Scene-Level Sketch-Based Image Retrieval with Minimal Pairwise Supervision

Ce Ge*, Jingyu Wang*, Qi Qi†, Haifeng Sun, Tong Xu, Jianxin Liao

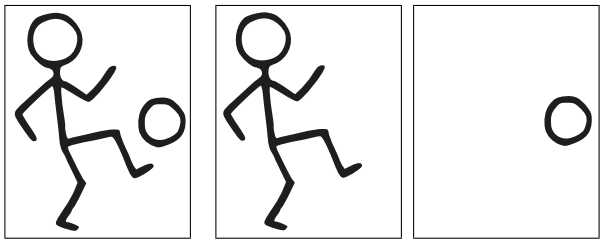
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing 100876, China
{nwlgc, wangjingyu, qiqi8266, hfsun}@bupt.edu.cn, xutong@ebupt.com, jxlbupt@gmail.com

Abstract

The sketch-based image retrieval (SBIR) task has long been researched at the instance level, where both query sketches and candidate images are assumed to contain only one dominant object. This strong assumption constrains its application, especially with the increasingly popular intelligent terminals and human-computer interaction technology. In this work, a more general scene-level SBIR task is explored, where sketches and images can both contain multiple object instances. The new general task is extremely challenging due to several factors: (i) scene-level SBIR inherently shares sketch-specific difficulties with instance-level SBIR (e.g., sparsity, abstractness, and diversity), (ii) the cross-modal similarity is measured between two partially aligned domains (i.e., not all objects in images are drawn in scene sketches), and (iii) besides instance-level visual similarity, a more complex multi-dimensional scene-level feature matching problem is imposed (including appearance, semantics, layout, etc.). Addressing these challenges, a novel Conditional Graph Autoencoder model is proposed to deal with scene-level sketch-images retrieval. More importantly, the model can be trained with only pairwise supervision, which distinguishes our study from others in that elaborate instance-level annotations (for example, bounding boxes) are no longer required. Extensive experiments confirm the ability of our model to robustly retrieve multiple related objects at the scene level and exhibit superior performance beyond strong competitors.

Introduction

Machine understanding of sketches has been widely studied in the fields of computer vision, computer graphics, and human-computer interaction (Herot 1976; Yu et al. 2017; Huang, Canny, and Nichols 2019). With the flourishing of touch screens (e.g., smartphones and tablets), sketching, as an efficient human-computer interaction, is becoming practical and ubiquitous. In particular, sketch-based image retrieval (SBIR) is a promising application that allows users to intuitively draw on screens and search for desired pictures without using clumsy text descriptions. SBIR is challenging because of the need to establish a correspondence between two distinct modalities of sparse line drawings and dense color pixels (Yu et al. 2016).



(a) Scene sketch (b) Instance sketches

Figure 1: Comparative illustration of scene sketch vs. instance sketches to represent a retrieval of "a soccer player is playing soccer".

Existing solutions mostly formalize SBIR as an embedding learning problem (Lei et al. 2019). The usual routine is to embed the query sketch and photos into a common vector space to compare similarities, and return the retrieval results through nearest neighbor search. Although recent approaches have achieved excellent performance, there is still a critical issue that they focused only on *instance-level* sketches (i.e., the query sketch is restricted to single object, such as a handbag, a bird, or a person), whereas *scene-level* SBIR (Zou et al. 2018) has been rarely investigated.

Instance-level SBIR is limited in application and is often ambiguous since it ignores contextual information and correlation between objects. Moreover, instance-level SBIR can only meet a small part of the demand considering the fact that scene-level images are more common in daily life. Fig. 1 shows an illustrative example of retrieving "a soccer player is playing soccer". Instance-level SBIR requires that the objects involved must be drawn one by one. Without seeing the ball, a retrieval system cannot infer much information from the stick figure alone. The figure may be a soccer player, a basketball player, or a dancer, and it may be kicking, running, dancing, etc. Similarly, it is difficult to extract useful information from just a sketchy circle. The ambiguity of instance SBIR makes it hard to accurately understand users' intentions.

Scene-level SBIR remedies these shortcomings and enables more powerful applications. On the one hand, it allows for joint retrieval of multiple objects where fine-grained pos-

*These authors contributed equally.
†Corresponding author.

tures and complex interactions that users intend to depict can be well represented (e.g., not just any person and any ball, but a soccer player kicking a soccer ball). On the other hand, it importantly progresses conventional sketch-based image retrieval paradigm from object-centric to true scene-level. This unique combination of sketch and scene can potentially trigger profound implications on the commercial adaptation of scene image retrieval—sketches as input (as opposed to images) offer a much needed degree of freedom when it comes to scene construction.

