Sketch Less for More:

On-the-fly Fine-grained Sketch-based Image Retrieval





Ayan Kumar Bhunia[a] Yongxin Yang[a] Timothy Hospedales[a,b] Tao Xiang[a] Yi-Zhe Song[a]

a. SketchX Lab, CVSSP, University of Surrey, UK
b. University of Edinburgh, UK
http://sketchx.ai

CVSSP Centre for Vision, Speech and Signal Processing SketchX

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

UNIVERSITY OF SURREY

Fine-grained SBIR

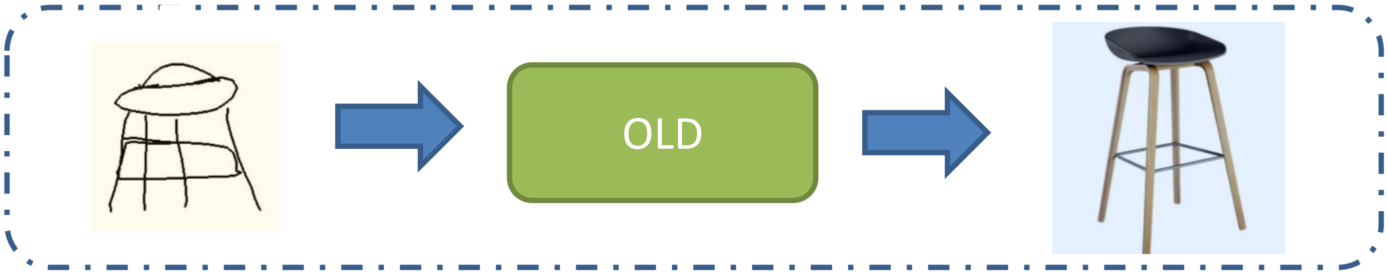
Sketch

Gallery Images

Problem – “I can’t sketch”

- **Time** taken to draw a *complete* sketch
- **Drawing skill** of the user

Old Setup: sketch first, *then* retrieve



New *On-the-fly* Setup: retrieve *as* you sketch

Less is more!

Why *On-the-fly*?

- **Natural**: incomplete sketches can *already* retrieve!
- **Faster**: *no need* to sketch the whole thing
- **More accurate**: modelling the *sketching process* does help

In most cases, we can retrieve with ~30% less strokes!

Why Challenging?

- Framework to model **dynamic sketching** for FG-SBIR
- Specific designs to handle **incomplete sketches**

Sketch Me *That* Shoe

Qian Yu[1] Feng Liu[1,2]
Yi-Zhe Song[1] Tao Xiang[1] Timothy M. Hospedales[1] Chen Change Loy[3]
Queen Mary University of London, London, UK[1]
Southeast University, Nanjing, China[2] The Chinese University of Hong Kong, Hong Kong, China[3]
{q.yu, feng.liu, yizhe.song, t.xiang, t.hospedales}@qmul.ac.uk ccloy@ie.cuhk.edu.hk

Abstract

We investigate the problem of fine-grained sketch-based image retrieval (SBIR), where free-hand human sketches are used as queries to perform instance-level retrieval of images. This is an extremely challenging task because (i) visual comparisons not only need to be fine-grained but also executed cross-domain, (ii) free-hand (finger) sketches are highly abstract, making fine-grained matching harder, and most importantly (iii) annotated cross-domain sketch-photo datasets required for training are scarce, challenging many state-of-the-art machine learning techniques.

In this paper, for the first time, we address all these challenges, providing a step towards the capabilities that would underpin a commercial sketch-based image retrieval application. We introduce a new database of 1,432 sketch-photo pairs from two categories with 32,000 fine-grained triplet ranking annotations. We then develop a deep triplet-ranking model for instance-level SBIR with a novel data augmentation and staged pre-training strategy to alleviate the issue of insufficient fine-grained training data. Extensive experiments are carried out to contribute a variety of insights into the challenges of data sufficiency and over-fitting avoidance when training deep networks for fine-grained cross-domain ranking tasks.

Figure 1. Free-hand sketch is ideal for fine-grained instance-level image retrieval.

