Context-based Sketch Classification

Jianhui Zhang
Hong Kong UST

Yilan Chen
City University of Hong Kong

Lei Li
Hong Kong UST

Hongbo Fu
City University of Hong Kong

Chiew-Lan Tai
Hong Kong UST

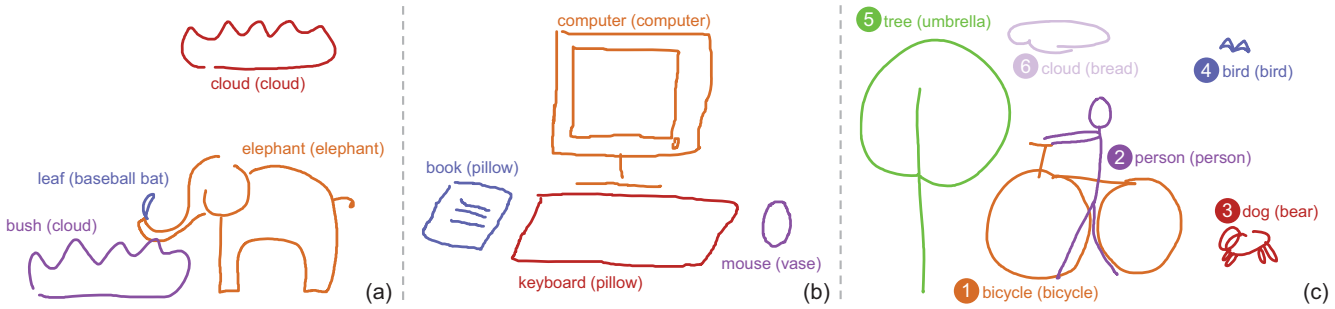

Figure 1: Given a sketched scene consisting of semantically segmented objects as input, our framework produces object categories with higher contextual compatibility compared to the predictions of the state-of-the-art single-object classification method (indicated within the parentheses). Our framework can be applied to both sketch co-classification (a)(b) and incremental sketch classification (c). The processing order is indicated next to the categories in (c).

ABSTRACT

We present a novel context-based sketch classification framework using relations extracted from scene images. Most of existing methods perform sketch classification by considering individually sketched objects and often fail to identify their correct categories, due to the highly abstract nature of sketches. For a sketched scene containing multiple objects, we propose to classify a sketched object by considering its surrounding context in the scene, which provides vital cues for alleviating its recognition ambiguity. We learn such context knowledge from a database of scene images by summarizing the inter-object relations therein, such as co-occurrence, relative positions and sizes. We show that the context information can be used for both incremental sketch classification and sketch co-classification. Our method outperforms a state-of-the-art single-object classification method, evaluated on a new dataset of sketched scenes.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → *Shape analysis*;

Expressive '18, August 17–19, 2018, Victoria, BC, Canada

ACM ISBN 978-1-4503-5892-7/18/08…$15.00
https://doi.org/10.1145/3229147.3229154

KEYWORDS

Sketch classification, context, object relations, co-analysis

ACM Reference Format:
Jianhui Zhang, Yilan Chen, Lei Li, Hongbo Fu, and Chiew-Lan Tai. 2018. Context-based Sketch Classification. In *Expressive '18: The Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering, August 17–19, 2018, Victoria, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3229147.3229154

1 INTRODUCTION

Freehand sketching is arguably one of the most accessible and efficient means for communicating ideas. However, semantically recognizing roughly-sketched objects remains an algorithmic challenge. Unlike natural images with rich texture details, sketches are a unique 2D art form that tends to be drawn at different levels of abstraction and in different styles, commonly with noticeable distortions. A robust sketch recognition approach capable of accommodating variations can facilitate better human-human communications and human-computer interactions in a high-level language and empower many practical downstream applications, such as graphical design, sketch-based retrieval or modeling.

Existing studies have mainly focused on identifying categories of individually sketched objects by casting the problem as an image classification task and applying available techniques from the natural image domain. Classifiers, such as Support Vector Machines (SVMs), are trained on certain image features that are extracted from a database of collected sketches [Eitz et al. 2012; Schneider and Tuytelaars 2014]. Recently, researchers have also employed deep Convolutional Neural Networks (CNNs) to learn more discriminant features for single sketched object classification [Sangkloy et al.

2016; Yu et al. 2017]. Although their accuracies (e.g., close to 80% on their tested datasets) surpass the performance of traditional approaches based on hand-crafted features (69%) [Schneider and Tuytelaars 2014], they still often produce wrong predictions due to the inherent interpretation ambiguity of individual sketches. Such a problem deteriorates considerably when the sketches are highly abstract and created in a short period, like millions of doodles in the QuickDraw dataset [Ha and Eck 2017].

We observe that users often sketch multiple objects together to compose a scene that visualizes a complex concept [Chen et al. 2009; Shin and Igarashi 2007; Xu et al. 2013]. Recognition uncertainty can be alleviated by considering relations among the objects in the scene as additional cues. For example, in Fig. 1-(b), it is very challenging for existing approaches to recognize each object individually (e.g., keyboard or mouse), since other categories in the training dataset (e.g., pillow or vase) may also contain sketches of similar shapes. However, if all the sketched objects are jointly considered, the compatibility of their predicted categories, in terms of co-occurrence and spatial relations, can strongly indicate that the drawing is more likely to be a workplace and that the ambiguous objects are more likely a keyboard and a mouse.

This motivates us to propose a new sketch classification framework that is endowed with relations of object categories to better address the above ambiguity issue. To this end, we need to solve two main challenges. Firstly the extraction of relation priors on categories of sketched objects requires voluminous sketch data. Unfortunately, there is a lack of large-scale datasets of sketched scenes at present. Existing sketch datasets, such as [Eitz et al. 2012; Ha and Eck 2017; Sangkloy et al. 2016], only comprise drawings of single objects. We resort to existing image datasets that contain rich annotations of objects in real-world scenes, and we show that it is a viable solution to transfer and apply the learned relation priors across different domains (i.e., image to sketch). The other challenge is to identify and quantify relations that are effective for ambiguity resolution, and then to unify them.

Taking a scene sketch composed of semantically segmented objects as input, we leverage existing CNNs, which are trained for single-object classification, to produce candidate categories for each sketched object in the scene. With the relation priors extracted, we define a compatibility function to evaluate the plausibility of specifically assigned categories of a pair of sketched objects. Then a prioritizing process is designed to select the most probable candidates that respect the extracted relation priors well so that a semantically valid scene is formed. Specifically, we present two context-based sketch classification algorithms that target different sketching scenarios: an incremental algorithm (Fig. 1-(c)) for recognizing a new object given confirmed categories of existing objects as context, and a co-classification algorithm (Fig. 1-(a)(b)) for jointly identifying the categories for all the objects in the scene at once. To test the performance of the proposed algorithms, we collect an evaluation dataset that contains 332 freehand scene sketches, covering various object categories and scene types. Overall our incremental classification algorithm is 6.7% higher and our co-classification algorithm is 6.3% higher than the CNN-only method [Sangkloy et al. 2016] in terms of classification accuracy. We will release this dataset to the community for future relevant research.