Scene-level SBIR is extremely challenging not only because of the shared difficulties with instance-level SBIR but also its unique scene-specific characteristics: (i) There is a domain gap between the two input modalities. Since scene sketches are rather abstract and iconic than natural images (e.g., caricatured and anthropomorphized objects), it is hard to learn cross-modal correspondence. (ii) People intend to depict only representative objects rather than drawing every detail exhaustively. The absence of some foreground objects and backgrounds makes the alignment even harder. (iii) A scene is not only determined by the object contained but also by their correlations (i.e., spatial layout). In essence, single instance-level sketches can be regarded as a narrow domain that has a limited and predictable variability in its appearance or category, whereas scene-level sketches lie in a much broader domain where complicated object interactions are further encoded (Smeulders et al. 2000).

We address these challenges by proposing a Conditional Graph Autoencoder (CGAE) model. By extracting object features and layout structure, scene sketches and natural images are abstracted as scene graphs. The model encoder is an augmented graph attention network with structural awareness and contextual condition. The variant decoder learns implicit semantic reconstruction to maintain distribution consistency. Last but not least, research on scene-oriented image retrieval usually requires detailed annotation of individual instances, which is a heavy burden in practical applications. We thus designed a minimally supervised learning paradigm to get rid of the dependence on instance-level category labels and bounding boxes. Using only image-level pairing information, our model can even beat recent solutions of utilizing more training supervision. The effectiveness of our design is confirmed by extensive experiments under multidimensional evaluation criteria.

The main contributions are summarized as follows:

- A novel pairwise supervised scene-level sketch-based image retrieval task is analyzed and studied.
- A Conditional Graph Autoencoder model is proposed to realize cross-modal scene-level retrieval by aggregating multi-level features including visual context, category semantics, and structural layout.
- Extensive experiments verified the superiority of our design and set a new state-of-the-art benchmark.

Related Work

Sketch-Based Image Retrieval

The research on sketch-based image retrieval can be traced back to the 1990s (Kato et al. 1992). Classical methods mainly focused on extracting various low-level visual descriptors of sketches and images (Kato et al. 1992). Photos were approximated as edge maps extracted by some form of edge detectors (Canny 1986). These methods are based on low-level or mid-level features, which lack robustness and adaptability to various drawing styles and different levels of abstraction.

The recent trend of sketch-based image retrieval is to learn deep visual semantic features (Shrivastava et al. 2011; Song et al. 2017) with the help of deep neural networks and object recognition techniques (Krizhevsky, Sutskever, and Hinton 2017). The features of input sketches and photos are jointly embedded into a common space to compare similarity. In order to alleviate the domain dependence problem, improved methods were proposed to enhance cross-category generalization (Song et al. 2018; Pang et al. 2019). Very recently, zero-shot SBIR draws increasing attention (Dutta and Akata 2019; Dey et al. 2019). While good performance has been achieved, it is noteworthy that all these methods are intentionally designed for single-instance sketches. Although some of them can be transformed to scene sketches, they are still agnostic to the structure of scenes. In addition to appearance characteristics (e.g., texture and poses), high-level semantic features (e.g., category labels and spatial layout) play a decisive role in measuring the similarity of scenes.

Scene-level Sketch Understanding

Early research on scene sketches mainly focused on industrial design and modeling (Donikian and Hégron 1992). Scene sketch-based 3D shape retrieval is another popular task (Yuan et al. 2019). Xu *et al.* (Xu et al. 2013) presented Sketch2Scene that allows co-retrieval and co-placement of 3D objects from scene sketches. Few studies have been devoted to scene-level sketch-based image retrieval. Dey *et al.* (Dey et al. 2018) implemented a sketch-related cross-modal multi-object retrieval system. However, they considered at most two objects and ignored the contextual environment. Jiang *et al.* (Jiang et al. 2017) investigated sketch-based aerial scene images retrieval, but their definition of "scene" is ambiguous as most images contain only one foreground object. Recently, Zou *et al.* (Zou et al. 2018) proposed the first large-scale scene sketch dataset and investigated segmentation and colorization tasks. We apply their proposed SketchyScene dataset to scene-level SBIR task. We noticed that the contemporaneous work SceneSketcher (Liu et al. 2020) and SceneSketcher-v2 (Liu et al. 2022) also employed GCN-based methods to achieve scene-level SBIR. The crucial difference is that our solution is the first attempt to get rid of laborious instance-level annotation and implement scene-level SBIR system using minimal image-level pairing information, while the SceneSketcher series are useless for this purpose.