However, existing SBIR works largely overlook such fine-grained details, and mainly focus on retrieving images of the same category [21, 22, 10, 2, 3, 27, 12, 19, 13, 28, 11], thus not exploiting the real strength of SBIR. This oversight pre-emptively limits the practical use of SBIR since text is often a simpler form of input when only category-level retrieval is required, e.g., one would rather type in the word "shoe" to retrieve one rather than sketching a shoe. The existing commercial image search engines have already done a pretty good job on category-level image retrieval. In contrast, it is when aiming to retrieve *a particular shoe* that sketching may be preferable than elucidating a long textual

[A]

Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval

Jifei Song* Qian Yu* Yi-Zhe Song Tao Xiang Timothy M. Hospedales
Queen Mary University of London University of Edinburgh
{j.song, q.yu, yizhe.song, t.xiang}@qmul.ac.uk, t.hospedales@ed.ac.uk

Abstract

Human sketches are unique in being able to capture both the spatial topology of a visual object, as well as its subtle appearance details. Fine-grained sketch-based image retrieval (FG-SBIR) importantly leverages on such fine-grained characteristics of sketches to conduct instance-level retrieval of photos. Nevertheless, human sketches are often highly abstract and iconic, resulting in severe misalignments with candidate photos which in turn make subtle visual detail matching difficult. Existing FG-SBIR approaches focus only on coarse holistic matching via deep cross-domain representation learning, yet ignore explicitly accounting for fine-grained details and their spatial context. In this paper, a novel deep FG-SBIR model is proposed which differs significantly from the existing models in that: (1) It is spatially aware, achieved by introducing an attention module that is sensitive to the spatial position of visual details; (2) It combines coarse and fine semantic information via a shortcut connection fusion block; and (3) It models feature correlation and is robust to misalignments between the extracted features across the two domains by introducing a novel higher order learnable energy function (HOLEF) based loss. Extensive experiments show that the proposed deep spatial-semantic attention model significantly outperforms the state-of-the-art.

Figure 1. FG-SBIR is challenging due to the misalignment of the domains (left) and subtle local appearance differences between a true match photo and a visually similar incorrect match (right).

bags by finger-sketching on a smart-phone screen.

FG-SBIR is a very challenging problem and remains unsolved. First, there is a large domain gap between sketch and photo – a sketch captures mainly object shape/contour information and contains no information on colour and very little on texture. Second, FG-SBIR is typically based on free-hand sketches which are drawn based on mental recollection of reference images shown moments before the drawing stage, making free-hand sketches distinctly more abstract than line tracings (human edgemaps). As a result, a sketch and its matched photo could have large discrepancies in shape and spatial misalignment both globally and locally. Finally, as an object instance recognition problem,

[B]

Generalising Fine-Grained Sketch-Based Image Retrieval

Kaiyue Pang[1,2]* Ke Li[1,3]* Yongxin Yang[1] Honggang Zhang[3]
Timothy M. Hospedales[1,4] Tao Xiang[1] Yi-Zhe Song[1]
[1]SketchX, CVSSP, University of Surrey [2]Queen Mary University of London
[3]Beijing University of Posts and Telecommunications [4]The University of Edinburgh
kaiyue.pang@qmul.ac.uk, {yongxin.yang, t.xiang, y.song}@surrey.ac.uk
{like1990, zhhg}@bupt.edu.cn, t.hospedales@ed.ac.uk

Abstract

Fine-grained sketch-based image retrieval (FG-SBIR) addresses matching specific photo instance using free-hand sketch as a query modality. Existing models aim to learn an embedding space in which sketch and photo can be directly compared. While successful, they require instance-level pairing within each coarse-grained category as annotated training data. Since the learned embedding space is domain-specific, these models do not generalise well across categories. This limits the practical applicability of FG-SBIR. In this paper, we identify cross-category generalisation for FG-SBIR as a domain generalisation problem, and propose the first solution. Our key contribution is a novel unsupervised learning approach to model a universal manifold of prototypical visual sketch traits. This manifold can then be used to paramaterise the learning of a sketch/photo representation. Model adaptation to novel categories then becomes automatic via embedding the novel sketch in the manifold and updating the representation and retrieval function accordingly. Experiments on the two largest FG-SBIR datasets, Sketchy and QMUL-Shoe-V2, demonstrate the efficacy of our approach in enabling cross-category generalisation of FG-SBIR.

