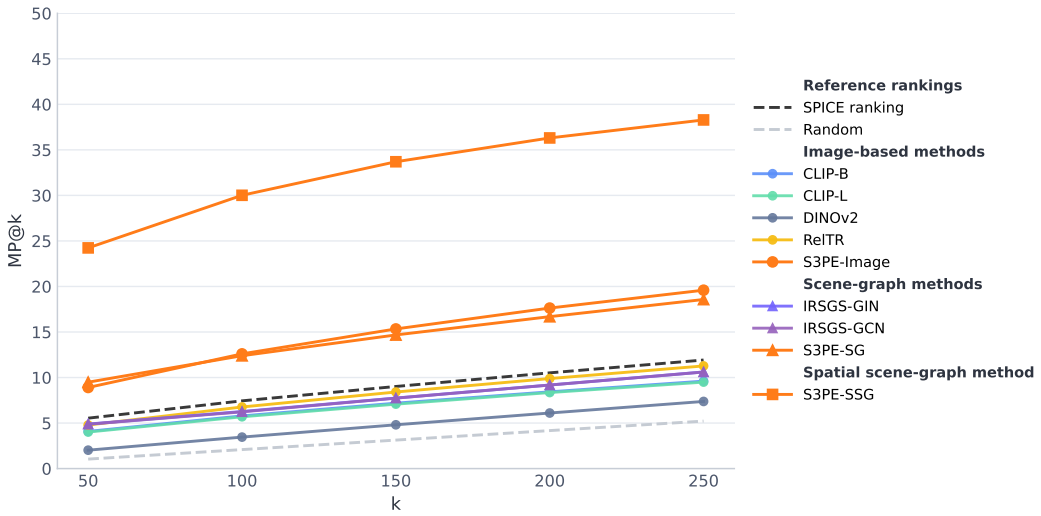

Figure 9: **Top-k spatial retrieval overlap on the PSG test set.** MP@k measures the mean proportion of the k retrieved samples that also belong to the k nearest neighbors according to SPOT. Higher values indicate a stronger overlap with the SPOT reference ranking. Results are reported for $k \in \{50, 100, 150, 200, 250\}$. The SPOT reference ranking is equal to 100%.

G.2 Qualitative results

We study the qualitative retrieval results of S3PE-Image on the test set of PSG (Yang et al., 2022) in Figure 10. For each query, the five closest images in the embedding space are retrieved. We selected a wide range of query images whose spatial scene graphs differ (e.g. from 2 objects for the topmost image to 24 objects for the second-to-last image).

Despite some retrieved images actually containing similar semantic content to the query image, most retrieved images display visual similarities that go beyond solely semantic content, and instead focus on the structural content of the scene and the spatial configuration of the objects.