Methodology

A novel Conditional Graph Autoencoder (CGAE) model is proposed in this work to learn scene-level sketch-based image retrieval as shown in Fig. 2. Scene-level sketch-based image retrieval is basically a cross-modal learning problem

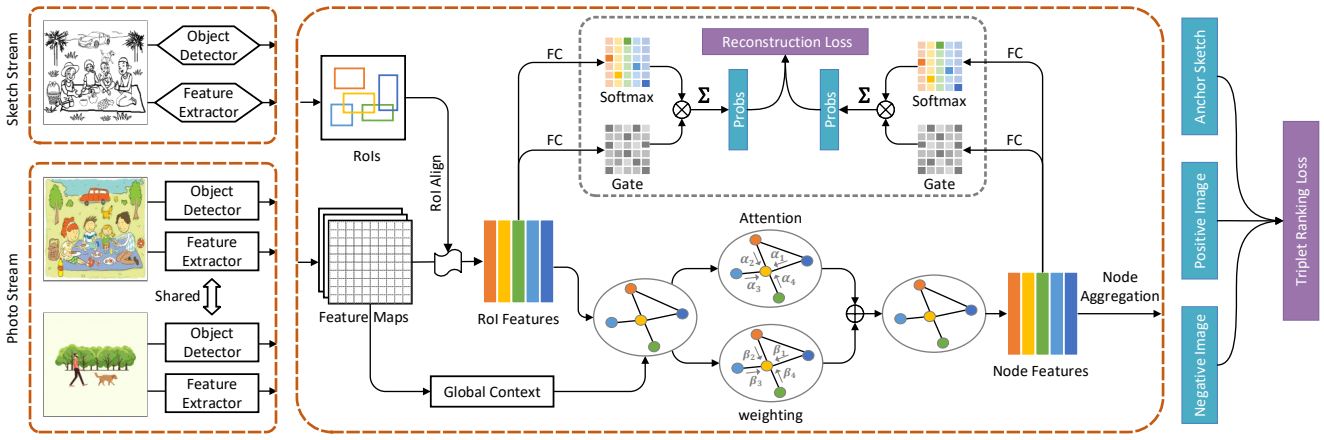


Figure 2: Overall architecture of the proposed Conditional Graph Autoencoder model.

between abstract line drawings and pixel-perfect natural images. Due to the inherent differences between the two input sources, heterogeneous feature extraction flows are needed. Given a triplet of anchor sketch, positive image, and negative image, their visual features and instance-level bounding boxes are separately extracted. These terms are then collectively fed into the shared graph model. The proposed CGAE model encodes scene content based on conditional graph attention structure and weakly supervised semantic reconstruction. The learned instance-level node features are aggregated to represent the whole image-level embedding. All these components are learned under the joint guidance of reconstruction loss and ranking loss.

Visual Representation Extraction

In recent years, the research on object detection of natural images and scene sketches has attracted much attention and has made remarkable progress. We extract instances from scene images with the help of mature object detectors. Compared with manual annotation, this greatly reduces labor and cost while showing good performance. Specifically, two Mask RCNN detectors (He et al. 2017) that trained separately on SketchyScene and MS COCO are employed to localize regions of interest (RoIs) from scene sketches and photos, respectively, as shown in the left part of Fig. 2. The detected RoIs are denoted as $\mathcal{B} = (b_1, b_2, \ldots, b_n)$. Meanwhile, the deep convolutional network backbones extract visual features from the whole sketch/image. By removing the last pooling and fully-connected layers, the backbones produce feature maps $\boldsymbol{F} = \phi(\boldsymbol{I}; \boldsymbol{\theta})$ of size $h \times w \times d$, where $\boldsymbol{I}$ is the input sketch/image and $\boldsymbol{\theta}$ is all the learnable weights.

Conditional Graph Autoencoder

Generally, instance-level SBIR can be realized simply based on categorization or visual similarity, viz., candidate photos match the query sketch if they are predicted to be the same category or the distance between their feature vectors are close. However, the actual similarity measurement is complicated since a scene is determined by multiple objects and their relations. Not only categorical semantics and inter-object relevance should be considered, but also the overall visual features (including background) are important.

The backbone of the our CGAE model is a graph attention network, which employs a graph model to mine correlations between instance nodes. We extended it to two branches by adding a weighting sub-network to explicitly embed prior knowledge of spatial layout. The global visual features are encoded and injected into graph nodes as a contextual condition. And a weakly supervised semantic reconstruction module is designed to implicitly maintain the class distribution. Each functional module and the training procedure will be detailed in subsequent sections.