Recent FG-SBIR methods [24, 36, 28, 22] address this issue by learning a deep network embedding of sketch and photo that makes them directly comparable. This embedding is often trained by a triplet ranking loss to ensure that the network embeds positive pairs nearby, and negative pairs farther apart. This line of work has made great progress, with state-of-the-art approaching human performance [22] on the Sketchy benchmark [24].

Nevertheless, existing work has thus far implicitly assumed that instance-level annotations of positive and negative pairs are available for every coarse category to be evaluated. This assumption limits the practical applicability of FG-SBIR. More specifically, as we shall show in this paper, in practice FG-SBIR generalises very poorly if training and testing categories are disjoint. This is of course unsatisfactory for potential users of FG-SBIR such as e-commerce, where it would be desirable to train a FG-SBIR system once on an initial set of product categories, and then have it deployed directly to newly added product categories – without needing to collect and annotate new data and retrain the FG-SBIR model. Compared to other category-level tasks such as object recognition in photo images, this annotation barrier is particularly high for FG-SBIR as instance-specific sketches are expensive and slow to collect.

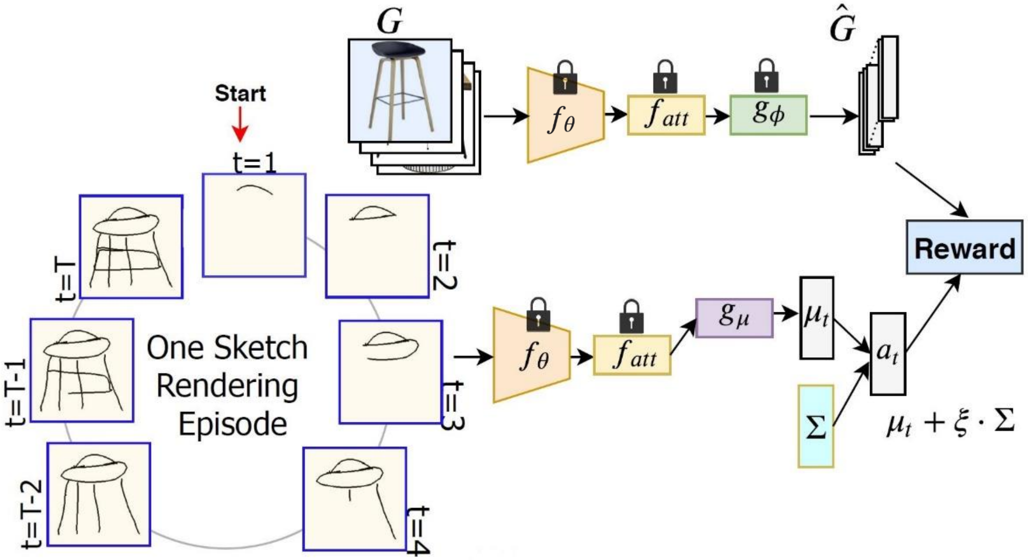
To understand why the existing FG-SBIR models have

[C]

A. Sketch Me That Shoe, Qian et al., CVPR 2016
B. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval, Song et al., ICCV 2017
C. Generalising fine-grained sketch-based image retrieval, Pang et al., CVPR 2019

Contributions

- **Reinforcement Learning** (RL) for cross-modal modelling.
- **Reward design** to encourage early retrieval
- **Rank optimization** over a complete sketch drawing episode