To sum up, our contributions in this work are three-fold: 1) extracting and transferring relation priors from the image domain to the sketch domain to alleviate recognition ambiguity in sketches; 2) a context-based sketch classification framework with two specific algorithms for different sketching scenarios, each achieving higher accuracy than the state-of-the-art CNN for single-object classification; 3) a new dataset of scene sketches for benchmarking the performance of relevant recognition algorithms.

2 RELATED WORK

The seminal work Sketchpad [Sutherland 1964] introduced sketching as a means of human-computer interaction in the earliest days of the computer graphics field. Since then, much research effort has been devoted to low-level understanding of sketches, such as identifying geometric primitives (e.g., lines or circles) [Paulson and Hammond 2008; Sezgin et al. 2001] or discovering repetitive patterns for gesture recognition [Donmez and Singh 2012]. Information like stroke ordering has been used in sketch recognition [Sezgin and Davis 2005]. Prior knowledge can also be utilized to facilitate the recognition of domain-specific drawings [Alvarado and Davis 2004], such as mathematical expressions [LaViola and Zeleznik 2004], chemical drawings [Ouyang and Davis 2011], electronic circuit diagrams [Kara and Stahovich 2004; Sezgin and Davis 2008] or architectural drawings [Lu et al. 2005].

Recently, high-level semantic understanding of sketches, in terms of objectness, has received an increasing amount of research attention. However, due to different levels of abstraction and styles in freehand drawings, recognizing a single sketched object semantically is challenging, especially for highly abstract ones. Given a dataset of sketched objects as training data, a common paradigm among existing studies is to first extract certain hand-crafted features (most of them originally proposed for natural images) from the sketch images and then to train a classifier (e.g., SVMs) on the feature representations to predict the categories of unseen sketched objects. The first work for large-scale recognition analysis of human sketched objects, by Eitz et al. [2012], used a bag-of-features representation [Sivic and Zisserman 2003] of sketch images and trained multi-class SVMs on a dataset of 250 object categories, collected via crowd-sourcing, achieving a recognition accuracy of 56%. In contrast, humans can recognize on average 73.1% of sketches in this dataset. Several follow-up studies differ mostly in the use of feature representations originated from natural images, for instance, by replacing bag-of-features with Fisher Vectors [Schneider and Tuytelaars 2014] (with the recognition accuracy improved to 68.9%) or by combining several types of local features [Li et al. 2015] (65.8%).

Visual recognition tasks [Russakovsky et al. 2015], such as object detection and classification in natural images, have advanced significantly thanks to the rapid development of deep learning. Similar approaches have been applied to single sketched object recognition. Instead of relying on hand-crafted features of sketch images, Yu et al. [2017] adopted a CNN architecture, matured in visual recognition tasks, to learn more discriminant hierarchical features for classifying sketches more robustly. Their carefully-designed sketch classification network outperforms humans, achieving an accuracy of 77.95% on the TU-Berlin dataset [Eitz et al. 2012]. To

address the lack of texture details in sketches, Zhang et al. [2016] introduced another CNN-based method that uses natural images as an intermedium for learning shared features of sketches more effectively. The recent work done by Sangkloy et al. [2016] reported the state-of-the-art accuracy (80.85%) using GoogLeNet [Szegedy et al. 2015] on the above dataset.

We reiterate that the aforementioned methods for high-level semantic recognition of sketches, including the ones with hand-crafted features and the ones with CNNs, are designed for individually sketched objects. In practice, multiple objects are often drawn together to depict a more complex scene, for example, for the composition of new images [Chen et al. 2009; Hu and Collomosse 2013], diagrams [Yesilbek and Sezgin 2017] or 3D scenes [Shin and Igarashi 2007; Xu et al. 2013]. In such scenarios, the relations among the sketched objects can provide strong cues for alleviating recognition uncertainty, which is the main focus of our work.

Leveraging object relations has been explored extensively in scene-object classification tasks in natural images, but rarely in scene sketches. Galleguillos and Belongie [2010] summarized three types of relations mainly used in image classification: co-occurrence, positions and sizes. Generally, such relation features are utilized after an independent categorization step to filter out incompatible results [Felzenszwalb et al. 2010; Rabinovich et al. 2007], or they are concatenated with other image features for joint classification [Li et al. 2009; Malisiewicz and Efros 2009; Mottaghi et al. 2014]. The latter however, requires a larger number of annotated images for training. As there is a lack of labeled scene-sketch data, we choose to transfer relations from the image domain to the sketch domain, and combine them with single sketched object recognition methods. Thus our work is more similar to the first approach.

Fisher et al. [Fisher and Hanrahan 2010] presented a context-based 3D model retrieval method, which uses relations between objects to predict potential candidates for subsequent retrived objects. Xu et al. [2013] proposed a sketch co-analysis approach, which is intended for finding semantically more meaningful 3D model combinations from individually retrieved results for sketch-based indoor scene modeling. Our work extends their co-analysis idea of utilizing context information to a more general sketch classification problem and aims to accommodate various object categories and scene types beyond indoor scenes. Similar to [Xu et al. 2013], we assume that an input scene sketch is composed of semantically segmented objects drawn one by one. Semantic sketch segmentation [Arandjelović and Sezgin 2011; Huang et al. 2014; Sun et al. 2012] is a challenging problem on its own. In this work, we mainly focus on the context-based recognition of sketched scene objects, expecting segmentations from user inputs.

Sketch-based image retrieval (SBIR), another line of studies, is related to sketch recognition, but generally aims to find visually the most relevant image content with respect to an input sketch [Cao et al. 2013; Chen et al. 2009; Eitz et al. 2011; Yu et al. 2016], instead of attempting on semantic understanding of the sketch. Sketch recognition, as demonstrated by [Li et al. 2013], can help to reduce the retrieval space significantly if the category of the input sketch is robustly identified. Recent works utilize Siamese CNN to minimize the distance between relevant sketches and images [Bui et al. 2017; Qi et al. 2016]. Several studies on SBIR, such as [Portenier et al. 2017; Qian et al. 2016], also investigated supplementing existing SBIR systems with a refinement post-process to group and re-rank the retrieval results. Despite sharing a similar refinement idea with our context-based framework for sketch classification, their designs heavily utilize the content and features of the retrieved real images while our work strives to use the learned relation priors to prioritize the candidate categories of input sketches that contain extremely limited feature information.

3 OVERVIEW

Our task is to take a scene sketch as input, which contains semantically segmented objects drawn one by one, and recognize the objects therein of unknown categories by considering both each sketched object itself and its context. We first use a pre-trained CNN for individual object recognition, which produces a set of candidate categories for each object. We then employ relation knowledge learned from a large-scale scene-image dataset, together with the CNN predictions, and select the most probable candidates by assessing the object category compatibility in the scene, yielding a more robust and accurate sketch classification framework (see Fig. 2).