Conditional Graph Construction. In order to incorporate instance-level features, the detected RoIs are projected onto the feature maps and pooled to d-dimensional feature vectors $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ via RoIAlign (He et al. 2017). We model the relationship of RoIs as a graph with n nodes $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$, where $\vec{x}_i$ is the feature vector of each instance node. Instance-level features are essential for learning the correspondence of foreground objects, but background information and overall statistical distribution are omitted. Therefore, we encode the global visual features to supplement contextual information. This is implemented by applying a global average pooling layer to suppress the feature maps $\boldsymbol{F}$ into a feature vector:

$$\vec{g} = \frac{1}{hw} \sum_{j=1}^{w} \sum_{i=1}^{h} \boldsymbol{F}_{i,j,:}. \tag{1}$$

The conditional node features are then denoted as $\boldsymbol{h} = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_n\}$, where $\vec{h}_i = \vec{x}_i \parallel \vec{g}$, and $\parallel$ represents the concatenation operation.

Structure-Aware Graph Attention Network. The graph attention network (GAT) (Velickovic et al. 2018) performs self-attention to learn node importance. The attention coefficient of node j's feature to node i is computed as follows:

$$e_{ij} = \text{LeakyReLU}(\vec{a}^T \cdot [\boldsymbol{W}\vec{h}_i \parallel \boldsymbol{W}\vec{h}_j]), \tag{2}$$

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}, \tag{3}$$

where $\boldsymbol{W}$ is the learnable weight matrix of linear transformation applied to every node, $\vec{a}$ is the learnable weight vector of attention mechanism, and $\mathcal{N}_i$ is the neighbors of node i. Note that since the RoIs are regarded as a complete graph and GAT drops all structural information, $\mathcal{N}_i$ is equivalent to all graph nodes (including i). However, the spatial structure of scene layout is important prior knowledge. We extended GAT to incorporate structural constraints by adding a parallel weighting sub-network. The weighting coefficient between node i and j is defined based on the normalized distance (reversed to express the meaning of similarity or importance):

$$\beta_{ij} = \max(1 - \frac{\mathrm{dist}(b_i, b_j)}{\sqrt{h^2 + w^2}}, 0), \tag{4}$$

where $\mathrm{dist}(\cdot, \cdot)$ is a distance metric such as the Euclidean distance. Putting them together, the node features are finally transformed to new representation $\boldsymbol{h}' = \{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_n\}$, where the output features for every node is the combination of the two sub-networks with a potentially nonlinearity:

$$\vec{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} (\alpha_{ij} \boldsymbol{W} \vec{h}_j + \beta_{ij} \boldsymbol{W} \vec{h}_j)\right). \tag{5}$$

Weakly Supervised Semantic Reconstruction. The new node representation $\boldsymbol{h}'$ has encoded both instance-level correlations and global visual context. In order to induce graph nodes to keep categorical information, we designed a variant autoencoder structure to apply semantic consistency. Instead of directly reconstructing the feature vector of each node, the feature autoencoding is implicitly learned through a weakly supervised learning paradigm to maintain distribution of classification prediction (Bilen and Vedaldi 2016). As shown in the upper part of Fig. 2, the semantic reconstruction sub-module is a symmetric two-branch structure, which is built from the initial RoI features and the encoded node features. The branch with softmax operation predicts class-wise classification scores along the category dimension as follows:

$$\boldsymbol{U} = \boldsymbol{W}^{(\mathrm{cls})} \boldsymbol{X} + \vec{b}^{(\mathrm{cls})},\ \boldsymbol{U} \in \mathbb{R}^{\mathcal{K} \times n}, \tag{6}$$

$$\boldsymbol{U}'_{ij} = \mathrm{softmax}_i(\boldsymbol{U}_{ij}) = \frac{\exp(\boldsymbol{U}_{ij})}{\sum_{k=1}^{\mathcal{K}} \exp(\boldsymbol{U}_{kj})}, \tag{7}$$

where $\mathcal{K}$ is the number of categories. The gated branch learns to select important instances along the instance dimension:

$$\boldsymbol{V} = \boldsymbol{W}^{(\mathrm{det})} \boldsymbol{X} + \vec{b}^{(\mathrm{det})},\ \boldsymbol{V} \in \mathbb{R}^{\mathcal{K} \times n}, \tag{8}$$

$$\boldsymbol{V}'_{ij} = \mathrm{sigmoid}(\boldsymbol{V}_{ij}) = \frac{1}{1 + \exp(-\boldsymbol{V}_{ij})}. \tag{9}$$

The outputs of the two branches are combined through the Hadamard product to perform element-wise cross-attention and then accumulated along the instance dimension as the final $\mathcal{K}$-dim classification scores:

$$\vec{p} = \mathrm{softmax}((\boldsymbol{U}' \odot \boldsymbol{V}') \cdot \vec{\mathbf{1}}_{[n]}), \tag{10}$$

where $\vec{\mathbf{1}}_{[n]}$ is a n-dim all-ones vector.