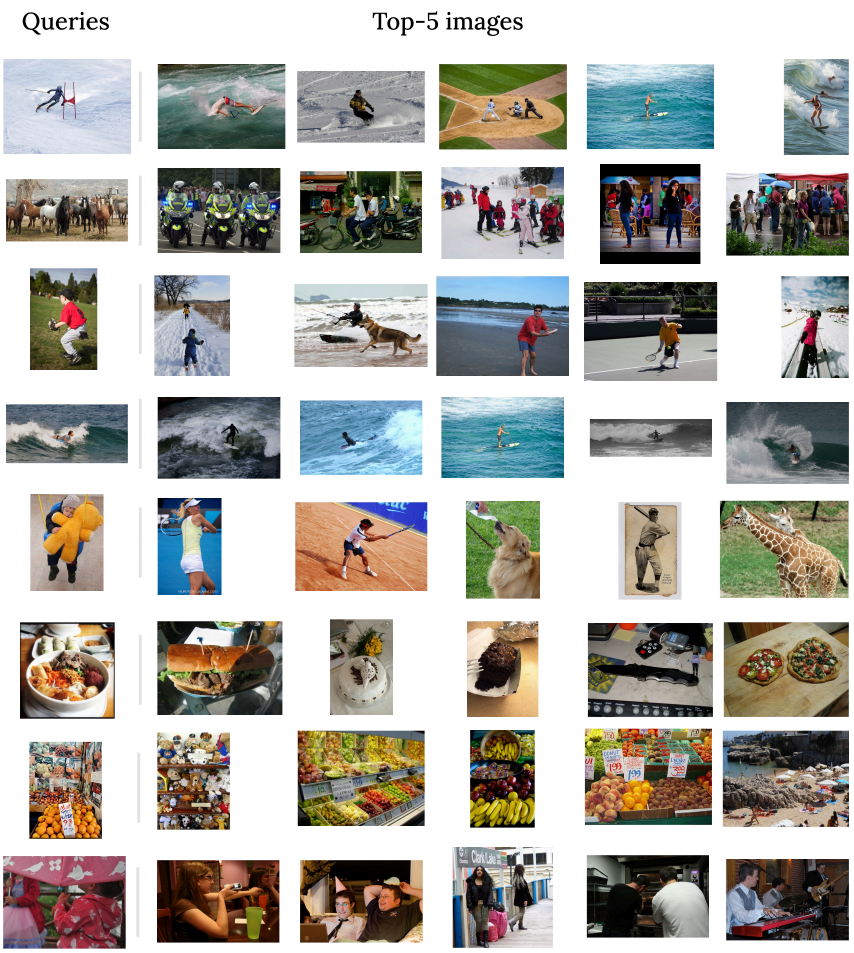

Figure 10: **Qualitative retrieval results with S3PE-Image on the PSG test set.** For each query image, the top-5 retrieved images are ranked according to decreasing cosine similarity between S3PE-Image embeddings. The examples illustrate that the learned image representations retrieve scenes with similar spatial organizations, even when their semantic contents differ.

H Cross-domain validation of SPOT and S3PE

To demonstrate the generalization capability of S3PE, we leverage STAR (Li et al., 2025), a remote sensing image dataset of 1,273 very-high resolution (VHR) images for scene graph generation. We leverage the STAR dataset to experimentally demonstrate the cross-domain generalization capabilities of SPOT, and S3PE with spatial scene graphs as inputs.

We demonstrate that the spatial configuration of objects can be leveraged to classify remote sensing scenes without direct access to visual data.

H.1 Dataset preprocessing

We leverage the ground truth scene graphs and use them as inputs to the S3PE-SSG pre-trained on PSG. STAR does not provide segmentation masks of the objects. However, oriented bounding boxes (OBB) coordinates are provided. Most remote sensing objects such as buildings, cars and runways can be approximated with oriented bounding boxes. We therefore leverage the oriented bounding boxes as the object masks for the computation of the Force Banners. Node inputs (i.e. bounding box coordinates) retain their representation (center coordinates $\mathbf{c} = (c_x, c_y)$ and size $\mathbf{s} = (w, h)$), without including the orientation θ value of the OBB. We leveraged axis-aligned width and height instead of oriented ones.

Table 3: **Per-class graph statistics of the STAR dataset used for cross-domain evaluation.** For each scene type, we report the number of images and the number of connected components (CC), objects, and relations per image. We leverage the S3PE-SSG embeddings of connected components: we report the average number of nodes (objects) and edges (relationships) of connected components for each class of STAR as well.

Scene Type	N images	CC/image	obj/image	rel/image	obj/CC	rel/CC
Toll Station	19	2.15	27.42	35.52	6.26	7.95
Construction Site	57	19.92	142.12	147.75	4.91	4.69
Wind Power Station	56	6.07	40.55	45.51	6.08	5.53
Service Area	86	8.86	100.68	197.67	8.08	12.03
Traffic Bridge	99	2.36	43.95	87.11	14.15	24.19
Port Scene	249	9.35	197.13	309.94	18.01	28.43
Airport	439	2.97	106.95	373.31	34.29	87.29
Sports	64	149.15	1383.89	1736.65	6.09	6.85
Nuclear Power Station	44	11.41	77.38	94.61	6.06	5.97
Thermal Power Station	98	7.84	55.08	53.09	5.67	5.44
Dam	62	1.90	29.11	40.69	12.72	13.72

Due to the image sizes (up to $24,000 \times 24,000$ pixel images), many objects and relationships are found within each image, leading to very large graphs and very small values of relative object sizes with respect to the image size.