Similar to [Sangkloy et al. 2016], we adopt the GoogLeNet [Szegedy et al. 2015] as the base of our CNN architecture due to its demonstrated state-of-the-art performance on sketches. Other single-object classification methods may be used alternatively. This network is pre-trained on ImageNet [Russakovsky et al. 2015] and then fine-tuned on a sketch dataset with the last layer modified to produce estimations over our used categories (Section 6.1).

While there is no large dataset of scene-sketches available for learning the relation knowledge of object categories, we use existing scene-image datasets that are of large volume and contain rich annotations [Krishna et al. 2017; Xiao et al. 2010]. We assume that, when drawing multiple objects, users still loosely follow the relations found in real scenes. We construct a graph to represent the extracted relations such as co-occurrence and spatial relations of object categories (Section 4).

Next we define a relation compatibility function that measures how well a pair of objects and their assigned categories respect the learned relation priors. We use this function as a building block and design context-based incremental classification and co-classification algorithms for recognizing the sketched objects in a scene (Section 5), catering two different sketching paradigms. We show that the proposed algorithms, leveraging both CNNs and the relation graph, are effective at alleviating recognition ambiguity existing in freehand sketches.

4 RELATION EXTRACTION

Now we describe the procedure of extracting relation knowledge from a large-scale scene-image dataset. We use the Visual Genome dataset [Krishna et al. 2017], which contains rich annotations such as attributes and relationships of object instances. We aim at constructing a directed relation graph $G = (V, E)$, where the nodes V denote the set of object categories of interest, and the edges E denote the set of pairwise relations among the categories. For an ordered category pair $\langle u, v\rangle$ $(u, v \in V)$, if u has certain valid relations with v (Fig. 3), a link edge e_{uv} from u to v is inserted into E, and in turn, so as is e_{vu} for $\langle v, u\rangle$. For simplicity, in the following

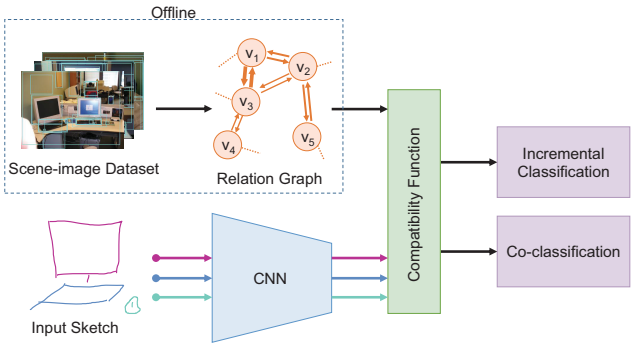

Figure 2: Pipeline of our context-based sketch classification framework.

sections, we use $\langle u, v\rangle$ to illustrate the extracted relations; the inverse relations can be easily derived according to our computational rules. To quantify the pairwise relations of $\langle u, v\rangle$, we extract two types of information from the dataset and associate them with e_{uv}: co-occurrence and spatial association. Next we give the detailed definitions for each type.

4.1 Co-occurrence

Categories u and v are likely to have a strong correlation if object instances of these two categories often appear together in the used scene-image dataset. We take advantage of the existing semantic annotations of objects in the Visual Genome dataset to obtain more reliable co-occurrence relations. In the dataset, if two objects of categories u and v are annotated with a semantic relation, for example, person *wearing* hat (Fig. 3-(a)(b)), we count them as an object relation pair. By collecting the statistics of object relation pairs in the dataset, we define the co-occurrence of $\langle u, v\rangle$ as

$$\eta_{uv} = \begin{cases} 0.5 \cdot \left(\frac{\Lambda(u,v)}{\Lambda(u)} + \frac{\Lambda(u,v)}{\Lambda(v)} \right) & \text{if } \Lambda(u,v) \geq 50 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\Lambda(u, v)$ counts the number of images containing object relation pairs of u and v, $\Lambda(u)$ counts the number of images containing u and so as does $\Lambda(v)$. Note that $\eta_{vu} = \eta_{uv}$. Instead of adopting an intersection-over-union form, the formulation in Eq. 1 can give stable results when one of the categories has a large number of object instances (e.g., person) yet the other category has a small number of object instances (e.g., backpack).

4.2 Spatial Association

To measure the spatial association of an ordered category pair $\langle u, v\rangle$, we again make use of the object relation pairs in the Visual Genome dataset and introduce a set of *discrete* attributes R_{uv} to describe the annotations (e.g., person *wearing* hat) with respect to the bounding boxes of objects. The attribute set, R_{uv}, includes relative position, relative size and overlap ratio of the objects of categories u and v. Specifically, R_{uv} includes:

- the objects of u are *above* (*below*) the objects of v w.r.t their bounding box centers;

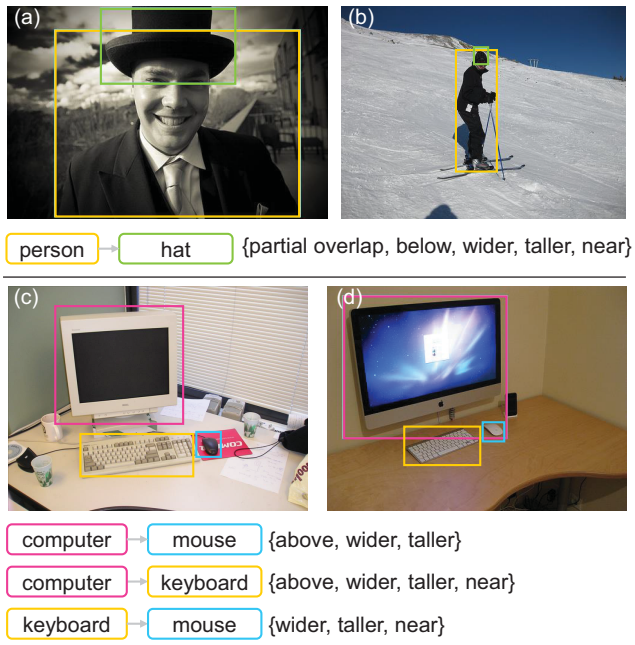

Figure 3: Example images with object annotations from the Visual Genome dataset. Images (a) and (b) contribute to the relations between person and hat, while images (c) and (d) contribute to the relations among computer, mouse and keyboard.

- the objects of u are *near* to (*far* from) the objects of v w.r.t their bounding box centers;
- the objects of u are *wider* (*narrower*) than the objects of v w.r.t their bounding box widths;
- the objects of u are *taller* (*shorter*) than the objects of v w.r.t their bounding box heights;
- the objects of u have *full* (*partial*, *no*) overlap regions with the objects of v.

The details of computation and thresholds are provided in the supplementary material. We iterate all the images with object relation pairs of u and v, and compute the frequencies for all the listed attributes. Only the attributes with high-frequency ($\geq$ 70% in our implementation) are kept in R_{uv}. For example, as shown in Fig. 3, since a mouse may appear above or below a keyboard with approximately equal frequency when being captured from arbitrary viewpoints, the category pair ⟨keyboard, mouse⟩ discards the *above/below* attribute that conveys unreliable vertical position information. We then associate the filtered attribute set R_{uv} with edge e_{uv}. Note that for edge e_{vu}, the attribute values in R_{vu} can be easily derived from R_{uv}. For example, in Fig. 3, the category pair ⟨hat, person⟩ has {*partial overlap, above, narrower, shorter, near*}. Unlike the *above* and *below* attributes, which often stem from physical supporting constraints, the *left* and *right* attributes may be easily affected by the camera viewpoints and are thus less reliable. Thus, we excluded the *left* and *right* attributes from spatial association.