Node Feature Aggregation. So far, we have obtained expressive node features that fuse instance features, global features, structure features, and semantic information. However, for retrieval purpose, they need to be aggregated into a single representation. A gated attention global pooling layer (Li et al. 2017) is appended to produce graph-level embedding as:

$$\vec{z} = \left(\mathrm{sigmoid}(\boldsymbol{W}^{(1)} \boldsymbol{h}' + b^{(1)}) \odot (\boldsymbol{W}^{(2)} \boldsymbol{h}' + b^{(2)})\right) \cdot \vec{\mathbf{1}}_{[n]}. \tag{11}$$

The reason we use gated *selective* pooling rather than *global* average pooling is that, as aforementioned, a major difficulty lies in scene-level SBIR is the misalignment of content between the two domains. Some objects in image domain may not be drawn in the corresponding sketch domain, or the artist can sometimes add extra imagination to sketches. The learnable partial aggregation of node features can make the model better adapt to two partially alignable domain.

Loss Function

Let $\mathcal{F}$ denote the whole network function as $\vec{z} = \mathcal{F}(\boldsymbol{I})$. Given a triplet t of an anchor sketch $\boldsymbol{S}$, a positive photo $\boldsymbol{P}^+$, and a negative photo $\boldsymbol{P}^-$, the following relation is expected:

$$\mathrm{dist}(\mathcal{F}(\boldsymbol{S}), \mathcal{F}(\boldsymbol{P}^+)) < \mathrm{dist}(\mathcal{F}(\boldsymbol{S}), \mathcal{F}(\boldsymbol{P}^-)). \tag{12}$$

This is achieved by adopting the margin-based ranking loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{tri}}(t) = \max(0, \Delta + \mathrm{dist}(\mathcal{F}(\boldsymbol{S}), \mathcal{F}(\boldsymbol{P}^+)) - \\ \mathrm{dist}(\mathcal{F}(\boldsymbol{S}), \mathcal{F}(\boldsymbol{P}^-))).\end{aligned} \tag{13}$$

The auxiliary semantic reconstruction loss is defined as the Kullback–Leibler divergence (Kullback and Leibler 1951) from the reconstructed to the original class distribution:

$$\mathcal{D}_{\mathrm{KL}}(\vec{p}_{\mathrm{ori}} \parallel \vec{p}_{\mathrm{rec}}) = -\sum \vec{p}_{\mathrm{ori}} \log \frac{\vec{p}_{\mathrm{rec}}}{\vec{p}_{\mathrm{ori}}}. \tag{14}$$

The final objective is to minimize the weighted sum:

$$\min_{\Theta} \left(\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathcal{L}_{\mathrm{tri}}(t) + \frac{1}{3N} \sum_{i=1}^{3N} \mathcal{D}_{\mathrm{KL}}(\vec{p}_{\mathrm{ori}} \parallel \vec{p}_{\mathrm{rec}})\right), \tag{15}$$

where $\mathcal{T}$ is all the triplets sampled from the $3N$ image batch.

Experiments

Experimental Setup

Datasets. SketchyScene (Zou et al. 2018) is the first and only usable large-scale scene sketch dataset to conduct scene-level SBIR. It consists of 7,264 human-created scene sketches and the reference images. Each scene sketch contains at least three object instances from 46 categories. After cleaning corrupted and duplicated data, we ended up with 5,616 sketch-photo pairs for training, 530 for validation, and 1,113 for testing. The recently proposed SketchyCOCO (Gao et al. 2020) is another scene-level sketch-image dataset, which was originally designed to study sketch to image generation task. However, due to its few categories (covering only 14 object classes) and poor instance-level alignment (missing major foreground objects in scene sketches),