RL for FG-SBIR

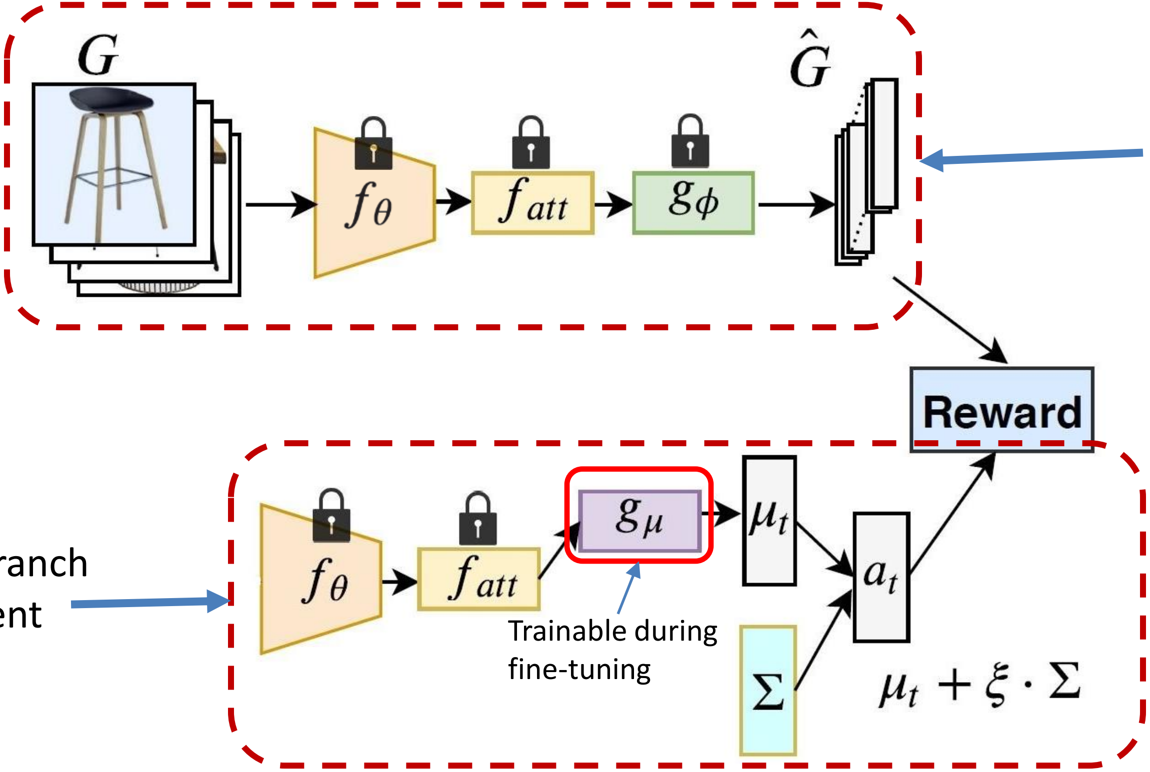


Photo branch fixed once trained

Sketch branch as RL agent

Reward Design

Total Reward

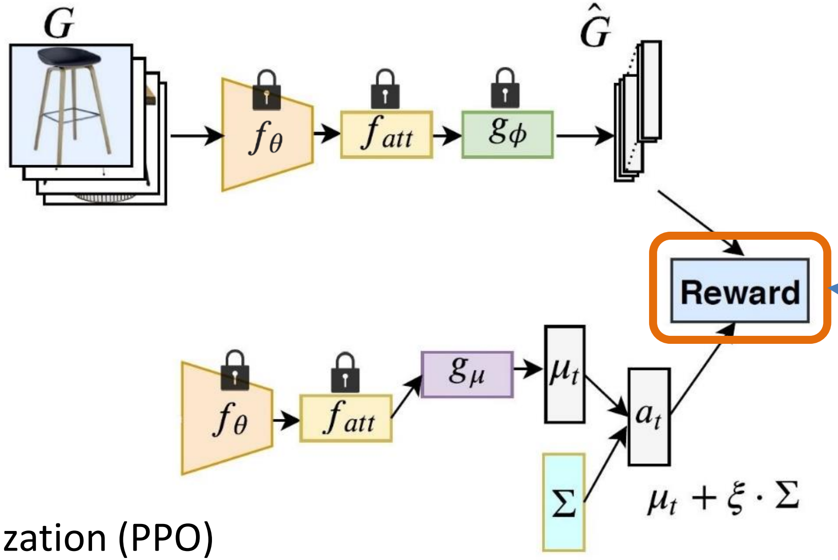
$$R_t = \gamma_1 \boxed{R_t^{Local}} + \gamma_2 \boxed{R_t^{Global}}$$

$$R_t^{Local} = \frac{1}{rank_t}$$

Local Reward

$$R_t^{Global} = -\max(0, \tau(L_t, L_{t+1}) - \tau(L_{t-1}, L_t))$$

Global Reward



Local to ensure early retrieval

Global as regularizer

Proximal Policy Optimization (PPO)