Discussions. For a category pair $\langle u, v\rangle$, we initially tried a different approach to quantifying the spatial association by estimating continuous probability density functions on relative positions and sizes of object instances. Specifically, we tried to use kernel density estimation to capture the distribution of object centers of u with respect to the object centers of v in the dataset. However, we found that for many category pairs, it is hard to estimate reliable probability models, possibly because the supporting samples of relative positions are often quite noisy. A similar problem was also encountered during the estimation of probability models for relative sizes. Our adopted approach based on discrete attributes, as described above, works well as shown in Section 6.2.

5 CONTEXT-BASED CLASSIFICATION

In this section, we first define a pairwise score function to assess the relation compatibility of two sketched objects being assigned with specific candidate categories (Section 5.1). The pairwise score function is built upon the extracted relation priors from Section 4. Then we present two context-based sketch classification algorithms for different scenarios: incremental classification (Section 5.2) and co-classification (Section 5.3) of sketches, showing the usefulness of employing relation knowledge to alleviate recognition ambiguity.

5.1 Relation Compatibility Function

Given two sketched objects o_1 and o_2 and their assigned object categories u and v as input, the relation compatibility function considers the co-occurrence and spatial association of the categories to produce a score, indicating the plausibility of $\langle o_1, o_2\rangle$ being labeled as $\langle u, v\rangle$ with respect to the extracted relation priors.

We first define a spatial association score ρ for measuring how well the bounding boxes of $\langle o_1, o_2\rangle$ satisfy the spatial attributes in R_{uv} as follows:

$$\rho(o_1, o_2, u, v) = \begin{cases} a^{\Delta(o_1, o_2, R_{uv})} \cdot b^{|R_{uv}| - \Delta(o_1, o_2, R_{uv})} & \text{if } |R_{uv}| > 0 \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

where $\Delta(o_1, o_2, R_{uv})$ counts the number of attributes in R_{uv} that are satisfied by the bounding boxes of o_1 and o_2, $|\cdot|$ denotes cardinality, a is a bonus factor and b is a penalty factor. We set $a = 1.5$ and $b = 0.5$. Intuitively, objects o_1 and o_2 are more compatible to each other if their bounding boxes satisfy more relation priors.

Next, we define the relation compatibility function ψ as

$$\psi(o_1, o_2, u, v) = \eta_{uv} \cdot \rho(o_1, o_2, u, v), \tag{3}$$

where the first term is the co-occurrence score and the second term is the spatial association score. In the following, we demonstrate the usage of ψ in two context-based sketch classification scenarios.

5.2 Incremental Classification

In an interactive scenario, the user may choose to confirm the category after sketching each object in the scene. Thus, to classify a newly sketched object, a definite scene context with confirmed categories of existing objects is given to the algorithm.

Suppose there are n context objects $\{o_i\}_{i=1}^n$ with the corresponding known categories $\{v_i\}_{i=1}^n$ ($v_i \in V$). Let o_{n+1} denote the newly sketched object that needs to be recognized. We first feed o_{n+1} into our pre-trained CNN to obtain the ranked top-k candidate categories $\{\bar{v}^j\}_{j=1}^k$. Next, we define a re-ranking score function to select the most compatible category for o_{n+1} with respect to the given context as follows:

$$\chi(o_{n+1}, \bar{v}^j | \{o_i\}, \{v_i\}) = p(\bar{v}^j | o_{n+1}) \cdot \max(m \sum_{i=1}^{n} \psi(o_{n+1}, o_i, \bar{v}^j, v_i), \lambda), \tag{4}$$

where $p(\bar{v}^j|o_{n+1})$ is the probability of o_{n+1} being of category $\bar{v}^j$, as predicted by the CNN, and m is the number of context objects whose categories $\{v_i\}$ have valid co-occurrence relations with $\bar{v}^j$ as defined in Eq. 1. For a candidate category that has zero or little compatibility with context objects, in order to maintain the influence of CNN, we assign a fixed value of re-ranking score $\lambda = 0.001$ in our implementation to it. By multiplying a factor m, we amplify the influence of aggregated contexts. In other words, in Eq. 4, we consider the probability of an object to be assigned with a certain category by its appearance as well as the relation compatibility of the whole scene. We iterate all the top-k candidate categories (see Section 6.2), compute the re-ranking scores, and classify the sketched object o_{n+1} as the category with the highest score.

We designed a simple user interface for sketching with our incremental classification algorithm. In our UI, we allow the user to confirm the category of each individually sketched object in a scene and fix wrong predictions if any. Please refer to the supplementary video for demos of our UI and algorithm.

5.3 Co-classification

Co-classification of sketched objects shares the same spirit with co-analysis. Instead of solving problems individually, co-analysis correlates the problems and aims at achieving a certain degree of consistency across the results [Sidi et al. 2011; Xu et al. 2013]. Different from incremental classification, with co-classification, all the objects in the input scene sketch do not have any pre-confirmed labelings and are analyzed jointly for their final category predictions. Such an algorithm allows the user to draw the objects continuously without interruptions to confirm the categories one by one, and is particularly useful for recognizing objects in sketched scenes created offline.

Given a set of n sketched objects $\{o_i\}_{i=1}^n$, the top-k candidate categories $\{\bar{v}_i^j\}_{j=1}^k$ for each object o_i are first obtained via the CNN. Next we define a score function δ to evaluate the plausibility of assigning each o_i with one of its candidates. Suppose one of the assignment combinations is $\{\hat{v}_i\}_{i=1}^n$, where $\hat{v}_i \in \{\bar{v}_i^j\}_{j=1}^k$. Let

$$\delta(\{o_i\}, \{\hat{v}_i\}) = \sum_{i=1}^{n} p(\hat{v}_i|o_i) \cdot \max(m \sum_{s \neq i} \psi(o_i, o_s, \hat{v}_i, \hat{v}_s), \lambda), \tag{5}$$

where $p(\hat{v}_i|o_i)$ is the probability, predicted by the CNN, of category $\hat{v}_i$ assigned to object o_i and λ is a small constant which is same as that in Eq. 4 to replace the context score when the candidate category has no relation with its context. Similar to the incremental classification method, the formula above aims to maximize the category compatibility of the assigned combination for each object, leaving other objects as context. However, a brute-force search strategy to maximize the score function δ is in $O(n^2 \cdot k^n)$,

which is computationally prohibitive even for a moderate number of sketched objects. Instead, we adopt a beam search algorithm similar to [Xu et al. 2013]. More specifically, we first consider all possible category pairs for every pair of sketched objects and compute the corresponding scores with δ (i.e., by setting $n = 2$). Each category pair is called a size-2 category set. The time complexity for enumerating all size-2 sets is $O(n^2k^2)$ and we leave only r sets with the highest scores for further processing. Next, we iteratively generate the optimal size-p sub-set via adding one new object to each size $(p-1)$ set or adding two new objects to each size $(p-2)$ set. At each iteration, the algorithm will produce at most $O(nkr)$ size p sets for size $(p-1)$ sets, and $O(r^2)$ sets for size $(p-2)$ sets. The above steps are repeated until the size of sub-sets reaches n. Finally, we leave the full-size set with the highest score of δ as our co-classification result.