Training Supervision	Method	acc.@1 ↑	MeanAcc ↑	MeanK ↓
Label	Categorization (He et al. 2016)	2.07%	7.15%	229
Label	SN (Sangkloy et al. 2016)	1.62%	7.56%	344
Triplet+Label	Sketchy (Sangkloy et al. 2016)	4.85%	15.60%	192
Triplet+Label+Box	SceneSketcher (Liu et al. 2020)	8.09%	20.15%	128
Triplet	Shoe (Yu et al. 2016)	10.78%	27.04%	118
	Hetero-CNN (Yu et al. 2016)	11.68%	30.49%	**92**
	GCN (Kipf and Welling 2017)	8.09%	21.34%	129
	GAT (Velickovic et al. 2018)	8.27%	22.69%	128
	CGAE (Ours)	**15.90%**	**33.09%**	97

Table 1: Comparison of scene-level sketch-based image retrieval on the SketchyScene datase (↑ means higher is better, while ↓ for the opposite).

the quality is not sufficient to support scene-level SBIR.

Evaluation Metrics. All compared methods are evaluated on the following three metrics:

- **acc.@K:** the most commonly used evaluation protocol for fine-grained SBIR that calculates the percentage of correct queries where the matching photo is included within the top-K recalled images.
- **MeanAcc:** the mean value of acc.@1 to acc.@K to reflect stable statistical performance. As convention, MeanAcc is calculated to acc.@10 as the top-10 results are more meaningful to users in practice.
- **MeanK:** the average ranking position of matching photos considering all queries. The acc.@K and MeanAcc metrics only measure the percentage of successful hits but cannot tell the severity of missed queries, whereas the MeanK is an overall estimate of expected performance.

Implementation Details. All the compared methods are re-implemented based on the same ResNet-50 (He et al. 2016) backbone. We adopt Mask R-CNN as the external object detectors, which do not participate in model training. Only the produced bounding boxes are used without any categorical labels. During training, input sketches and images are resized to 448×448 and randomly horizontally flipped to augment data. The dimensions of node features and graph embedding are set to 512. Models are all trained using the Adam optimizer (Kingma and Ba 2015) for up to 800 epochs. The early stopping strategy is employed to combat overfitting. The batch size and learning rate are set to 12 and 1e-4, respectively. All the experiments were conducted on NVIDIA Tesla P100 GPU with 16GB memory.

Competitors. We are investigating a cutting-edge new task and thus specially-designed solutions are rare. While independently developing scene-level SBIR, we noticed that Liu et al. explored a similar task, but with different supervision setting. We consider SceneSketcher (Liu et al. 2020) the most relevant state-of-the-art competitor. As in SceneSketcher, we adapt representative SBIR methods from instance-level to scene-level and perform training under fair configurations. These competitors are listed in Table 1.

- **Categorization:** a vanilla implementation of "retrieval by categorization" that compares the score vectors of multi-label classification.
- **SN:** another "retrieval by categorization" baseline that extracts the penultimate network layer activation of the backbone as the feature vectors.
- **Sketchy:** a heterogeneous cross-domain embedding method that includes a main triplet ranking loss and an auxiliary classification loss (Sangkloy et al. 2016).
- **Shoe:** a classic fine-grained SBIR method proposed in *Sketch Me That Shoe* (Yu et al. 2016), which employs a Siamese network to compare sketches and edge maps.
- **Hetero-CNN:** the heterogeneous version of "Shoe" that directly learns joint embedding of scene sketches and natural images instead of edge maps.
- **GCN:** applying a standard graph convolutional network (Kipf and Welling 2017) on the scene graph of RoIs to encode new node features.
- **GAT:** similar to "GCN" but instead applying a graph attention network (Velickovic et al. 2018). Note that the node features learned by "GCN" and "GAT" need to be aggregated to graph-level embeddings via a gated pooling layers as in our CGAE.
- **SceneSketcher:** a recently proposed GCN-based method to deal with scene-level SBIR that relies on sufficient instance-level dataset annotations including bounding boxes and ground-truth labels.

Results

The experimental results are listed in Table 1. From the results we made the following observations. (i) Our CGAE model surpasses all competitive methods on the acc.@1 and MeanAcc metrics and achieves state-of-the-art performance 15.90% and 33.09%, respectively. (ii) Although Sketchy and SceneSketcher utilized more training supervision (involving image triplets, instance class labels, and instance bounding boxes), they still perform much worse than ours, which

Backbone	Structure Weighting	Context Condition	Semantic Reconstr.	acc.@1 ↑	MeanAcc ↑	MeanK ↓
✓				8.27%	22.69%	128
✓		✓	✓	12.49%	26.96%	108
✓	✓		✓	13.28%	27.39%	107
✓	✓	✓		14.19%	32.17%	103
✓	✓	✓	✓	15.90%	33.09%	97

Table 2: Ablation study of the proposed CGAE model.