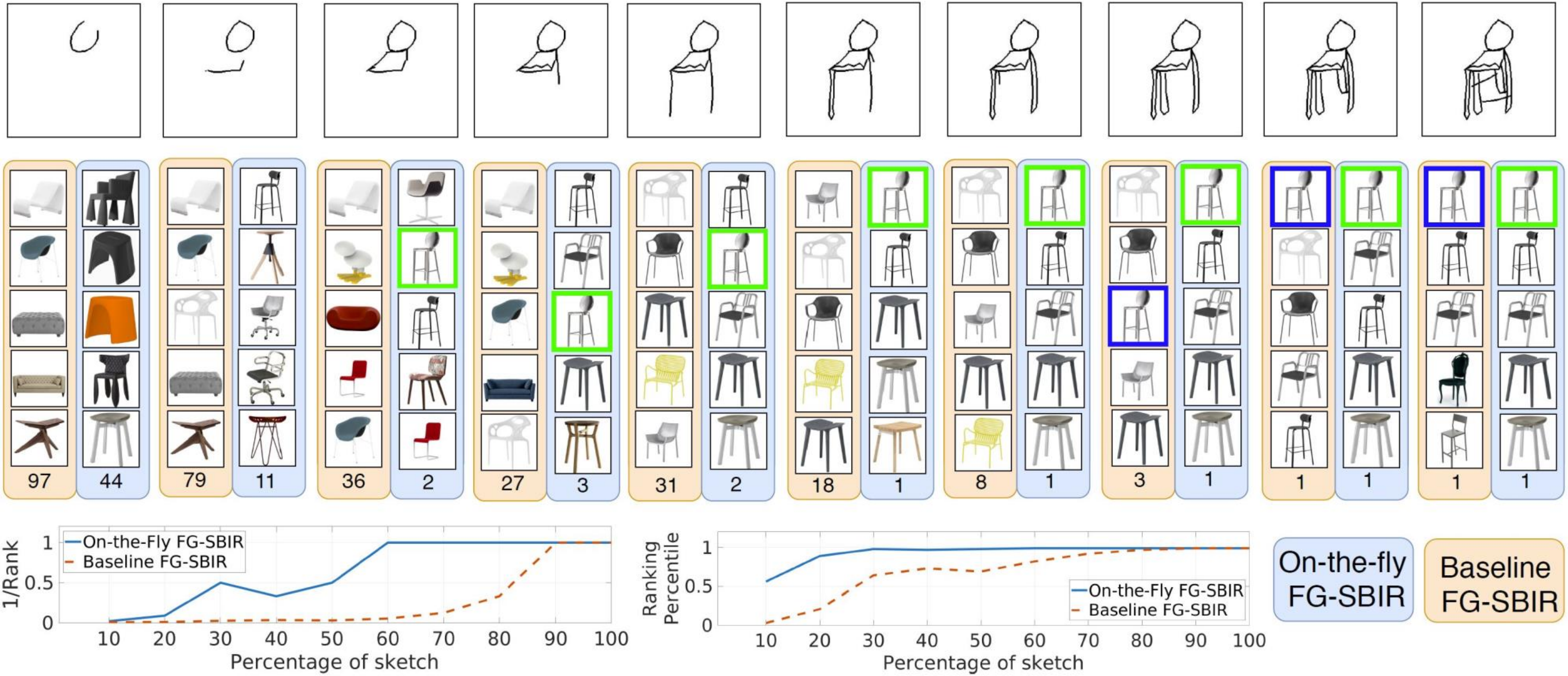
Experiments

- **Datasets:** QMUL-Shoe-V2 & QMUL-Chair-V2

- **Evaluation Metric:**
  - top-q accuracy (A@q)
  - area under ranking percentile *vs* percentage of sketch (m@A)
  - area under 1/rank *vs* percentage of sketch (m@B)
- Baselines:
  - basic triplet loss models [A, B]
  - a triplet model that *uses all intermediate incomplete sketches* as training data.
  - 20 different models each dealing with a specific percentage of sketch (e.g., 5%, 10%, ..., 100%)
  - [C] as a generalized solution to approximate rankings