6 EVALUATION

To validate our proposed relation-supported sketch classification, we first collected a scene-sketch evaluation dataset along with the ground truth labelings (Section 6.1). Then we performed a series of experiments on the collected dataset, comparing both the incremental classification and co-classification algorithms with the state-of-the-art CNN for single sketched object recognition, to show the effectiveness of incorporating relations for ambiguity resolution (Section 6.2).

6.1 Data Collection

We used the QuickDraw dataset [Ha and Eck 2017] which contains 345 object categories and 70K individually sketched objects per category to train the CNN, and used the Visual Genome dataset [Krishna et al. 2017] which contains 108K scene-images, 80K object categories, 3.8 million object instances to extract relation knowledge. Furthermore, to get a set of object categories with less redundancy and noise, we first obtained the intersection of categories of the QuickDraw and Visual Genome datasets, and then removed the categories that have no co-occurrence. We used this filtered set of categories(in total 70 categories) to construct our sketch scene dataset. The category selection step allows us to unify the pre-trained CNN and the extracted relation priors that are used in our proposed framework. We used a simple sketching interface similar to the one for incremental classification (Section 5.2) for data collection.

We recruited 15 participants with basic drawing skills and asked them to quickly draw simple scene sketches that are semantically meaningful to humans. There were no strict time limits, though we found that most participants finished one scene sketch within three minutes. The participants needed to draw in two different ways: 1) freely choosing categories from the full set to compose a scene; 2) choosing categories only from a given subset to compose a specific type of scene. For the second case, we roughly divided the used object categories into six subsets, representing different scene types, which include dining room, home, outdoor activity, outdoor nature, street view and workplace. The purpose of dividing into different scene types is to test whether our context-based method can perform better under specific scene types where the relations among objects tend to be stronger compared to freely drawn scenes.

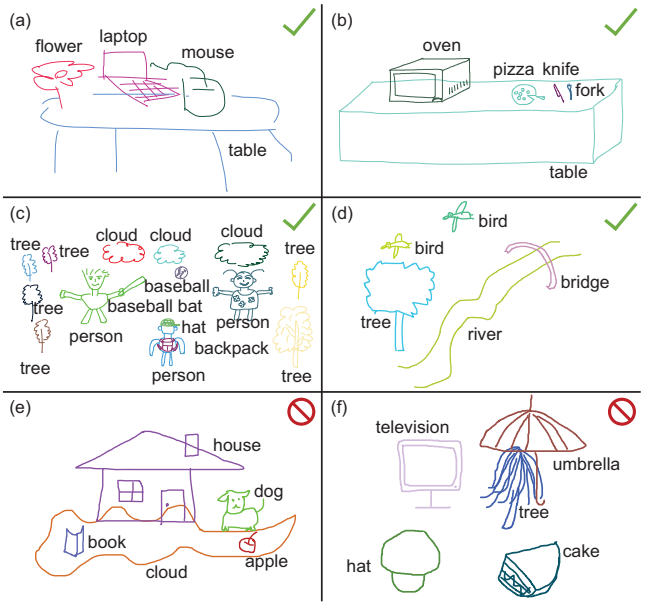

Figure 4: Examples of our collected scene sketches along with ground truth labelings. Sketches (a)-(d) are kept in the dataset while sketches (e)-(f), representing invalid scenes, are discarded from the dataset after a data verification process.

More specifically, each participant was asked to draw seven groups of sketches in total (i.e., one group of the freely composed scene type and six groups of pre-defined scene types) and three to four sketches per group, and in each scene sketch, at least three objects and ground truth labelings were required.

After the data collection, similar to [Eitz et al. 2012; Sangkloy et al. 2016], we manually reviewed all the sketches for verification. One of the users seemed to have misconceptions of our requirements and created sketches that mostly do not represent real-world scenes (Fig. 4-(e)(f)), which violate our assumption that users loosely follow the relations found in real scenes during sketching. We thus removed the collected data of this user and kept the sketches of all the remaining 14 users. Fig. 4 shows some examples of our collected data. As a result, we retained 332 out of 354 collected sketches, which cover all 70 object categories and contain 1,568 individual objects. We then randomly selected 1/3 sketches from each group of sketched scenes drawn by each user as the validation set and left the rest as the test set, resulting in 110 validation samples and 222 test samples. The validation set was used for parameter fine-tuning. Please refer to the supplementary material for the summarized statistics of our dataset.

6.2 Experiments

We evaluate the performance of our proposed context-based sketch classification framework on the test set of the above collected scene-sketch dataset, and as a baseline, we also compare it with the state-of-the-art CNN for single sketched object classification [Sangkloy et al. 2016].

Table 1 shows the top-1 accuracies of the CNN, our incremental classification and co-classification algorithms in different scenes. The accuracy is calculated as the number of objects with correctly predicted categories over the total number of objects of a specific scene type in the test set. To evaluate the CNN-only method, we feed each individually sketched object into the CNN and test the top-1 prediction against the ground truth. To evaluate our incremental classification algorithm, we simulate the interactive sketching scenario, that is, we progressively classify each object in a scene in the drawing order given by the users during data collection. (See *User Input* in Table 1). For the first object in an input scene, only the CNN is used for prediction and no context computation is involved. Then for each subsequent object, we assign the ground truth to the already tested objects that serve as context, and perform predictions as described in Section 5.2. Specifically, we use all the object categories ranked by the CNN as the candidates, since Eq. 4 can be evaluated efficiently. To evaluate our co-classification algorithm, the categories of all the objects in an input scene are jointly predicted, as described in Section 5.3, without considering the drawing order. For each object, we use top-10 categories predicted by the CNN as candidates. Fig. 8 shows the classification results of several sketched scenes with the above three methods.

Table 1: Top-1 accuracy (%) of each method for each scene. "User input" means the use of the input order of sketched objects and "Random" means the random drawing order of objects.

	CNN	Increment.		Co-class.
		User Input	Random	
Dining room	80.1	**88.5**	82.4	**88.5**
Home	69.1	**78.1**	76.3	75.6
Outdoor activity	65.7	69.7	64.1	**70.4**
Outdoor nature	69.4	**79.1**	74.0	78.4
Street view	79.1	**83.2**	81.0	81.5
Workplace	71.2	79.7	73.9	**83.8**
Free	81.3	**86.7**	84.0	85.3
All	74.5	**81.2**	78.1	80.8

Overall our incremental classification algorithm is 6.7% higher and our co-classification algorithm is 6.3% higher than the CNN-only baseline. The one-tail, paired t-tests confirm that both the incremental classification and the co-classification algorithms significantly outperform the CNN-only method (p-value < 0.05). Our incremental classification and co-classification algorithms have corrected 26.2% and 25.8% of the wrong predictions by the CNN, respectively, and made mistakes only for 1.1% and 1.5% of the correct predictions by the CNN.