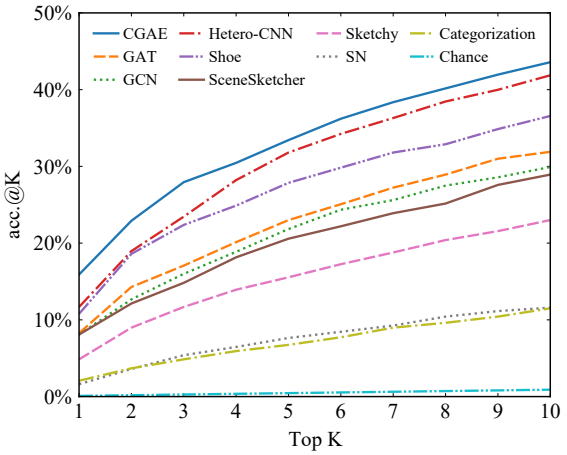


Figure 3: Detailed retrieval accuracy from top-1 to top-10.

confirms the effectiveness of our model design. (iii) As regards the MeanK indicator, CGAE still has obvious advantages over other competitors (the lower the better), except that only Hetero-CNN is slightly better than ours. This reflects that the average performance of Hetero-CNN is satisfactory, but it struggles to push the matching image to the top. (iv) Various graph-based native methods (e.g., GCN, GAT, SceneSketcher) have similar performance and are inferior to CNN-based feature extractors (e.g., Shoe and Hetero-CNN). Also based on graph model, our design principles have proven to be very effective in the scene-level sketch-based image retrieval task. Fig. 3 shows the detailed retrieval accuracy of each method from top1 to top10. Clearly, our model (CGAE) consistently outperforms all other competitors in different recalls. This indicates that our model has excellent and stable performance in top retrieval.

Ablation Study

We separately removed each module to analyze its contribution to the final model. The results are presented in Table 2. The first entry of backbone alone is the same as the GAT in Table 1, which serves as a baseline method. It can be observed from the table that the performance of the residual model is still better than GAT even if any module is removed. Among the three modules, the structure-weighting sub-network has greatest impact. Its ablation results in 3.41% and 6.13% degradation on the acc.@1 and MeanAcc metrics, respectively. Secondly, the global contextual information plays an important role, and its contribution is very close to that of the structure weighting. The semantic reconstruction makes improvements limited as expected since it is implicitly learned under weak supervision. In summary, the ablation studies confirmed the effectiveness of the sub-modules in our model.

Qualitative Results

Several retrieval examples of our method are presented in Fig. 4. The first column is the input query sketch, and the next six columns are the top retrievals. In the first three examples, the exact matching pictures are returned first. Our model is capable of handling complex scenes, even if they contain considerable object instances, such as the first example. In the last two examples, although the matching pictures are not retrieved first, the recalled top results are yet similar to the target. In other words, the model gains the ability of distinguish different scene types. Specifically, the first two recalled pictures in the penultimate example are much satisfactory under manual judgment. In fact, the current evaluation metrics only measure the exact matched images and ignore inherent intra-class similarities. Therefore, reconstruction of datasets and redesign of evaluation criteria are particularly important to promote field development. Nevertheless, our model has shown its ability to abstract general scene patterns.

Conclusion and Future Work

In this work, we studied a novel scene-level sketch-based image retrieval paradigm with minimal pairwise supervision. The sketch-specific and scene-specific challenges are discussed and analyzed. A Conditional Graph Autoencoder model is proposed to learn scene embedding from both the instance and global levels. The graph attention network is extended with contextual condition and structural constraint, and a variant semantic reconstruction decoder is designed adapt to weak supervised learning. The experimental results are very inspiring and reveal the great potential of the task.

This work is a preliminary investigation and there are many valuable problems worth exploring. On the one hand, scene pictures convey rich information and more characteristics could be further exploited, e.g., co-occurrence, orientation, and viewpoint. On the other hand, object-centric

Figure 4: Retrieval examples from the SketchyScene test set.

scenes are studied in this work, and scene graphs can be easily constructed with RoIs. However, this does not apply to landscapes containing non-rigid objects, e.g., sky and ocean. In the future, we will explore background-aware models to perceive semantics from general scenes.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (62071067, 62201072, 62171057, 62101064, 62001054), China Postdoctoral Science Foundation under Grant 2022M710468, in part by the Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center, in part by BUPT innovation and entrepreneurship support program.