A. Sketch Me That Shoe, in CVPR 2016
B. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval, in ICCV 2017
C. SoDeep: A Sorting Deep Net to Learn Ranking Loss Surrogate, in CVPR 2019

Results



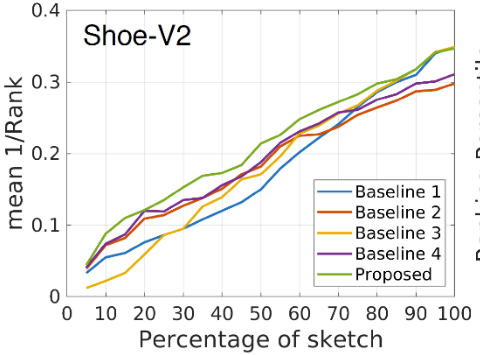
97
44
79
11
36
2
27
3
31
2
18
1
8
1
3
1
1
1
1
1
1/Rank
On-the-Fly FG-SBIR
Baseline FG-SBIR
Percentage of sketch
Ranking Percentile
On-the-Fly FG-SBIR
Baseline FG-SBIR
Percentage of sketch
On-the-fly FG-SBIR
Baseline FG-SBIR

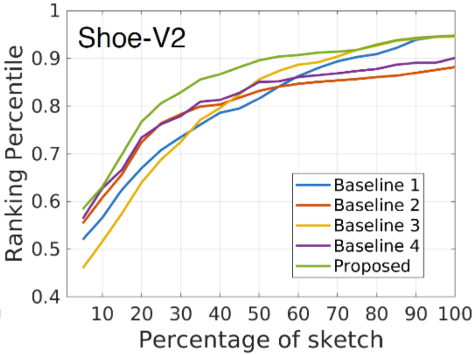
Results

Quantitative Results on Different Baselines (A@q, m@A, and m@B)

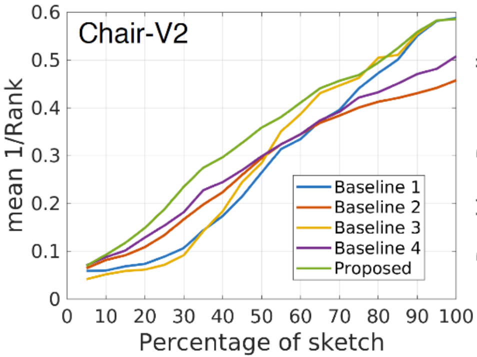
	Chair-V2				Shoe-V2			
	m@A	m@B	A@5	A@10	m@A	m@B	A@5	A@10
B1	77.18	29.04	**76.47**	88.13	80.12	18.05	65.69	**79.69**
B2	80.46	28.07	74.31	86.69	79.72	18.75	61.79	76.64
B3	76.99	30.27	**76.47**	88.13	80.13	18.46	65.69	79.69
B4	81.24	29.85	75.14	87.69	81.02	19.50	62.34	77.24
TS	76.01	27.64	73.47	85.13	77.12	17.13	62.67	76.47
Ours	**85.44**	**35.09**	76.34	**89.65**	**85.38**	**21.44**	**65.77**	79.63

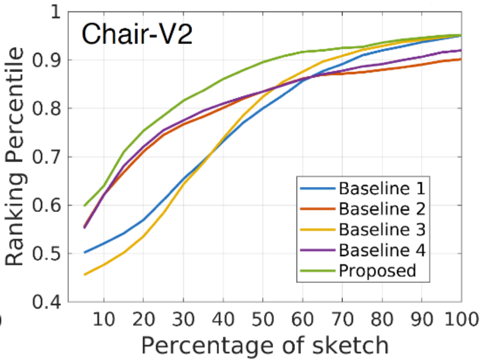
Percentage-wise Results for Shoe-V2 (m@A, and m@B)