In terms of improvements for specific scene types, we find that our incremental classification and co-classification algorithms work well particularly in home (9.0%, 6.5% higher than the CNN), dining room (8.4%, 8.4%), outdoor nature (9.7%, 9.0%) and workplace (8.5%, 12.6%), The performance improvement for outdoor activity (4.0%, 4.7%) and street view (4.1%, 2.4%) is relatively smaller. The reason for the decline is that the selected object categories of these scene types share fewer relations and that users tend to draw from perspectives different from photos of these scene types. For the free scene type where object relations may be less strong, accuracy improvements (5.4%, 4.0%) are still observed.

Since our incremental classification algorithm re-ranks the category candidates predicted by the CNN, we can calculate and compare the top-1 to top-10 accuracies on the dataset, which are plotted in Fig. 6. (Note that there is no re-ranking of candidate categories for each object involved in our co-classification algorithm where a bottom-up greedy search is used.) We can see that our context-based algorithm also improves the CNN results under such measurement. In addition, since drawing order is involved in the accuracy computation of our incremental classification algorithm, we also evaluate our algorithm with the random drawing order of objects. Specifically, we randomly permute the drawing order of objects in an input scene and then classify the objects with the new order. We repeat this procedure for 10 times and calculate the average accuracy. Table 1 (*Random*) shows the experiment results. It is expected that the accuracy declines with random order. We deduce that this is because every two objects drawn sequentially tend to have stronger relations. The randomized drawing order may cut this underlying connection in the incremental classification scenario.

To validate the effectiveness of our relation formulation, we disable the spatial association term in Eq. 3 by setting it to 1 and then calculate the corresponding accuracies of our modified classification methods. As shown in Table 2, together with Table 1, the overall accuracy of all the sketches declines by 1.1% and 2.0% respectively for the incremental classification and co-classification algorithms without the spatial association term. The t-tests shows that both the incremental classification and co-classification algorithms outperform the ones without the spatial term with p-value < 0.05.

Table 2: Top-1 accuracy (%) of our context-based classification methods without the spatial association term.

	Increment.	Co-class.
Dining room	87.0	88.5
Home	77.2	75.6
Outdoor activity	68.4	67.1
Outdoor nature	77.7	75.0
Street view	83.2	79.2
Workplace	77.1	79.6
Free	86.2	84.4
All	80.1	78.8

Different levels of drawing skills of the recruited users result in varying quality of the collected sketches, which can affect the recognition performance. In Fig. 5, we show the accuracies of the three classification methods applied to all the sketched scenes of a specific user. Among the results, the sketches of User 2 receive the lowest recognition accuracy with all the three classification methods. Fig. 7-(U2) shows an example sketch from User 2. Nevertheless, for User 2, our incremental and co-analysis classification algorithms are still able to perform reasonable context inference

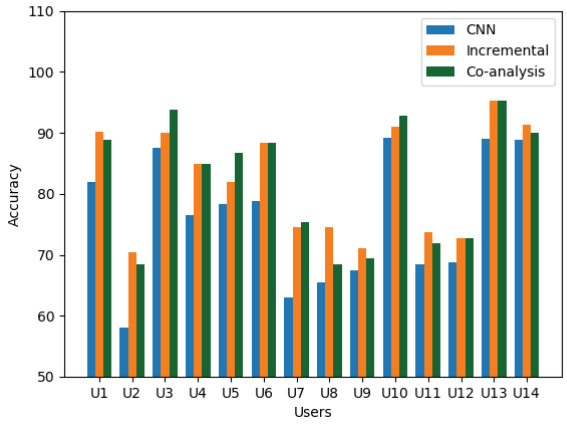

Figure 5: Top-1 Accuracy (%) for sketches drawn by different users.

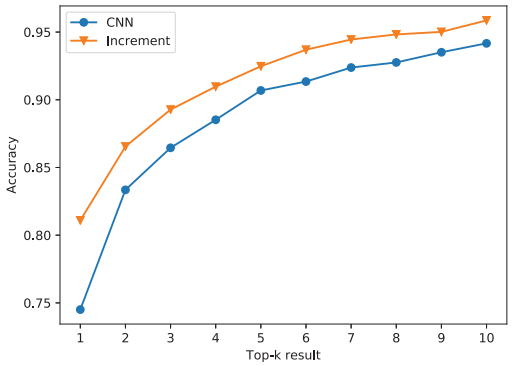

Figure 6: Top-k accuracy (%) of incremental classification.

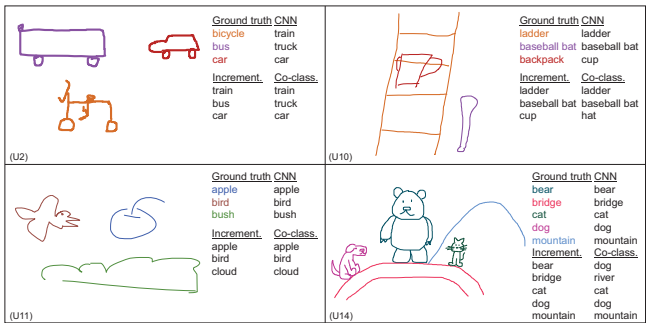

Figure 7: Representative sketches from Users 2, 10, 11 and 14, on which our classification algorithms do not perform as well as expected.

to improve the overall classification accuracy (12.4%, 10.4% higher than the CNN).

The sketches of Users 10, 11 and 14 gain less improvement in recognition accuracy with our co-classification algorithm. It is partially because the objects in the scenes drawn by these users rarely appear together in the real world or strongly violate our spatial term. The drawing intuition of people sometimes disagrees with the real-world scenes, that is, people may tend to draw objects of similar categories together, even though they may not usually appear together in the real world (e.g., the animals shown in Fig. 7-(U14)). One possible improvement for our algorithms is to manually enrich our relation graph with such kinds of relations. The example scenes casually composed by the users without reasonable spatial relations are shown in Fig. 7-(U10)(U11). Such cases make it hard for our algorithm to correct the CNN predictions.

In Fig. 8, we demonstrate more results of both our incremental and co-classification methods. These results show that our algorithm can still work when CNN produces several incorrect predictions.

7 DISCUSSION AND FUTURE WORK

In this work, we have developed a context-based framework for sketch classification. We propose to co-analyze the objects within a scene sketch and optimize their compatibility with respect to relation priors extracted from an existing image dataset. Our approach can be applied to two scenarios, namely incremental classification and co-classification. We demonstrate the superior performance of our approach over the state-of-the-art individual sketch classification method through quantitative evaluation on a newly collected dataset of scene sketches.