We instead partition the image with respect to the connected components (CC) of the scene graphs. The cropped images (with respect to each CC) are instead used as the image size for the normalization of node inputs and process each connected component through S3PE-SSG, leading to various embeddings per image.

STAR (Li et al., 2025) does not include scene classes (e.g. ports, Thermal Power Station). To assign these labels, we leverage the semantic classes of objects to automatically annotate each image (e.g. Stadiums tend to be in Sports scenes, runways tend to be in Airport scenes). After automatic annotation of the whole dataset, we manually assessed the validity of the assigned classes of a portion of the dataset (approximately 50%).

Dataset statistics, alongside class-specific numbers of objects, edges, and connected components are provided in Table 3. STAR is highly imbalanced, with most images showing airports and port scenes, with few toll stations and nuclear power stations. We randomly sample 40 scene graphs for each class for fine-tuning and evaluate on the remaining scene graphs. Toll Stations are removed from the dataset due to the small number of image-graph samples.

H.2 Training protocol

For each subgraph $\mathcal{G}_c^I$ from an image I with $c \in CC$ a connected component, we consider the S3PE-SSG embedding $\mathbf{z}_c = E(\mathcal{G}_c^I)$ as a node representation of a complete graph $\mathcal{G}^I$ that we use as input to a graph neural network. Self-loops are added to $\mathcal{G}^I$ to account for the few images from STAR with only one connected component. We append to each node representation the centers and sizes of the connected component normalized with respect to the image size, and project them into a latent space with a feed-forward layer.

The model then performs message passing, followed by a mean pooling function and a feed-forward network with a classification head. We experiment with different message passing methods: Transformer convolutions, Graph Isomorphism Network (GIN), GATv2, and graph convolution networks (GCN), as well as baseline methods that do not include message passing and instead perform average or max pooling on the S3PE-SSG embeddings directly.

Each model is trained for 200 epochs with a learning rate of 10^{-4} with weight decay (10^{-4}) and a batch size of 16. We additionally include data augmentations such as node dropout (with a probability $p = 0.1$), and node perturbations (zero-mean gaussian noise with a scale of 5% of feature-wise standard deviation).

Table 4: **Classification scores on the test set of STAR (Li et al., 2025).** We report both the Accuracy metric and the macro-averaged F1 Score. Best scores are **bolded**. Second best scores are underlined.

Model	Acc.	F1 Score
Random	10.00	10.00
Avg. Pool. + FFN	44.08	49.44
Max. Pool. + FFN	46.60	49.56
Transformer Conv.	<u>58.75</u>	<u>59.95</u>
GIN	58.43	59.87
GATv2	57.81	59.36
GCN	**59.13**	**60.30**

H.3 RESULTS

As the dataset is imbalanced, we evaluate the different models with both the accuracy metric and the F1 Score. We emphasize that each model only leverages the spatial configuration of the different objects of the ground truth scene graphs, without access to the semantic classes of the objects. Experimental results are provided in Table 4 and demonstrate that the spatial configuration of objects in different remote sensing scenes enables an effective scene classification, especially for the GCN architecture that achieves 60.30% F1 Score. Baseline models (without message passing) still achieve strong classification results, but are outperformed by all message passing method.

To qualitatively inspect the learned embedding space of the best model (GCN), we provide a t-SNE representation of the embedding of the GCN for each data point of the test set of STAR in Figure 11. The t-SNE representation induces two main clusters of airports (green points). The visualization of data points in both clusters reveals distinct patterns of airports with respect to their orientation. The third cluster demonstrates that our method can identify scenes from the same location although not exactly identical (zoomed in and with a small translation).

However, cluster 4 demonstrates the limitations of our approach: as the model does not have access to the visual content, some scenes from different classes can have similar spatial configurations (i.e. dams and traffic bridges).

I AI USE STATEMENT

In this work, generative AI tools were used solely to improve the readability, grammar and wording of the manuscript. They were not used to develop the methodology, formulate mathematical claims, design experiments, implement methods, analyze data or interpret results. All AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of this work.