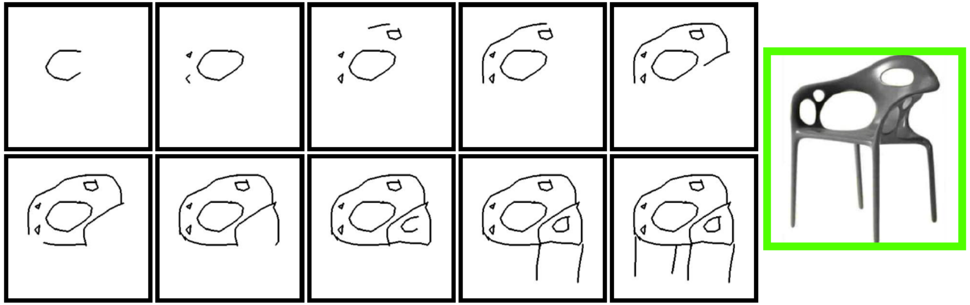


Percentage-wise Results for Chair-V2 (m@A, and m@B)

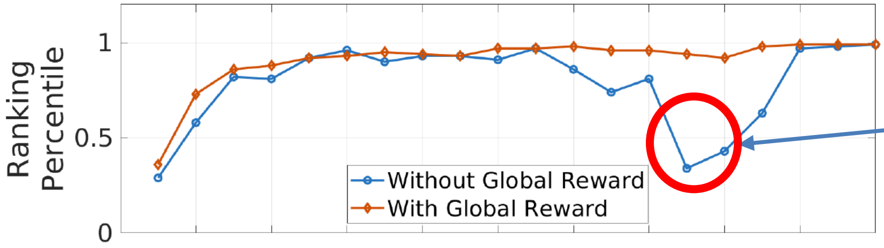




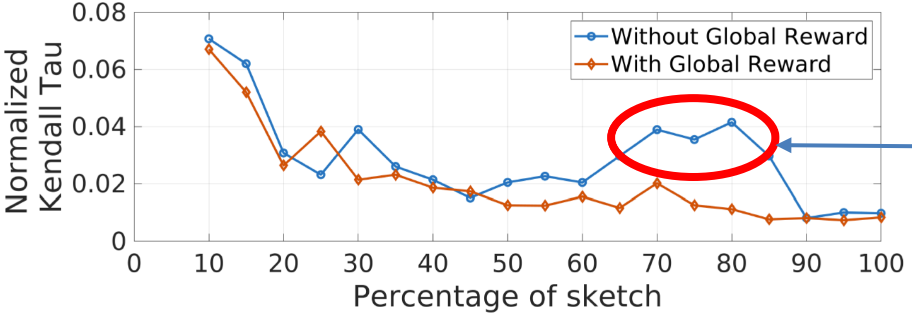
Ablation Study



Progressive order of sketching



Drop in ranking percentile



Corresponding explosive increase of *Kendall-Tau distance*

Ablation Study

Comparative Study with Different RL Methods (m@A, and m@B)

RL Methods	Chair-V2		Shoe-V2	
	m@A	m@B	m@A	m@B
Vanilla Policy Gradient	80.36	32.34	82.56	19.67
PPO-AC-Clipping	81.54	33.71	83.47	20.84
PPO-AC-KL Penalty	80.99	32.64	83.84	20.04
PPO-A-KL Penalty	81.34	33.01	83.51	20.66
TRPO	83.21	33.68	83.61	20.31
PPO-A-Clipping (Ours)	**85.44**	**35.09**	**85.38**	**21.44**

Comparative Study with Candidate Reward Designs (m@A, and m@B)

Reward Schemes	Chair-V2		Shoe-V2	
	m@A	m@B	m@A	m@B
$\text{rank} \leq 1 \Rightarrow \text{reward} = 1$	82.99	32.46	82.24	19.87
$\text{rank} \leq 5 \Rightarrow \text{reward} = 1$	81.36	31.94	81.74	19.37
$\text{rank} \leq 10 \Rightarrow \text{reward} = 1$	80.64	30.57	80.87	19.08
$-\text{rank}$	83.71	32.84	83.81	20.71
$\frac{1}{\sqrt{\text{rank}}}$	83.71	33.97	83.67	20.49
$\frac{1}{\text{rank}}$	84.33	34.11	84.07	20.54
Ours (Eq. 4)	**85.44**	**35.09**	**85.38**	**21.44**