Our context-based sketch classification framework will be useful for various applications. For example, incremental classification can be seamlessly integrated with interactive drawing systems to better support human-computer interactions, as shown in the supplementary video. In addition to labeling, our method will also facilitate transforming a sketch into other interesting forms, such as clip-art illustrations. More specifically, given a user-sketched scene, multiple objects can be more reliably recognized and then replaced with homogeneous clip-art pieces to synthesize artistic images. Our approach could also be extended to enable a suggestive interface, where users can specify a location in the scene with simple gestures and then compatible objects will be automatically suggested. This could be achieved by simply ignoring the CNN prediction term in Eq. 4 and ranking all the available categories in the relation graph by the computed scores. Such suggestive interfaces would be of help to design exploration.

We have tried an end-to-end learning approach. The main challenge is that we do not have sufficient sketch scene data to train a deep neural network. Therefore, similar to [Lun et al. 2017], we used the bounding boxes and labels in natural images as templates and filled in sketched objects to generate scene sketches for training. For each scene sketch image, we randomly chose an object as target and the others as context. We then conducted a joint training method to classify the target object with the context information as additional input. The result was promising but not competitive to our work. This is partially because the layouts of natural images are different from those of sketch images, and it may be challenging for the neural network to learn transferable relationships directly. In the future, we will consider to collect a database of sketched scenes

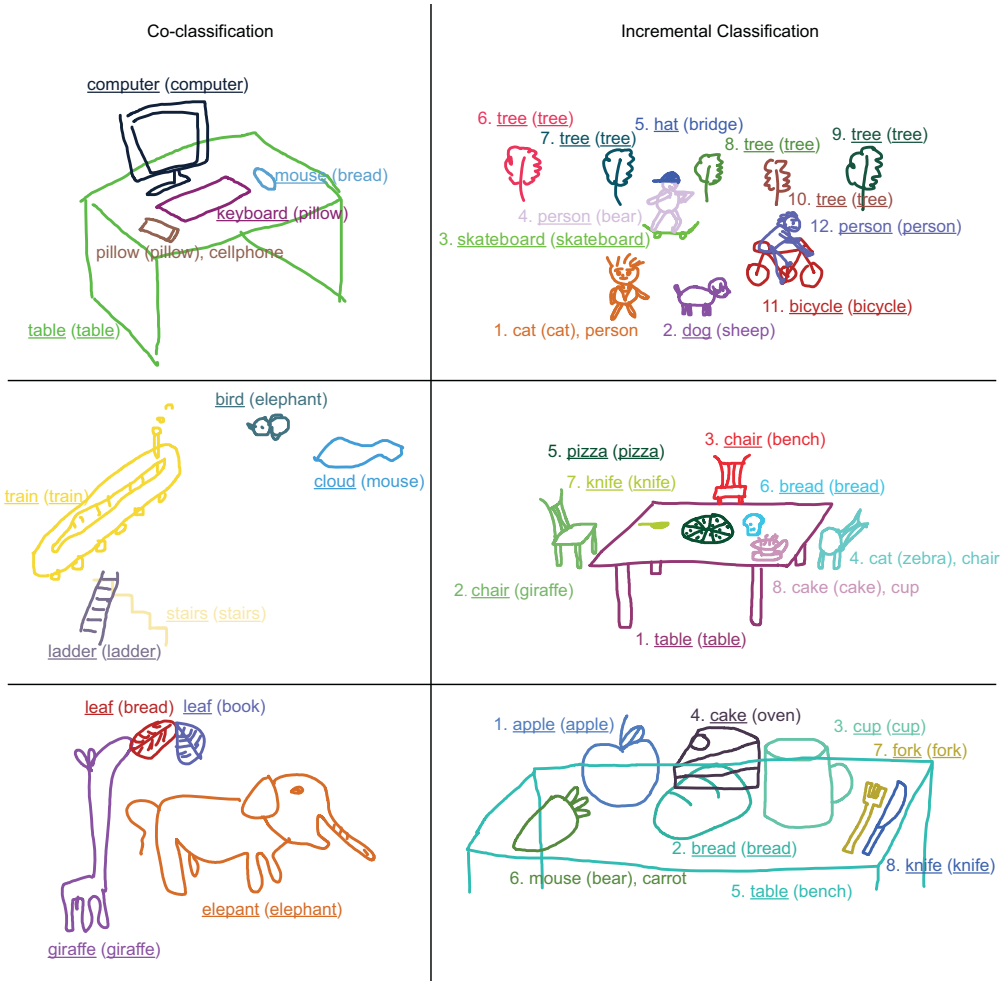

Figure 8: More classification results. Left column: co-classification results; Right column: incremental classification results. The labeled text indicates the prediction of our algorithm and the prediction of the CNN within the parentheses. For incremental classification results, the sequence of drawing is indicated before each label. The underlined labels are consistent with the given ground truth. If none of the predictions are correct, we show the given ground truth at the last.

consisting of various objects and learn the context information directly from such a database. Another interesting future work is to unify the object-level sketched scene segmentation and sketch recognition into a deep learning framework.

ACKNOWLEDGMENTS

This work was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU11300615, CityU11204014, HKUST16201315), and ACIM-SCM.

REFERENCES

Christine Alvarado and Randall Davis. 2004. SketchREAD: A Multi-domain Sketch Recognition Engine. In *Proc. ACM UIST*. ACM, 10. https://doi.org/10.1145/1029632.1029637

Relja Arandjelović and Tevfik Metin Sezgin. 2011. Sketch recognition by fusion of temporal and image-based features. *Pattern Recognition* 44, 6 (2011), 1225–1234.

T Bui, L Ribeiro, M Ponti, and John Collomosse. 2017. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. *Computer Vision and Image Understanding* 164 (2017), 27–37.

Xiaochun Cao, Hua Zhang, Si Liu, Xiaojie Guo, and Liang Lin. 2013. SYM-FISH: A Symmetry-Aware Flip Invariant Sketch Histogram Shape Descriptor. In *Proc. IEEE ICCV*.

Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. 2009. Sketch2Photo: Internet Image Montage. *ACM TOG* 28, 5, Article 124 (Dec. 2009), 10 pages. https://doi.org/10.1145/1618452.1618470

N. Donmez and K. Singh. 2012. Concepture: A Regular Language Based Framework for Recognizing Gestures with Varying and Repetitive Patterns. In *Proc. SBIM*. Eurographics Association, 9.

Mathias Eitz, James Hays, and Marc Alexa. 2012. How Do Humans Sketch Objects? *ACM TOG* 31, 4, Article 44 (July 2012), 10 pages. https://doi.org/10.1145/2185520.2185540

M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. 2011. Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors. *IEEE TVCG* 17, 11 (Nov 2011), 1624–1636. https://doi.org/10.1109/TVCG.2010.266

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE TPAMI* 32, 9 (Sept 2010), 1627–1645. https://doi.org/10.1109/TPAMI.2009.167

Matthew Fisher and Pat Hanrahan. 2010. Context-based Search for 3D Models. *ACM TOG* 29, 6, Article 182 (Dec. 2010), 10 pages. https://doi.org/10.1145/1882261.1866204

Carolina Galleguillos and Serge Belongie. 2010. Context Based Object Categorization: A Critical Survey. *CVIU* 114, 6 (June 2010), 712–722. https://doi.org/10.1016/j.cviu.2010.02.004

David Ha and Douglas Eck. 2017. A Neural Representation of Sketch Drawings. *CoRR* abs/1704.03477 (2017).