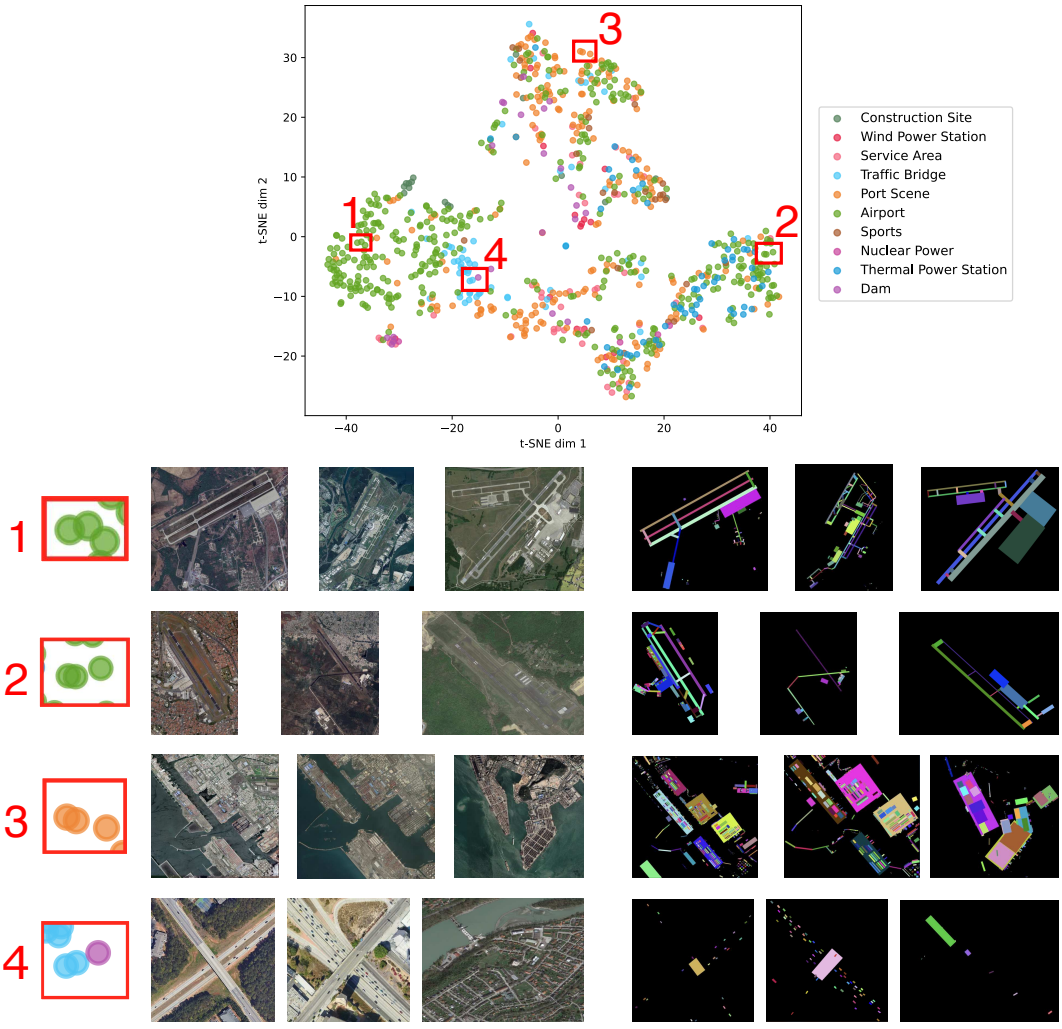

Figure 11: **t-SNE visualization of the best classification model (Graph Convolution Networks) on the test set of STAR.** We investigate four clusters of the t-SNE. Clusters 1 and 2 highlight differences found inside the Airport classes, Cluster 3 showcases similar Port scenes that are consistent with the translation and scaling properties of the Force Banner (Appendix A). Finally, Cluster 4 highlights the limitations of solely using spatial configurations of objects to classify remote sensing scenes: the spatial configuration of objects can be similar across scene types.