References

Bilen, H.; and Vedaldi, A. 2016. Weakly Supervised Deep Detection Networks. In *CVPR*.

Canny, J. 1986. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-8(6): 679–698.

Dey, S.; Dutta, A.; Ghosh, S. K.; Valveny, E.; Llados, J.; and Pal, U. 2018. Learning Cross-Modal Deep Embeddings for Multi-Object Image Retrieval Using Text and Sketch. In *ICPR*.

Dey, S.; Riba, P.; Dutta, A.; Llados, J. L.; and Song, Y.-Z. 2019. Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval. In *CVPR*.

Donikian, S.; and Hégron, G. 1992. The Kernel of a Declarative Method for 3D Scene Sketch Modeling. *Program. Comput. Softw.*, 18.

Dutta, A.; and Akata, Z. 2019. Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-Based Image Retrieval. In *CVPR*.

Gao, C.; Liu, Q.; Xu, Q.; Wang, L.; Liu, J.; and Zou, C. 2020. SketchyCOCO: Image Generation From Freehand Scene Sketches. In *CVPR*.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-Cnn. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. In *ECCV*.

Herot, C. F. 1976. Graphical Input Through Machine Recognition of Sketches. In *SIGGRAPH*.

Huang, F.; Canny, J. F.; and Nichols, J. 2019. Swire: Sketch-Based User Interface Retrieval. In *CHI*.

Jiang, T.-B.; Xia, G.-S.; Lu, Q.-K.; and Shen, W.-M. 2017. Retrieving Aerial Scene Images With Learned Deep Image-Sketch Features. *J. Comput. Sci. Technol.*, 32(4): 726–737.

Kato, T.; Kurita, T.; Otsu, N.; and Hirata, K. 1992. A Sketch Retrieval Method for Full Color Image Database-Query by Visual Example. In *ICPR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet Classification With Deep Convolutional Neural Networks. *Commun. ACM*, 60(6): 84–90.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1): 79–86.

Lei, J.; Song, Y.; Peng, B.; Ma, Z.; Shao, L.; and Song, Y.-Z. 2019. Semi-Heterogeneous Three-Way Joint Embedding Network for Sketch-Based Image Retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 1–1.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2017. Gated Graph Sequence Neural Networks. In *ICLR*.

Liu, F.; Deng, X.; Zou, C.; Lai, Y.; Chen, K.; Zuo, R.; Ma, C.; Liu, Y.; and Wang, H. 2022. SceneSketcher-v2: Fine-Grained Scene-Level Sketch-Based Image Retrieval Using Adaptive GCNs. *IEEE Trans. Image Process.*, 31: 3737–3751.

Liu, F.; Zou, C.; Deng, X.; Zuo, R.; Lai, Y.-K.; Ma, C.; Liu, Y.-J. L.; and Wang, H. 2020. SceneSketcher: Fine-Grained Image Retrieval with Scene Sketches. In *ECCV*.

Pang, K.; Li, K.; Yang, Y.; Zhang, H.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2019. Generalising Fine-Grained Sketch-Based Image Retrieval. In *CVPR*.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Trans. Graphics*, 35(4): 1–12.

Shrivastava, A.; Malisiewicz, T.; Gupta, A.; and Efros, A. A. 2011. Data-Driven Visual Similarity for Cross-Domain Image Matching. *ACM Trans. Graphics*, 30(6): 1–10.

Smeulders, A.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12): 1349–1380.

Song, J.; Pang, K.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2018. Learning to Sketch With Shortcut Cycle Consistency. In *CVPR*.

Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In *ICCV*.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*.

Xu, K.; Chen, K.; Fu, H.; Sun, W.-L.; and Hu, S.-M. 2013. Sketch2Scene: Sketch-Based Co-Retrieval and Co-Placement of 3D Models. *ACM Trans. Graphics*, 32(4): 1.

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C. C. 2016. Sketch Me That Shoe. In *CVPR*, 799–807.

Yu, Q.; Yang, Y.; Liu, F.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Sketch-A-Net: A Deep Neural Network That Beats Humans. *Int. J. Comput. Vision*, 122(3): 411–425.

Yuan, J.; Abdul-Rashid, H.; Li, B.; and Lu, Y. 2019. Sketch/Image-Based 3D Scene Retrieval: Benchmark, Algorithm, Evaluation. In *MIPR*.

Zou, C.; Yu, Q.; Du, R.; Mo, H.; Song, Y.-Z.; Xiang, T.; Gao, C.; Chen, B.; and Zhang, H. 2018. SketchyScene: Richly-Annotated Scene Sketches. In *ECCV*.