Rui Hu and John Collomosse. 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding* 117, 7 (2013), 790–806.

Zhe Huang, Hongbo Fu, and Rynson W. H. Lau. 2014. Data-driven Segmentation and Labeling of Freehand Sketches. *ACM TOG* 33, 6, Article 175 (Nov. 2014), 10 pages. https://doi.org/10.1145/2661229.2661280

Levent Burak Kara and Thomas F. Stahovich. 2004. Hierarchical Parsing and Recognition of Hand-sketched Diagrams. In *Proc. ACM UIST*. ACM, 13–22. https://doi.org/10.1145/1029632.1029636

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV* 123, 1 (01 May 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

Joseph J. LaViola, Jr. and Robert C. Zeleznik. 2004. MathPad2: A System for the Creation and Exploration of Mathematical Sketches. *ACM TOG* 23, 3 (Aug. 2004), 432–440. https://doi.org/10.1145/1015706.1015741

Bo Li, Yijuan Lu, and R. Fares. 2013. Semantic sketch-based 3D model retrieval. In *ICMEW*. 1–4. https://doi.org/10.1109/ICMEW.2013.6618316

L. J. Li, R. Socher, and L. Fei-Fei. 2009. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proc. IEEE CVPR*. 2036–2043. https://doi.org/10.1109/CVPR.2009.5206718

Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. 2015. Free-hand sketch recognition by multi-kernel feature learning. *CVIU* 137 (2015), 1 – 11. https://doi.org/10.1016/j.cviu.2015.02.003

Tong Lu, Chiew-Lan Tai, Feng Su, and Shijie Cai. 2005. A New Recognition Model for Electronic Architectural Drawings. *CAD* 37, 10 (2005), 1053 – 1069. https://doi.org/10.1016/j.cad.2004.11.004

Zhaoliang Lun, Changqing Zou, Haibin Huang, Evangelos Kalogerakis, Ping Tan, Marie-Paule Cani, and Hao Zhang. 2017. Learning to group discrete graphical patterns. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 225.

Tomasz Malisiewicz and Alyosha Efros. 2009. Beyond categories: The visual memex model for reasoning about object relationships. In *Advances in neural information processing systems*. 1222–1230.

R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, R. Urtasun, and A. Yuille. 2014. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Proc. IEEE CVPR*. 891–898. https://doi.org/10.1109/CVPR.2014.119

Tom Y. Ouyang and Randall Davis. 2011. ChemInk: A Natural Real-time Recognition System for Chemical Drawings. In *Proc. ACM IUI*. ACM, 10. https://doi.org/10.1145/1943403.1943444

Brandon Paulson and Tracy Hammond. 2008. PaleoSketch: Accurate Primitive Sketch Recognition and Beautification. In *Proc. ACM IUI*. ACM, 10. https://doi.org/10.1145/1378773.1378775

Tiziano Portenier, Qiyang Hu, Paolo Favaro, and Matthias Zwicker. 2017. SmartSketcher: Sketch-based Image Retrieval with Dynamic Semantic Re-ranking. In *Proc. SBIM*. ACM, Article 1, 12 pages. https://doi.org/10.1145/3092907.3092910

Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. 2016. Sketch-based image retrieval via siamese convolutional neural network. In *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2460–2464.

X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang. 2016. Enhancing Sketch-Based Image Retrieval by Re-Ranking and Relevance Feedback. *IEEE TIP* 25, 1 (Jan 2016), 195–208. https://doi.org/10.1109/TIP.2015.2497145

A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. 2007. Objects in Context. In *Proc. IEEE ICCV*. 1–8. https://doi.org/10.1109/ICCV.2007.4408986

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115, 3 (01 Dec 2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM TOG* 35, 4, Article 119 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925954

Rosália G. Schneider and Tinne Tuytelaars. 2014. Sketch Classification and Classification-driven Analysis Using Fisher Vectors. *ACM TOG* 33, 6, Article 174 (Nov. 2014), 9 pages. https://doi.org/10.1145/2661229.2661231

Tevfik Metin Sezgin and Randall Davis. 2005. HMM-based efficient sketch recognition. In *Proceedings of the 10th international conference on Intelligent user interfaces*. ACM, 281–283.

Tevfik Metin Sezgin and Randall Davis. 2008. Sketch recognition in interspersed drawings using time-based graphical models. *Computers & Graphics* 32, 5 (2008), 500–510.

Tevfik Metin Sezgin, Thomas Stahovich, and Randall Davis. 2001. Sketch Based Interfaces: Early Processing for Sketch Understanding. In *Proceedings of Workshop on Perceptive User Interfaces*. ACM, 8. https://doi.org/10.1145/971478.971487

HyoJong Shin and Takeo Igarashi. 2007. Magic Canvas: Interactive Design of a 3-D Scene Prototype from Freehand Sketches. In *Proceedings of Graphics Interface*. ACM, 63–70. https://doi.org/10.1145/1268517.1268530

Oana Sidi, Oliver van Kaick, Yanir Kleiman, Hao Zhang, and Daniel Cohen-Or. 2011. Unsupervised Co-segmentation of a Set of Shapes via Descriptor-space Spectral Clustering. *ACM TOG* 30, 6, Article 126 (Dec. 2011), 10 pages. https://doi.org/10.1145/2070781.2024160

J. Sivic and A. Zisserman. 2003. Video Google: a Text Retrieval Approach to Object Matching in Videos. In *Proc. IEEE ICCV*. https://doi.org/10.1109/ICCV.2003.1238663

Zhenbang Sun, Changhu Wang, Liqing Zhang, and Lei Zhang. 2012. Free hand-drawn sketch segmentation. In *European Conference on Computer Vision*. Springer, 626–639.

Ivan E. Sutherland. 1964. Sketchpad: A Man-machine Graphical Communication System. In *DAC*. ACM, 6.329–6.346. https://doi.org/10.1145/800265.810742

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going Deeper With Convolutions. In *Proc. IEEE CVPR*.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. IEEE CVPR*. https://doi.org/10.1109/CVPR.2010.5539970

Kun Xu, Kang Chen, Hongbo Fu, Wei-Lun Sun, and Shi-Min Hu. 2013. Sketch2Scene: Sketch-based Co-retrieval and Co-placement of 3D Models. *ACM TOG* 32, 4, Article 123 (July 2013), 15 pages. https://doi.org/10.1145/2461912.2461968

Kemal Tugrul Yesilbek and T Metin Sezgin. 2017. Sketch recognition with few examples. *Computers & Graphics* 69 (2017), 80–91.

Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. 2016. Sketch Me That Shoe. In *Proc. IEEE CVPR*.

Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2017. Sketch-a-Net: A Deep Neural Network that Beats Humans. *IJCV* 122, 3 (01 May 2017), 411–425.

H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. 2016. SketchNet: Sketch Classification with Web Images. In *Proc. IEEE CVPR*. https://doi.org/10.1109/CVPR.2016.125